Economics 686 Lecture Notes

Gordon R. Fisher[1]

Department of Economics

Concordia University

1455 de Maisonneuve West

Montreal, Quebec

H3G 1M8

May 20, 2001

Contents

1 Basic Matrix Algebra

1.1 Vectors in R^n

Much of the discussion will be concerned with real numbers. The set of real numbers is denoted by R. If x is a real number, then x is an element of R, which is written $x \in R$. An array of n real numbers, $x_1; x_2; ¢¢¢; x_n$, arranged as a row $[x_1; x_2; ¢¢¢; x_n]$ or as a column

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

is called an n-tuple. An n-tuple, $n = 1; 2; 3; ¢¢¢$, is a special case of a vector: a $(1 \times n)$ row-vector or an $(n \times 1)$ column vector. Generally, n-tuples are elements of an n-dimensional vector space de¯ned over the ¯eld of real numbers, R. Here, the set of real numbers and the ¯eld of real numbers are synonymous. If x represents an n-tuple of real numbers, then x is said to be an element of n-dimensional vector space de¯ned over the ¯eld of real numbers, R. This vector space is written R^n whereupon

$x \in R^n$ and

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

$x_i \in R^n$; $i = 1; 2; ¢¢¢; n$. The column vector, x, may also be represented as a row by transposing it to yield $x^>$:

$$x^> = \begin{bmatrix} x_1 & x_2 & ¢¢¢ & x_n \end{bmatrix}.$$

Normally, if $x \in R^n$, then x is to be thought of as a column or $(n £ 1)$ vector. If w and z are each elements of R^n, they are said to be conformable, whereupon their sum, $w + z$, and their scalar product may be de¯ned. If the elements of w and z are w_i and z_i, $i = 1; 2; ¢¢¢; n$, then

$$w + z = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} + \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} w_1 + z_1 \\ w_2 + z_2 \\ \vdots \\ w_n + z_n \end{bmatrix}.$$

The scalar (or inner) product is de¯ned by

$$w^>z = \sum_i w_i z_i \in R.$$

Of course, because $w^{>}z$ is a real number (referred to as a scalar, in general),

$$w^{>}z = z^{>}w.$$

The scalar product of two (conformable) vectors plays an important role because it is a means of measuring length. For $x \in R^n$, the (Euclidean) length of x is written $\|x\|$ and is defined by

$$\|x\| = \sqrt{(x^{>}x)} = \sqrt{\sum x_i^2},$$

thus ensuring that length is always a positive number, or zero.

Consider now $k < n$ vectors $x_{:1}; x_{:2}; ¢¢¢; x_{:k}$, from R^n. These may be arranged as n rows and k columns denoted by X

$$[x_{:1}; x_{:2}; ¢¢¢; x_{:k}] = \begin{bmatrix} x_{11} & x_{12} & ¢¢¢ & x_{1k} \\ x_{21} & x_{22} & ¢¢¢ & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \vdots\vdots\vdots & x_{nk} \end{bmatrix} = X.$$

X is an $(n \pounds k)$ matrix. X may be expressed as k columns of n elements $x_{:j}$ in R^n, $j = 1; 2; ¢¢¢; k$, or as n rows of k elements $x_{i:}$ in R^k, $i = 1; 2; ¢¢¢; n$:

$$X = \begin{bmatrix} x_{1:} \\ x_{2:} \\ \vdots \\ x_{n:} \end{bmatrix}.$$

Combining the two alternative vector notations, $x_{:j}$ and $x_{i:}$, the nk individual real numbers comprising X are x_{ij}, i = 1; 2; ¢¢¢; n; j = 1; 2; ¢¢¢; k, the subscripts i; j denoting position in the i'th row and j'th column.

Some discussion will be centered on the linear model

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + ¢¢¢ + x_{ik}\beta_k + \varepsilon_i \tag{1.1}$$

in which i = 1; 2; ¢¢¢; n, n > k as above. Equation (1.1) may also be written as

$$y_i = \sum_j x_{ij}\beta_j + \varepsilon_i \tag{1.2}$$

where j = 1; 2; ¢¢¢; k; or as

$$y = x_{:1}\beta_1 + x_{:2}\beta_2 + ¢¢¢ + x_{:k}\beta_k + \varepsilon \tag{1.3}$$

where y, the k vectors $x_{:j}$ and " are all n-tuples of elements y_i, x_{ij} and ε_i respectively. Finally, let $\beta \in R^k$ be the (k £ 1) vector of elements $\beta_1; \beta_2; ¢¢¢; \beta_k$, then equations (1:1), (1:2) and (1:3) may be written compactly as

$$y = X\beta + \varepsilon, \tag{1.4}$$

X being an (n £ k) matrix. In the usual interpretation, y; $x_{:1}$; $x_{.2}$; ¢¢¢; $x_{:k}$ are vectors of n observations on each of the k + 1 variables, β is a vector of k coe± cients and " is regarded as a vector of n unknown errors. Given y; X and β, then " = y ¡ $X\beta$. In practice, neither β nor " is known and hence must be estimated in some way. To begin to study this problem, it is useful to understand some of the theory of linear spaces. As a ¯rst step, some of the algebra of matrices is reviewed.

1.2 Some Simple Matrices

Generally, vectors will be denoted by lower case letters and matrices by upper case letters: thus x will generally be a column vector and X a matrix. Consider the $(n \times m)$ matrix A of elements $a_{ij} \in R;\ i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m$. It is common to write $A = [a_{ij}]$ to describe a matrix and its elements. If $n \neq m$, then A is said to be rectangular; if $n = m$, then A is said to be square. Matrices of elements from R are said to be real matrices, meaning matrices comprising real elements. A matrix comprising zero elements is called the null matrix.

When A is square and $a_{ij} = \pm_{ij}$ where $\pm_{ij}$ is Kronecker's delta (i.e. $\pm_{ij} = 1$ for $i = j$ and $\pm_{ij} = 0$ for $i \neq j,\ i, j = 1, 2, \ldots, n$) then

$$A = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} = I.$$

I is the identity matrix. When I is $(n \times n)$, it is common to write the identity matrix as I_n so that its order, $(n \times n)$, is recognized explicitly in the notation. Similarly, an $(n \times m)$ matrix is said to be of order $(n \times m)$. If a matrix is simply of order n, then it is square with n rows and n columns.

As in the case of the matrix X in section 1.1, the notation

$$A^{>} = [a_{ji}] = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}.$$

Thus, $A^{>}$ is $(m \times n)$.

If A and B are of the same order, then $A + B$ and $A - B$ are defined as the matrices C and D respectively as follows

$$A + B = [a_{ij} + b_{ij}] = C$$

$$A - B = [a_{ij} - b_{ij}] = D.$$

Matrices that are not of the same order cannot be added together or subtracted from one another. If $\lambda \in R$, i.e. λ is a scalar, then

$$A\lambda = \lambda A = [\lambda a_{ij}].$$

This operation is called scalar multiplication.

The product of two matrices A and B, say AB, is only defined when the number of columns of A is equal to the number of rows of B. Thus if A is $(n \times k)$ and B is $(k \times m)$ then

$$AB = \left[\sum_{q=1}^{k} a_{iq} b_{qj} \right]$$

for $q = 1; 2; ¢¢¢; k$; $i = 1; 2; ¢¢¢; n$ and $j = 1; 2; ¢¢¢; m$. The product, AB, then has order $(n £ m)$. Writing down the orders explicitly:

$$\underset{(n £ k)}{A}\ \underset{(k £ m)}{B} = \underset{(n £ m)}{C} = \left[\sum_{q} a_{iq} b_{qj} \right] :$$

Thus, the inner orders must be the same (k) and the outer orders (n; m) determine the order of the product. It follows that, if AB is de¯ned, BA is not de¯ned unless the outer orders are the same $(n = m)$. When $(n = m)$, AB is $(n £ n)$ and BA is $(k £ k)$, whereupon AB 6= BA.

Notice that, for vectors x and y in R^n, while the scalar, or inner, product $x^> y$ is a scalar, the outer product $xy^>$ is an $(n £ n)$ matrix.

Matrix multiplication and addition have the distributive property

$$(A + B) C = AB + AC$$

and both operations are associative

$$(A + B) + C = A + (B + C)$$

$$(AB) C = A (BC) .$$

Since AB 6= BA in general, matrix multiplication is not commutative. For any $(n £ m)$ matrix A,

$$I_n A = A = A I_m.$$

If $A = A^{>}$, A is said to be symmetric, because $a_{ij} = a_{ji}$, except when i = j. A cannot, therefore, be symmetric unless it is square. The rule for the transpose of a matrix product is

$$(AB)^{>} = B^{>}A^{>}. \tag{1.5}$$

Notice that both $A^{>}A$ and $AA^{>}$ are both symmetric, since by applying rule (1:5)

$$\begin{aligned} \left(A^{>}A\right)^{>} &= A^{>}A \\ \left(AA^{>}\right)^{>} &= AA^{>}. \end{aligned}$$

Turning back to equation (1:4), a matrix of some importance in connection with the estimation of this equation is

$$X^{>}X = \begin{bmatrix} x_{:1}^{>}x_{:1} & x_{:1}^{>}x_{:2} & \cdots & x_{:1}^{>}x_{:k} \\ x_{:2}^{>}x_{:1} & x_{:2}^{>}x_{:2} & \cdots & x_{:2}^{>}x_{:k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{:k}^{>}x_{:1} & x_{:k}^{>}x_{:2} & \cdots & x_{:k}^{>}x_{:k} \end{bmatrix}$$

$$= \begin{bmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \cdots & \sum x_{i1}x_{ik} \\ \sum x_{i2}x_{i1} & \sum x_{i2}^2 & \cdots & \sum x_{i2}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ik}x_{i1} & \sum x_{ik}x_{i2} & \cdots & \sum x_{ik}^2 \end{bmatrix}.$$

If the first column of X is $x_{i1} = 1$ for all i, then equation (1:1) becomes

$$y_i = \beta_1 + x_{i2}\beta_2 + \cdots + x_{ik}\beta_k + \varepsilon_i$$

a linear model with a constant term, β_1. In this case,

$$X^{\top}X = \begin{bmatrix} n & \sum x_{i2} & \sum x_{i3} & \cdots & \sum x_{ik} \\ \sum x_{i2} & \sum x_{i2}^2 & \sum x_{i2}x_{i3} & \cdots & \sum x_{i2}x_{ik} \\ \sum x_{i3} & \sum x_{i3}x_{i2} & \sum x_{i3}^2 & \cdots & \sum x_{i3}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ik} & \sum x_{ik}x_{i2} & \sum x_{ik}x_{i3} & \cdots & \sum x_{ik}^2 \end{bmatrix}.$$

which is clearly symmetric.

If A and B are square matrices of order n, and $AB = I_n$, then B is said to be the inverse of A, and A is the inverse of B: $B = A^{-1}$ and $A = B^{-1}$. When the inverse of A exists, A is said to be invertible. An invertible matrix is also said to be non-singular. If A and B are square matrices of the same order and AB is invertible, $(AB)^{-1} = B^{-1}A^{-1}$ for then

$$AB(AB)^{-1} = ABB^{-1}A^{-1} = I_n,$$

$$(AB)^{-1}AB = B^{-1}A^{-1}ab = I_n.$$

1.3 Some Special Types of Matrices

A square matrix $A = [a_{ij}]$ is said to be diagonal if $a_{ij} = 0$ for $i \neq j$ and some or all of the diagonal elements are non-zero. If all elements above the diagonal are zero and some or all elements on and below the diagonal are non-zero, then the matrix is said to be lower triangular. It is then clear how an upper triangular matrix is constructed.

A square matrix, A, is idempotent if $A = A^2$. An idempotent matrix is often called a projection matrix for reasons that will become obvious. If an idempotent matrix is also symmetric, $A = A^> = A^2$, then it is said to be an orthogonal projection matrix. Projection matrices play a crucial role in the statistical analysis of linear models like (1:4).

If x is a vector and A is square such that $x^>Ax = x^>A^>x > 0 \quad 8x \neq 0$, then A is said to be positive de¯nite (pd); if $x^>Ax = x^>A^>x ¸ 0 \quad 8x \neq 0$, then A is non-negative de¯nite (nnd) or positive semi-de¯nite. If $¡ x^>Ax = ¡ x^>A^>x > 0 \quad 8x \neq 0$, then A is negative de¯nite. If $A = A^> = A^2$, then A is nnd unless it is pd, whereupon $A = I$.

1.4 Orthogonal Matrices and Eigenvalues

An orthogonal matrix, M, is square and invertible such that $M^{¡1} = M^>$. If A is a real, square symmetric matrix, then there exists an orthogonal matrix, M, such that

$$M^>AM = ¤ \tag{1.6}$$

in which ¤ is diagonal having diagonal elements $¸_i$ (say, $i = 1; 2; ¢¢¢; n$). If A is invertible, then all the $¸_i$ will be non-zero. If the largest square sub-matrix of A that is invertible has order $r < n$, then r of the $¸_i$ will be non-zero and the remaining $(n ¡ r)$ will be zero. The non-zero elements $¸_i$ are called eigenvalues or characteristic

roots or proper values. Notice that

$$M^{>}AM = ¤ \quad , \quad AM = M¤. \tag{1.7}$$

Let the ¯rst column of M be m. Then, if ¸ is the ¯rst eigenvalue of ¤, (1:7) implies Am = m¸ and hence

$$(A ¡ I¸) m = 0. \tag{1.8}$$

Equation (1:8) will hold for any column, m, of M choosing the appropriate ¸ from the diagonal elements of ¤. Thus (1:8) is quite general in the sense that for each ¸, there will be a corresponding m. Indeed, the solution to m in (1:8) is called the eigenvector corresponding to the eigenvalue (or characteristic or proper value) ¸. More on this later.

If A is pd, it is invertible and all eigenvalues in ¤ will be positive. Taking the reciprocal of the square root of each and forming a corresponding diagonal matrix, say $¤^{¡\frac{1}{2}}$, it follows from (1:6) that

$$¤^{¡\frac{1}{2}}M^{>}AM¤^{¡\frac{1}{2}} = ¤^{¡\frac{1}{2}}¤ \; ¤^{¡\frac{1}{2}} = I.$$

Putting $M¤^{¡\frac{1}{2}} = Q$, this last equation may then be written as

$$Q^{>}AQ = I. \tag{1.9}$$

Equation (1:9) may be regarded as a de¯nition of a pd matrix and, since M and ¤ are invertible, Q is also invertible.

1.5 A Special Matrix Product

The Kronecker product of two matrices, A and B, is written $A \otimes B$ and is defined by

$$A \otimes B = [a_{ij}B].$$

If A is $(n \times m)$ and B is $(p \times q)$, then $A \otimes B$ is $(np \times mq)$. The following are properties of the Kronecker product:

$$(A \otimes B)(C \otimes D) = AC \otimes BD,$$

whereupon A and C and B and D must be conformable pairs,

$$A \otimes (B + C) = (A \otimes B) + (A \otimes C),$$

$$(A \otimes B) = (A \otimes I)(I \otimes B);$$

when A and B are invertible,

$$[A \otimes B]^{-1} = \left(A^{-1} \otimes B^{-1}\right).$$

1.6 Scalar Functions of Square Matrices

The trace of a square matrix, A, is the sum of its diagonal elements. This is written:

$$\operatorname{tr} A = \sum a_{ii}.$$

If A, B and C form a square matrix product, for example A is $(n \times p)$, B is $(p \times r)$ and C is $(r \times n)$ so that ABC is $(n \times n)$, then $\operatorname{tr} ABC = \operatorname{tr} BCA = \operatorname{tr} CAB$. In regard

to a Kronecker product of square matrices, tr (A - B) = tr A tr B. Quite clearly, the trace of a square matrix is a scalar.

The determinant of a square matrix may be calculated in a variety of equivalent ways. The classical expansion is based upon the (2 £ 2) case. If

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

then the determinant of A, written det A or | A |, is given by

$$\det A = a_{11}a_{22} - a_{12}a_{21}.$$

Then if A is (3 £ 3)

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

which equals

$$= a_{11}\mathring{A}_{11} - a_{12}\mathring{A}_{12} + a_{13}\mathring{A}_{13} \tag{1.10}$$

in which $\mathring{A}_{ij}$ equals the determinant obtained by eliminating row i and column j (i; j = 1; 2; 3). If now

$$A_{ij} = (-1)^{i+j}\mathring{A}_{ij} \tag{1.11}$$

then (1:11) allows (1:10) to be written as:

$$\det A = \sum_j a_{1j} A_{1j}. \tag{1.12}$$

Equation (1:12) is a way of calculating the determinant of A by expansion of the first row. Equally, the same result would be obtained by expanding the second row or column and following the same rules. In the present (3 × 3) case,

$$\begin{aligned} \det A = & -a_{21}(a_{12}a_{33} - a_{32}a_{13}) \\ & + a_{22}(a_{11}a_{33} - a_{31}a_{13}) \\ & - a_{23}(a_{11}a_{32} - a_{31}a_{12}) \end{aligned} \tag{1.13}$$

which reduced to (1:10). Generally, let $A = [a_{ij}]$ and $(-1)^{i+j}\mathring{A}_{ij} = A_{ij}$ where $\mathring{A}_{ij}$ is the determinant formed by deleting the i'th row and the j'th column of A $(i, j = 1, 2, \cdots, n)$, then

$$\begin{aligned} \det A &= \sum_j a_{ij} A_{ij} \text{ for any fixed } i \\ &= \sum_i a_{ij} A_{ij} \text{ for any fixed } j. \end{aligned}$$

For any two conformable square matrices, A and B, $\det AB = \det A \times \det B$. If two columns (rows) of A are interchanged, $\det A$ merely changes sign. If a proportion of one row (column) is added to or subtracted from another, then the determinant remains unchanged. This last property is a reflection of the role played by linear dependence in determining the value of a determinant. Let the columns of A be

$a_{:1}; a_{:2}; \cdots; a_{:m}$. Then these columns are said to be linearly independent if for scalars $\alpha_1; \alpha_2; \cdots; \alpha_m$ the equation

$$\alpha_1 a_{:1} + \alpha_2 a_{:2} + \cdots + \alpha_m a_{:m} = 0 \tag{1.14}$$

implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_m = 0$. If, on the contrary, equation (1:14) implies that not all α_i are zero, then $a_{:1}; a_{:2}; \cdots; a_{:m}$ are said to be linearly dependent. Assuming linear dependence, say of the form $\alpha_1; \alpha_2; \alpha_3$ are non-zero, $\alpha_4; \alpha_5; \cdots; \alpha_m$ are each zero, then

$$\begin{aligned} a_{:1} &= -\frac{\alpha_2}{\alpha_1} a_{:2} - \frac{\alpha_3}{\alpha_1} a_{:3} \\ &= \beta_2 a_{:2} + \beta_3 a_{:3} \end{aligned}$$

in an obvious notation. In this case, taking β_2 times column two from column one, and then β_3 times column three from the remainder yields:

$$\det A = \det \begin{bmatrix} 0 & a_{:2} & a_{:3} & \cdots & a_{:m} \end{bmatrix}.$$

Expanding by the first column then yields $\det A = 0$. Thus a square matrix containing linearly dependent columns (rows) has determinant equal to zero.

1.7 The Rank of a Matrix

A square matrix whose determinant is zero is said to be singular. Non-singular matrices are invertible; singular matrices are not invertible. For any matrix, square or

rectangular, square matrices may be formed by eliminating rows and columns and then corresponding determinants may be calculated. If the largest non-vanishing determinant, formed of the elements of a matrix by eliminating rows and columns, is of order r, then the matrix is said to have rank r. If the matrix is $(n \times n)$ of rank r, then $r \leq n$ and the matrix has r linearly independent rows and columns, and the remaining $(n - r)$ rows and columns respectively are linearly dependent on them. If the matrix is rectangular of order $(n \times m)$ of rank r, then $r \leq \min\{n, m\}$. As an example, consider the $(n \times k)$ matrix X of equation (1:4) with $k < n$. This matrix can never have rank greater than k; normally its rank is k, in which case X is said to be of full rank, i.e. its rank is as large as it can be.

1.8 Determinants as the Product of Eigenvalues

In regard to square, symmetric matrices like the matrix A in equation (1:6), notice that $\det A = \det\left[M^{\top} A M\right]$ because M is an orthogonal matrix implying that $M^{\top} = M^{-1}$ and

$$\begin{aligned} \det\left[M^{\top} A M\right] &= \det M^{\top} \det A \det M \\ &= \det^{-1} M \det A \det M. \end{aligned} \tag{1.15}$$

Now the determinant of any square matrix is the same as the determinant of its transpose; and the determinant of the inverse of a matrix is equal to the reciprocal of the determinant of the matrix itself. Thus (1:15) reduces to $\det A$. Since also

$\det\left[M^{>}AM\right] = \det\Lambda = \prod_{i=1}^{n}\lambda_i$, the product of the eigenvalues of A, the determinant of A is seen to be the product of its eigenvalues.

When A is square but not symmetric, the same rule (that det A is the product of its eigenvalues) holds, but for a different reason: there will exist a non-singular (but not necessarily real) matrix Q, such that

$$Q^{-1}AQ = T, \tag{1.16}$$

T being an upper triangular (not necessarily real) matrix whose diagonal elements are the eigenvalues, λ_i, of A. Since the determinant of an upper (or lower) triangular matrix is simply the product of its diagonal elements, $\det A = \det T = \prod_{i=1}^{n}\lambda_i$ as before.

1.9 Determinants as Areas and Volumes

An important application of determinants, which plays a vital role in multiple integration theory and hence in multivariate probability, is the evaluation of areas of parallelograms in R^2 and of hypervolumes of n-dimensional parallelepipeds in R^n. An example in R^2 is given in Figure 1. Here, interest centres on the area of the parallelogram defined by the sum of the vectors z and w, that is the area OABC. The sum, w + z, is represented by $\vec{OB}$. The area of the parallelogram OABC is the same as the area of the rectangle OHGC defined by the sum of the vectors r and w. The vector r has been constructed so that r and w are orthogonal, or at right angles to one

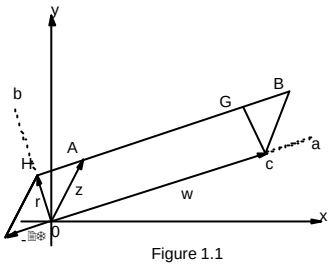

Figure 1.1

Figure 1:

another. When vectors are orthogonal, their scalar product is zero. Thus $r = z - w\mu$ for some positive fraction μ and hence

$$w^{>}(z - w\mu) = 0$$

implies

$$w^{>}z = w^{>}w\mu$$

or

$$w\mu = w\left(w^{>}w\right)^{-1}w^{>}z.$$

Setting $w\left(w^{>}w\right)^{-1}w^{>} = P$, it follows that $P = P^{>} = P^2$, so P is an orthogonal projection matrix. Thus

$$r = (I_2 - P)z.$$

Recall, first, that the matrix formed by w and z, say A, has a determinant whose

value is unaffected by replacing z by $(z - w\mu) = (I_2 - P)z$:

$$\det A = \det[w : z] = \det[w : z - w\mu], \tag{1.17}$$

because a determinant is unaffected by taking away from any one column a fixed proportion of another. Recall also that while the columns of A are not orthogonal, w and $(I_2 - P)z$ are orthogonal by construction. Now the area of interest is OHGC and this is obviously $\|w\| \cdot \|r\|$ or $\|w\| \cdot \|(I_2 - P)z\|$. Writing $B = [w : (I_2 - P)z]$,

$$|A| = \det[w : (I_2 - P)z] = |B|$$

and

$$\begin{aligned} |B| &= \det^{\frac{1}{2}}\left[B^{\prime}B\right] \\ &= \det^{\frac{1}{2}}\begin{bmatrix} w^{\prime}w & 0 \\ 0 & z^{\prime}(I_2 - P)z \end{bmatrix}, \end{aligned}$$

because $(I_2 - P) = (I_2 - P)^{\prime} = (I_2 - P)^2$. Thus

$$\begin{aligned} |A| &= \sqrt{\{w^{\prime}w : z^{\prime}(I_2 - P)z\}} \\ &= \|w\| \cdot \|(I_2 - P)z\| \\ &= \|w\| \cdot \|r\|. \end{aligned}$$

Thus the area of OABC equals the area of OHGC equals $\|w\| \cdot \|r\|$, the determinant of the matrix comprising the co-ordinates of the vectors w and z.

The same result could have been obtained by applying equation (1:16), or (1:6) if appropriate. This would be equivalent to forming the vector r and then rotating the axes (x; y) about 0 to become the axes (a; b). When this is done, w and r respectively have co-ordinate vectors of the form

$$\begin{bmatrix} a_1 \\ 0 \end{bmatrix} \text{ and } \begin{bmatrix} 0 \\ b_2 \end{bmatrix}$$

in the new co-ordinate system. That these vectors would have these co-ordinates in the system (a; b) is obvious, since w lies on the a-axis and r lies on the b-axis.

As an example, let the matrix A be given by:

$$A = \begin{bmatrix} 6 & 2 \\ 2 & 3 \end{bmatrix} \Rightarrow |A| = 14.$$

The matrix A is represented in ¯gure 1 by the two vectors w and z:

$$w = \begin{bmatrix} 6 \\ 2 \end{bmatrix}; \quad z = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

yielding A as speci¯ed below. The vector r may be determined from $r = z - w\mu$ and $w^{\top} r = 0$. Hence $w^{\top} r = 6(2 - 6\mu) + 2(3 - 2\mu) = 0$. Solving for μ yields:

$$r = \begin{bmatrix} -0.7 \\ 2.1 \end{bmatrix}.$$

It then follows that $\|w\|^2 \|r\|^2 = (w^{\top} w)(r^{\top} r) = (36 + 4)(0.7^2 + 2.1^2) = 196$. Thus $\|w\| \|r\| = \sqrt{196} = 14$, as already determined for $|A|$. Taking the alternative route,

$r = (I_2 - P)z$ in which

$$P = w(w^\top w)^{-1}w^\top$$
$$= \begin{bmatrix} 0.9 & 0.3 \\ 0.3 & 0.1 \end{bmatrix}.$$

Thus

$$(I_2 - P)z = \begin{bmatrix} 0.1 & -0.3 \\ -0.3 & 0.9 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$
$$= \begin{bmatrix} -0.7 \\ 2.1 \end{bmatrix},$$

as before.

1.10 Vectorization

It is occasionally useful to transform a matrix into a vector. As will become evident in chapter two, the axioms of a vector space, which define what may be regarded as a vector, apply also to matrices. Therefore it is quite legitimate to regard a matrix as a vector, in the sense that a matrix is an element of a vector space.

If A is an $(n \times m)$ matrix of m columns $a_{:j} \in R^n$ and n rows $a_{i:} \in R^m$ for

$i = 1; 2; ¢¢¢; n$ and $j = 1; 2; ¢¢¢; m$, then

$$\text{vec}A = \begin{bmatrix} a_{:1} \\ a_{:2} \\ \vdots \\ a_{:m} \end{bmatrix} \in R^{nm}$$

and

$$\text{vec}A^{>} = \begin{bmatrix} a_1^{>}: \\ a_2^{>}: \\ \vdots \\ a_n^{>}: \end{bmatrix} \in R^{mn}.$$

Of course, (vecA) is $(1 £ nm)$ which cannot be equal to $\text{vec}A^{>}$ which is $(mn £ 1)$. Also

$$\text{vec}A^{>} = \begin{bmatrix} a_1:; & a_2:; & ¢¢¢; & a_n: \end{bmatrix}$$

whereas

$$(\text{vec}A)^{>} = \begin{bmatrix} a_1^{>}:; & a_2^{>}:; & ¢¢¢; & a_m^{>}: \end{bmatrix}.$$

An important result for later application is the vectorization of a real matrix product. If A, B and C are respectively $(n £ m)$, $(m £ p)$ and $(p £ k)$ real matrices, their matrix product is of order $(n £ k)$ and hence vec(ABC) must be an nk-tuple.

This nk-tuple is obtained from

$$\text{vec}(ABC) = \left(C^{\top} \otimes A\right)\text{vec}B.$$

The matrix $C^{\top} \otimes A$ is of order $(kn \times pm)$ and vecB is $(mp \times 1)$, thus vec(ABC) must be $(kn \times 1)$ or an nk-tuple. In the case of a matrix product AB, this may be written $AB = ABI_p = I_nAB$ whereupon, applying the rule,

$$\begin{aligned} \text{vec}(ABI_p) &= [I_p \otimes A]\text{vec}B \\ \text{vec}(I_nAB) &= \left(B^{\top} \otimes I_n\right)\text{vec}A \end{aligned}$$

it must clearly be the case that

$$\begin{aligned} \text{vec}(AB) &= [I_p \otimes A]\text{vec}B \\ &= \left(B^{\top} \otimes I_n\right)\text{vec}A. \end{aligned}$$

For the matrix product ABCD,

$$\begin{aligned} \text{vec}(ABCD) &= \left(D^{\top} \otimes AB\right)\text{vec}C \\ &= \left(D^{\top} \otimes A\right)\text{vec}(BC). \end{aligned}$$

1.11 An Important Partitioned Inverse

Let A and C be symmetric, non-singular matrices of order $(m \times m)$ and $(k \times k)$, and let B be an $(m \times k)$ matrix of rank k. Using these matrices, let the $(mk \times mk)$

symmetric matrix D be de¯ned by

$$D = \begin{bmatrix} A & B \\ B^{>} & C \end{bmatrix}. \tag{1.18}$$

It is taken that all the inverses in expressions to be developed below exist.

Before proceeding, it should be made explicit that matrices that are partitioned, like D, may be treated as if the sub-matrices are elements. For example, if

$$E = \begin{bmatrix} P & Q \\ R & S \end{bmatrix}$$

then, provided the products exist

$$DE = \begin{bmatrix} A & B \\ B^{>} & C \end{bmatrix} \begin{bmatrix} P & Q \\ R & S \end{bmatrix} = \begin{bmatrix} AP + BR & AQ + BS \\ B^{>}P + CR & B^{>}Q + CS \end{bmatrix}.$$

Similarly

$$E^{>} = \begin{bmatrix} P^{>} & R^{>} \\ Q^{>} & S^{>} \end{bmatrix}.$$

Returning to D of (1:18), a formula for $D^{i\,1}$ is

$$D^{i\,1} = \begin{bmatrix} {\scriptstyle i}A \,{\scriptstyle ¡}\, BC^{i\,1}B^{>}{\scriptstyle ¢}^{i\,1} & {\scriptstyle i} {\scriptstyle i}A \,{\scriptstyle ¡}\, BC^{i\,1}B^{>}{\scriptstyle ¢}^{i\,1}BC^{i\,1} \\ {\scriptstyle ¡}\, {\scriptstyle i}C \,{\scriptstyle ¡}\, B^{>}A^{i\,1}B{\scriptstyle ¢}^{i\,1}B^{>}A^{i\,1} & {\scriptstyle i}C \,{\scriptstyle ¡}\, B^{>}A^{i\,1}B{\scriptstyle ¢}^{i\,1} \end{bmatrix} \tag{1.19}$$

In (1:19) it should be noticed that, since D is symmetric, so is $D^{i\,1}$ and hence

$${\scriptstyle i}A \,{\scriptstyle ¡}\, BC^{i\,1}B^{>}{\scriptstyle ¢}^{i\,1}BC^{i\,1} = {\scriptstyle n}{\scriptstyle i}C \,{\scriptstyle ¡}\, B^{>}A^{i\,1}B{\scriptstyle ¢}^{i\,1}B^{>}A^{i\,1}{\scriptstyle o}^{>}$$

or

$$\left(A - BC^{-1}B^{>}\right)^{-1}BC^{-1} = A^{-1}B\left(C - B^{>}A^{-1}B\right)^{-1}. \tag{1.20}$$

That D^{-1} in (1:19) is indeed the inverse of D is easily checked by post-multiplying D^{-1} by D. This yields

$$\begin{aligned} & \left(A - BC^{-1}B^{>}\right)^{-1}A - \left(A - BC^{-1}B^{>}\right)^{-1}BC^{-1}B^{>} \qquad (1.21)\\ = \; & \left(A - BC^{-1}B^{>}\right)^{-1}\left(A - BC^{-1}B^{>}\right) \\ = \; & I_m \end{aligned}$$

$$\left(A - BC^{-1}B^{>}\right)^{-1}B - \left(A - BC^{-1}B^{>}\right)^{-1}BC^{-1}C = 0 \tag{1.22}$$

$$-\left(C - B^{>}A^{-1}B\right)^{-1}B^{>}A^{-1}A + \left(C - B^{>}A^{-1}B\right)^{-1}B^{>} = 0 \tag{1.23}$$

$$\begin{aligned} & -\left(C - B^{>}A^{-1}B\right)^{-1}B^{>}A^{-1}B + \left(C - B^{>}A^{-1}B\right)^{-1}C \qquad (1.24)\\ = \; & \left(C - B^{>}A^{-1}B\right)^{-1}\left(C - B^{>}A^{-1}B\right) \\ = \; & I_k \end{aligned}$$

1.12 Useful Inverses and Projection Matrices

Re-writing equation (1:21) using equation (1:20) yields

$$\left(A - BC^{-1}B^{>}\right)^{-1}A - A^{-1}B\left(C - B^{>}A^{-1}B\right)^{-1}B^{>} = I_m:$$

Post-multiplying throughout by A^{-1}, there results

$$\left(A - BC^{-1}B'\right)^{-1} = A^{-1} + A^{-1}B\left(C - B'A^{-1}B\right)^{-1}B'A^{-1}. \qquad (1.25)$$

Equation (1:25) finds application in the random-coefficients model associated with the linear model (1:4), i.e. model (1:4) augmented by the relation

$$\gamma = A\delta + \alpha \qquad (1.26)$$

in which A is a $(k \times m)$ $(m \leq k)$ known matrix of rank m, δ is an m-tuple of coefficients and α is a $(k \times 1)$ vector of errors, independent of ε in (1:4). Associated with α is an $(m \times n)$ pd matrix Ψ. Setting

$$\begin{bmatrix} \Sigma & X \\ X' & -\Psi^{-1} \end{bmatrix} = \begin{bmatrix} A & B \\ B' & C \end{bmatrix}$$

and applying (1:25),

$$\left(\Sigma + X\Psi X'\right)^{-1} = \Sigma^{-1} - \Sigma^{-1}X\left(\Psi^{-1} + X'\Sigma^{-1}X\right)^{-1}X'\Sigma^{-1}. \qquad (1.27)$$

The matrix on the left-hand side of (1:27) is the inverse of the matrix associated with the composite error, $\varepsilon + X\alpha$, obtained by substituting (1:26) into (1:4).

Two applications of equation (1:19) arise in least squares estimation theory applied to model (1:4). First, it is often useful to have available a particular form of the pd

matrix $(X^{\top}X)^{-1}$ when $X = [X_1 \vdots X_2]$. Applying (1:19),

$$(X^{\top}X)^{-1} = \begin{bmatrix} X_1^{\top}X_1 & X_1^{\top}X_2 \\ X_2^{\top}X_1 & X_2^{\top}X_2 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} (X_1^{\top}M_2X_1)^{-1} & -(X_1^{\top}M_2X_1)^{-1}X_1^{\top}X_2(X_2^{\top}X_2)^{-1} \\ -(X_2^{\top}M_1X_2)^{-1}X_2^{\top}X_1(X_1^{\top}X_1)^{-1} & (X_2^{\top}M_1X_2)^{-1} \end{bmatrix}$$

in which $M_i = I_n - X_i(X_i^{\top}X_i)^{-1}X_i^{\top}$; $i = 1; 2$. Setting

$$P_i = X_i(X_i^{\top}X_i)^{-1}X_i^{\top},$$

notice that $P_i = P_i^{\top} = P_i^2$ and hence that $M_i = M_i^{\top} = M_i^2$. Thus P_i and M_i are orthogonal projection matrices. Also, $P_i^{\top}M_i = P_iM_i = 0$ and $P_i + M_i = I_n$. For any $y \in R^n$, $P_iy + M_iy = y$. Let $P_iy = u$ and $M_iy = v$. Then $u^{\top}v = v^{\top}u = 0$ and u and v are orthogonal to one another. Finally, the rank of an orthogonal projection matrix is equal to its trace and $\mathrm{tr}P_i = \mathrm{tr}\left[X_i(X_i^{\top}X_i)^{-1}X_i^{\top}\right] = \mathrm{tr}\left[(X_i^{\top}X_i)^{-1}X_i^{\top}X_i\right] = \mathrm{tr}I_k = k$ and $\mathrm{tr}M_i = \mathrm{tr}(I - P_i) = (n - k)$. Moreover, $k + (n - k) = n$.

A second application of equation (1:19) is in terms of evaluating the projection $X(X^{\top}X)^{-1}X^{\top}$ in terms of X_1 and X_2. Using the inverse (1:19),

$$X(X^{\top}X)^{-1}X^{\top} = X_1(X_1^{\top}M_2X_1)^{-1}X_1^{\top}M_2 + X_2(X_2^{\top}M_1X_2)^{-1}X_2^{\top}M_1. \qquad (1.28)$$

With

$$F_1 = X_1 \left(X_1^{>} M_2 X_1\right)^{¢_{i\,1}} X_1^{>} M_2$$

$$F_2 = X_2 \left(X_2^{>} M_1 X_2\right)^{¢_{i\,1}} X_2^{>} M_1$$

$$P = X \left(X^{>} X\right)^{¢_{i\,1}} X^{>},$$

equation (1:28) becomes

$$P = F_1 + F_2. \tag{1.29}$$

Here, $P = P^{>} = P^2$, $F_1 = F_1^2$ and $F_2 = F_2^2$ but neither F_1 nor F_2 is symmetric. While P is referred to as an orthogonal projection, F_1 and F_2 are referred to as oblique projections. In model (1:4)

$$y = X\bar{} + "$$

$$= X_1\bar{}_1 + X_2\bar{}_2 + "$$

where $\bar{} = \left(\bar{}_1^{>}, \bar{}_2^{>}\right)^{>}$. Py is the least squares estimator of $X\bar{}$, F_1y and F_2y are the corresponding least squares estimators of $X_1\bar{}_1$ and $X_2\bar{}_2$ respectively.

2 Vector Spaces

2.1 Introduction

The aim of chapter one was to review some of the theory of matrices from the classical viewpoint of manipulative algebra. Along with a number of specific definitions and properties of matrices, three points were noted:

1. A vector in R^n may be regarded as a special case of an $(n \times m)$ matrix simply by setting $m = 1$.

2. Any $(n \times m)$ matrix, A, defined over the field R, may be transformed into a $(nm \times 1)$ vector by the operation of vectorization. Thus, vecA is an $(nm \times 1)$ vector.

3. Let M be the set of all $(n \times m)$ matrices defined over R and let $\alpha \in R$. Then for any pair, A and B in M, $A + B \in M$ and $\alpha A \in M$.

Point 1 suggests that matrices are more general objects than vectors, but this seems to be contradicted by point 2, since any matrix may be written as a vector. As will become apparent, matrices and vectors are merely special examples of a class of objects which obey the property recognized in point 3. This property is fundamental to the definition of a linear or vector space.

It is the purpose of this chapter to introduce the theory of linear or vector spaces and to explain, in reasonable detail, a particular type of vector space with which there will be especial concern. This is a Euclidean space and such a space is the setting for the linear model outlined in equations (1:1){(1:4). A Euclidean space is a vector space endowed with a means of measuring the length of vectors and the angles between them. The means of measuring lengths and angles is the scalar product.

The first step is to characterize point 3 above in the form of a definition.

Definition 1 A set V of elements $x; y; z; ¢¢¢$ is said to form a vector space over the field F if:

(a) For every two elements x and y in V, there is associated an element z in V called the sum of x and y, written $x + y = z$.

(b) For every x in V and every $¸$ in F, there is an element $¸x$ in V called the product of x and $¸$.

These operators must obey the following axioms.

(1) Vector operations:

Commutativity: $x + y = y + x$

Associativity: $(x + y) + z = x + (y + z)$

Zero: V contains an element, 0,

such that $x + 0 = x$ $8x$ in V.

Negativity: For every x in V, there exists an

element $(¡ x)$ such that $x + (¡ x) = 0$.

(2) Scalar Multiplication:

of a vector: $1x = x$

$® (¯x) = ®¯ (x)$ $®; ¯ 2 F;$

of a sum of vectors $(® + ¯) x = ®x + ¯x$

or scalars $® (x + y) = ®x + ®y.$ ■

The theory of vector spaces is much more general than is suggested by the examples of R^n. Two examples illustrate the point.

Example 1 The set of all polynomials of degree not exceeding some natural number, n, constitutes a vector space, whereas the set of polynomials of degree exactly n does not. To see this, the following example su± ces:

$$\mathbf{i}_{t^3 + t}^{\mathbf{¢}} + \mathbf{i}_{¡ t^3 + t}^{\mathbf{¢}} = 2t$$

This is not a polynomial of degree $n = 3$, but it is a polynomial of degree not exceeding $n = 3$. ■

Example 2 The set of all real matrices of order $(n \times m)$, $A; B; C; \ldots$ constitutes a vector space since $A + B = C$ lies in the same set and so does αA where $\alpha \in R$. For example

$$\begin{bmatrix} 1 & 4 \\ 2 & 1 \\ 3 & 3 \end{bmatrix} + \begin{bmatrix} 6 & 1 \\ 2 & 1 \\ 6 & 1 \end{bmatrix} = \begin{bmatrix} 7 & 5 \\ 4 & 2 \\ 9 & 4 \end{bmatrix};$$

and

$$2\begin{bmatrix} 1 & 4 \\ 2 & 1 \\ 3 & 3 \end{bmatrix} = \begin{bmatrix} 2 & 8 \\ 4 & 2 \\ 6 & 6 \end{bmatrix}.$$

The answer in each case is a (3×2) matrix of real numbers. This example merely illustrates point 3 in the opening paragraph of this sections. ■

The elements of a vector space are generally called vectors. There is some confusion caused by this terminology in examples 1 and 2. Thus a matrix is an element of a vector space, namely the set of all real matrices of the same order. This implies that a matrix is a vector; and, in a formal sense, it is. Yet there will be no serious confusion so long as it is recognized that a matrix will be called a matrix and that such a matrix is an element of a vector space. As will become apparent below, vector spaces may be distinguished by their dimensionality. In the case of the vectors from R^n, R^n is said to be an n-dimensional vector space. In the case of $(n \times m)$ matrices of

real elements, these are elements of R^{nm}, notwithstanding the fact that such matrices may be thought of as m vectors drawn from R^n. Of course, if A is an $(n \pounds m)$ matrix, then $vecA \; 2 \; R^{nm}$. In respect of example 1, a terminology may be used corresponding to that of example 2. Thus polynomials of degree not exceeding n may be recognized as polynomials while also appreciating that such polynomials are elements of a vector space.

Vectors de¯ned over the ¯eld R are called real vectors; the spaces in which these vectors lie are called real vector spaces. In econometrics, real vector spaces are of great importance, especially real vector spaces endowed with a scalar product. Before introducing a scalar product, three important features of general vector spaces are introduced. These are dimensionality, basis and co-ordinates.

2.2 Dimensionality, Basis and Co-ordinates

If $x; y; z; ¢¢¢; w$ are vectors in a vector space and $®; ¯; °; ¢¢¢; µ$ is a corresponding set of scalars, then

$$®x + ¯y + °z + ¢¢¢ + µw$$

is said to be a linear combination of the vectors $x; y; z; ¢¢¢; w$ with coe± cients $®; ¯; °; ¢¢¢; µ$.

De¯nition 2 Let V be a vector space. The vectors $x; y; z; ¢¢¢; w$ in V are said to be linearly dependent if there exists a corresponding set of scalars $®; ¯; °; ¢¢¢; µ$, not

all zero, such that the linear combination

$$\alpha x + \beta y + \gamma z + \cdots + \mu w = 0. \tag{2.1}$$

The vectors $x; y; z; \cdots; w$ are said to be linearly independent if the equality

$$\alpha x + \beta y + \gamma z + \cdots + \mu w = 0$$

implies that $\alpha = \beta = \gamma = \cdots = \mu = 0$. ■

Let the vectors $x; y; z; \cdots; w$ be linearly dependent. Then there exists scalars such that equation (2:1) exists with at least one coefficient non-zero. Let $\alpha \neq 0$. Then equation (2:1) may be re-written

$$x = -\frac{\beta}{\alpha}y - \frac{\gamma}{\alpha}z - \cdots - \frac{\mu}{\alpha}w. \tag{2.2}$$

Putting $-\frac{\beta}{\alpha} = \lambda; -\frac{\gamma}{\alpha} = \tau; \cdots; -\frac{\mu}{\alpha} = \phi$, then

$$x = \lambda y + \tau z + \cdots + \phi w \tag{2.3}$$

and x may be expressed as a linear combination of the remaining vectors. Going the other way round, if equation (2:3) holds, then equation (2:1) will also hold by setting $\alpha = -1; \beta = \lambda; \gamma = \tau; \cdots; \mu = \phi$. Thus if a set of vectors is linearly dependent, then one vector can be expressed as a linear combination of the others and if one vector may be expressed as a linear combination of the others, then those vectors are linearly dependent.

Definition 3 A vector space V is said to be n-dimensional if it contains n linearly independent vectors and any $(n + 1)$ vectors in V are linearly dependent. ■

The idea of a basis follows directly from the idea of dimensionality.

Definition 4 Any set of n linearly independent vectors in an n-dimensional space V is called a basis of V. ■

From the definition of a basis comes the idea of co-ordinates. Co-ordinates are a product of the following theorem.

Theorem 1 Every vector in an n-dimensional vector space V can be represented as a linear combination of basis vectors.

Proof. Let $e_1; e_2; ¢¢¢; e_n$ be a basis of V. For any arbitrary x in V, $x; e_1; e_2; ¢¢¢; e_n$ represents a linearly dependent set of vectors. Therefore there will exist scalars $®_0; ®_1; ¢¢¢; ®_n$, not all zero, such that

$$®_0x + ®_1e_1 + ®_2e_2 + ¢¢¢+ ®_ne_n = 0.$$

Now $®_0 6= 0$ for if it were zero, then the basis vectors would be linearly dependent which they cannot be. Thus $®_0 6= 0$ and

$$x = ¡ \frac{®_1}{®_0}e_1 ¡ \frac{®_2}{®_0}e_2 ¡ ::: ¡ \frac{®_n}{®_0}e_n.$$

This proves that x may be expressed as a linear combination of the basis vectors. To prove uniqueness, let there be two linear combinations of the same basis vectors which equate with x:

$$x = \lambda_1 e_1 + \lambda_2 e_2 + \cdots + \lambda_n e_n$$

and

$$x = \mu_1 e_1 + \mu_2 e_2 + \cdots + \mu_n e_n.$$

Subtracting, it follows that

$$0 = (\lambda_1 - \mu_1) e_1 + (\lambda_2 - \mu_2) e_2 + \cdots + (\lambda_n - \mu_n) e_n.$$

Since $e_1, e_2, \cdots, e_n$ are linearly independent vectors, it follows that $(\lambda_1 - \mu_1) = (\lambda_2 - \mu_2) = \cdots = (\lambda_n - \mu_n) = 0$, implying that $\lambda_1 = \mu_1, \lambda_2 = \mu_2, \cdots, \lambda_n = \mu_n$. Thus, the representation is unique. ■

In theorem 1, it may be taken that, relative to the basis $e_1, e_2, \cdots, e_n$, x is determined by the unique set of coefficients $\lambda_1, \lambda_2, \cdots, \lambda_n$. These coefficients are called the co-ordinates of x relative to the basis $e_1, e_2, \cdots, e_n$. This basis may be defined as an $(n \times n)$ matrix since each e_i is an $(n \times 1)$ vector. Let this matrix, called a basis matrix, be defined as

$$[e_1, e_2, \cdots, e_n] = E,$$

and let

$$\begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix} = \xi$$

be the co-ordinate vector relative to E. Thus

$$x = E\xi.$$

Let another basis of V be $[g_1; g_2; \cdots; g_n] = G$ which again is an $(n \times n)$ matrix, and let the co-ordinates of x relative to the basis matrix G be μ, another $(n \times 1)$ co-ordinate vector of components $\mu_1; \mu_2; \cdots; \mu_n$:

$$x = G\mu.$$

The two bases must be linearly dependent, because they together represent 2n vectors of an n-dimensional vector space. Thus each g_i may be represented uniquely as a linear combination of the basis vectors comprising E, i.e.

$$g_i = E\bar{\ }_i \quad \forall i = 1; 2; \cdots; n \tag{2.4}$$

in which $\bar{\ }_i$ is an $(n \times 1)$ vector of co-ordinates. Equation (2:4) may be expressed as

$$[g_1; g_2; \cdots; g_n] = E[\bar{\ }_1; \bar{\ }_2; \cdots; \bar{\ }_n] \tag{2.5}$$

whereupon, writing $B = [\beta_1; \beta_2; \cdots; \beta_n]$, since B is another $(n \times n)$ matrix and

$$G = EB: \tag{2.6}$$

It follows from the representation of x that

$$x = E\xi = G\mu = EB\mu \tag{2.7}$$

and immediately $\xi = B\mu$: B must be invertible because equation (2:6) indicates that the rank of G (which must be n since it has n linearly independent columns, by definition) is equal to the rank of EB. But E also has rank n, since it is a basis matrix. Hence, B also has rank n. Therefore, $\mu = B^{-1}\xi$. If the matrix B (whose columns represent the co-ordinates of the basis matrix G relative to the basis matrix E) is known, then G = EB and the co-ordinates, μ, of x relative to the basis G are also known from the co-ordinates, ξ, of x relative to the basis E via $\mu = B^{-1}\xi$.

Finally note that if $x = E\xi$ and $y = E\mu$, then $x + y = E\xi + E\mu = E(\xi + \mu)$, i.e. given the same basis, addition of vectors corresponds to the addition of their co-ordinates.

2.3 Euclidean Space

A Euclidean space is a real vector space endowed with a scalar product, i.e. a mapping $R^n \to R$. The purpose of a scalar (or inner) product is to provide the means to measure length and angle in the space. Vectors have length and direction and these ideas may be given numerical expression once the scalar product is defined.

The natural scalar product for any two vectors x and y in R^n is

$$x^{>}y = \sum_{i} x_i y_i \tag{2.8}$$

for all i = 1; 2; ¢¢¢; n, the x_i and y_i being the co-ordinates of x and y relative to the basis $[e_1; e_2; ¢¢¢; e_n] = I_n$. Notice the following properties of $x^{>}y$ for every pair of vectors in R^n:

i) $x^{>}y = y^{>}x$;

ii) $(¸x)^{>}y = ¸(x^{>}y)$; $¸ \in R$;

iii) $(x + z)^{>}y = x^{>}y + x^{>}z$;

iv) $x^{>}x ¸ 0$ with equality only if x = 0.

The scalar product $x^{>}y$ is just a special case of $x^{>}Ay$ where $A = I_n$. A more general scalar product would be

$$x^{>}Ay = \sum_{i} \sum_{j} x_i a_{ij} y_j$$

for i; j = 1; 2; ¢¢¢; n, subject to the following conditions corresponding to (i){(iv) immediately above for every pair of vectors in R^n:

i) $x^{>}Ay = y^{>}Ax$;

ii) $(¸x)^{>}Ay = ¸ \, i x^{>}Ay ¢$; $¸ \in R$

iii) $(x + z)^{>}Ay = x^{>}Ay + x^{>}Az$;

iv) $x^{>}Ax ¸ 0$ with equality only if x = 0.

Condition (i) implies that $y^{>}A^{>}x = y^{>}Ax$ and hence that A is a symmetric matrix. Condition (iv) implies that A is a pd matrix. This last condition is required to measure length as $+\sqrt{(x^{>}Ax)}$ and so $x^{>}Ax$ must be a pd quadratic form for a real square root to exist. In practice, A may be chosen as a matter of convenience.

Because there are many possibilities for de¯ning a scalar product, a more general notation is required. Usually, this takes the form $(x; y)$ or $hx; yi$; sometimes $[x; y]$. To indicate a particular product, for example the one that takes the form $x^{>}Ax$ above, it would be usual to de¯ne this relative to $x^{>}y$ in the following notation: $(:; :)$ is the natural scalar and $h:; :i = (:; A:)$. A general de¯nition of a scalar product is now set out as

De¯nition 5 If for every pair of vectors $x; y$ in R^n there is associated a real number $(x; y)$ such that

i) $(x; y) = (y; x)$

ii) $(¸x; y) = ¸(x; y); \quad ¸ 2 R$

iii) $(x_1 + x_2; y) = (x_1; y) + (x_2; y)$

iv) $(x; x) ¸ 0$ with equality only when $x = 0$,

then $(x; y)$ is called the scalar product de¯ned on R^n and R^n is then referred to as an n-dimensional Euclidean space.

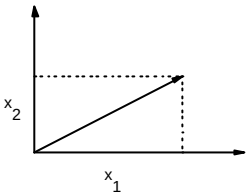

Figure 2.1

Figure 2:

Definition 6 Let x be a vector in Euclidean space in which the scalar product $(\cdot;\cdot)$ is defined. By the length of x is meant the positive number $\|x\| = \sqrt{(x;x)}$, unless $x = 0$, whereupon $\|x\| = 0$.

Example 3 In 2-dimensional Euclidean space, the vector x has co-ordinates x_1 and x_2 as indicated in figure 2.1. The scalar product defined on this space is $(x;y) = x^{\top}y$. Thus

$$x = \begin{matrix} x_1 \\ x_2 \end{matrix}$$

and

$$\|x\|^2 = x^{\top}x = x_1^2 + x_2^2.$$

It follows that $\|x\| = \sqrt{(x_1^2 + x_2^2)}$, which is the same length as according to Pythagoras' Theorem. ■

Let x and y be two non-zero vectors in R^n on which the scalar product $(\cdot;\cdot)$ is

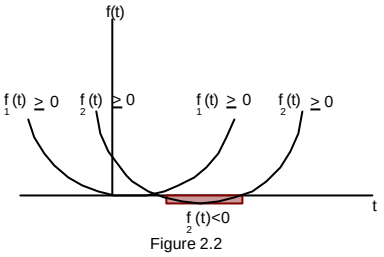

Figure 2.2

Figure 3:

de¯ned. Let t be a real number. Axiom 4 of de¯nition 5 requires

$$(x ¡ ty; x ¡ ty) ¸ 0$$

with equality only if x = ty. Consequently

$$(x; x) ¡ t(x; y) ¡ t(y; x) + t^2(y; y) ¸ 0.$$

From axiom 1 of de¯nition 5, this last expression becomes

$$(y; y)t^2 ¡ 2(x; y)t + (x; x) = f(t) ¸ 0. \qquad (2.9)$$

Thus the value of t must be such that f(t), a quadratic form in t, is always positive or zero, and never negative; thus f(t) may be like $f_1(t)$ or $f_3(t)$ in ¯gure 2, but not like $f_2(t)$. The quadratic f(t), therefore, can never cross the t-axis (diagram 2.2) i.e. f(t) can never have two distinct real roots. If f(t) cannot have two distinct real roots, then in equation (2:9) it must have either two coincident roots or imaginary roots. Hence

$$4(x; y)^2 · 4(y; y)(x; x) \qquad (2.10)$$

i.e. in $ax^2 + bx + c$; $b^2 = 4ac$ when coincident roots emerge, or $b^2 < 4ac$ and hence the roots are imaginary. Thus from equation (2:10)

$$(x; y)^2 \cdot (x; x)\ (y; y). \tag{2.11}$$

Inequality (2:11) is known as the Cauchy-Schwartz Inequality. It implies

$$¡ 1 \cdot \frac{(x; y)}{kxk\ kyk} \cdot 1. \tag{2.12}$$

For this reason, it is possible to state

De¯nition 7 By the angle μ between two non-zero vectors in n-dimensional Euclidean space is meant

$$\mu = \cos^{¡1}\left(\frac{(x; y)}{kxk\ kyk}\right). \blacksquare$$

The bounds expressed by equation (2:12) are the bounds of the cosine function and this is the justi¯cation for using De¯nition 6. It is helpful, nevertheless, to explore the geometry in R^2 that lies behind De¯nition 6.

Example 4 Consider x and y in R^n on which the natural scalar product $x^>y = (x; y)$ is de¯ned. The notation is described in ¯gure 2.3.

The basis in the space is

$$[e_1\ e_2] = \begin{matrix} 2 & 3 \\ 6 & 7 \\ 4 & 5 \end{matrix} \begin{matrix} \\ 1\ \ 0 \\ \\ 0\ \ 1 \end{matrix} = I_2.$$

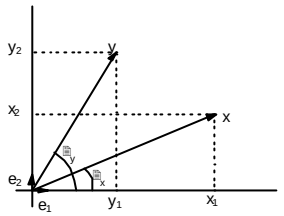

Figure 2.3

Figure 4:

Relative to the basis

$$x = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

and

$$y = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

The lengths of x and y are quite natural in view of Pythagoras' Theorem

$$\|x\|^2 = x_1^2 + x_2^2 \ , \ \|x\| = \sqrt{(x_1^2 + x_2^2)},$$
$$\|y\|^2 = y_1^2 + y_2^2 \ , \ \|y\| = \sqrt{(y_1^2 + y_2^2)}.$$

Moreover

$$\sin\mu_x = \frac{x_2}{\|x\|}; \cos\mu_x = \frac{x_1}{\|x\|};$$
$$\sin\mu_y = \frac{y_2}{\|y\|}; \cos\mu_y = \frac{y_1}{\|y\|}.$$

Now interest centres on the angle between x and y, that is $\mu = \mu_y - \mu_x$ and $\cos(\mu_y -$

$\mu_x) = \cos\mu_y \cos\mu_x + \sin\mu_y \sin\mu_x$: Therefore

$$\cos\mu = \frac{x_1y_1 + x_2y_2}{\|x\| \|y\|}$$
$$= \frac{x^{>}y}{\|x\| \|y\|}.$$

It follows that the Cauchy-Schwartz inequality is vindicated in this case and that

$$x^{>}y = \|x\| \|y\| \cos\mu.$$

When $\mu = \frac{¼}{2} = 90^{±}$, $\cos\mu = 0$ and hence $x^{>}y = 0$; x and y are said to be orthogonal. If $\mu = 0$ $\cos\mu = 1$ and if $\mu = ¼ = 180^{±}$, $\cos\mu = ¡ 1$: In these two cases x and y are co-linear. For example, when $\mu = 0$

$$x^{>}y = \|x\| \|y\|$$
$$= \|x\|^2 \frac{\|y\|}{\|x\|}$$
$$= x^{>}x\frac{\|y\|}{\|x\|}$$

Setting $¸ = \frac{\|y\|}{\|x\|}$, $¸ \in R$ and $x^{>}y = x^{>}(¸x)) y = ¸x$: If $\mu = ¼$; then $¸ < 0$ and y would be opposite in direction to x. ■

De¯nition 8 If for any pair of vectors x and y in Euclidean space

$$\|x + y\|^2 = \|x\|^2 + \|y\|^2,$$

then x and y are said to be orthogonal.

De¯nition 8 implies that (x; y) = 0 and is in fact a straightforward statement of Pythagoras' Theorem. Notice also that the statement holds for any scalar product. If for example h:; :i = (:; A:), then

$$kxk = \sqrt{(x^{>}Ax)}.$$

Moreover, if $\left\langle x^{>}Ay \right\rangle = 0$ then x and y are orthogonal relative to the scalar product h:; :i. Thus relative to h:; :i

$$\begin{aligned} kx + yk^2 &= (x + y)^{>} A (x + y) \\ &= x^{>}Ax + x^{>}Ay + y^{>}Ax + y^{>}Ay \\ &= x^{>}Ax + 2x^{>}Ay + y^{>}Ay \\ &= kxk^2 + kyk^2 + 2x^{>}Ay. \end{aligned}$$

Thus $kx + yk^2 = kxk^2 + kyk^2$ only when $x^{>}Ay = 0$, i.e.

$$hx; yi = (x; Ay) = 0.$$

De¯nition 9 In n-dimensional Euclidean space, the vectors $x_1; x_2; ¢¢¢; x_n$ are said to be pairwise orthogonal if

$$kx_1 + x_2 + ::: + x_nk^2 = kx_1k^2 + kx_2k^2 + ¢¢¢ + kx_nk^2.$$

If the vectors $x_1; x_2; ¢¢¢; x_n$ constitute a set of non-zero, pairwise orthogonal vectors in n-dimensional Euclidean space, these vectors constitute a basis of the space, called an orthogonal basis. ■

Notice that any vector x in Euclidean space may be normalized to have length one, by the simple device of scalar multiplication by $\alpha = \frac{1}{\|x\|}$. Thus if $z \leftarrow \frac{x}{\|x\|}$ then

$$\begin{aligned} \|z\| &= \sqrt{(z; z)} \\ &= \sqrt{\frac{(x; x)}{\|x\|^2}} \\ &= 1. \end{aligned}$$

Definition 10 The non-zero vectors $e_1; e_2; \cdots; e_n$ of an n-dimensional Euclidean space are said to form an orthonormal basis if they are pairwise orthogonal and if, in addition, each has unit length; that is if $(e_i; e_j) = \delta_{ij}$ where δ_{ij} is Kronecker's delta $i; j = 1; 2; \cdots; n$ (i.e. $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ij} = 1$ for $i = j$). ■

Before moving on to discuss orthogonal bases, one theorem and one definition will be introduced.

Theorem 2 In Euclidean space, let x and y be two vectors. Then $\|x+y\| \leq \|x\|+\|y\|$.

Proof.

$$\begin{aligned} \|x + y\|^2 &= (x + y; x + y) \\ &= (x; x) + 2(x; y) + (y; y). \end{aligned}$$

Now $(x; y) \leq \|x\| \|y\|$: Hence

$$\begin{aligned} \|x + y\|^2 &= (x + y; x + y) \\ &\leq (x; x) + 2\|x\|\|y\| + (y; y). \end{aligned}$$

But

$$\begin{aligned}(x; x) + 2kxkkyk + (y; y) &= kxk^2 + 2kxkkyk + kyk^2 \\ &= (kxk + kyk)^2.\end{aligned}$$

Thus

$$kx + yk \cdot kxk + kyk. \blacksquare$$

De¯nition 11 *The distance between the vectors x and y in Euclidean space is de¯ned as* $kx ¡ yk$. ■

Every n-dimensional Euclidean vector space contains orthogonal bases, and hence orthonormal bases. Let e_i $(i = 1; 2; ¢¢¢; n)$ be an $(n £ 1)$ vector with zeros everywhere except in the i'th position which is unity. Then $e_1; e_2; ¢¢¢; e_n$ constitutes an orthonormal basis:

$$[e_1; e_2; ¢¢¢; e_n] = I_n.$$

Moreover, $\sum e_i ´ e$ which has one in every position; e is called the equiangular line in R^n. Every n-tuple x in R^n may be written $x = I_n x$, whereupon x_i is seen to be the i'th co-ordinate of x relative to the orthonormal basis matrix I_n. Finally, consider

$$\begin{aligned}(e_i; x) &= e_i^{>} x \\ &= x_i.\end{aligned}$$

Also $\left(e_i^{>} e_i\right)^{-1} = 1$. Hence

$$e_i \left(e_i^{>} e_i\right)^{-1} e_i^{>} x = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ x_i \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

$e_i \left(e_i^{>} e_i\right)^{-1} e_i^{>}$ is an example of an orthogonal projection. Before turning our attention to these, it is important to introduce the concept of a subspace.

2.4 Subspaces

A subset of vectors in a vector space is called a subspace if it forms a vector space according to De¯nition 1. Consider the basis matrix I_n of R^n. The ¯rst k columns of I_n form an $(n \times k)$ matrix E_k and the remaining $(n - k)$ columns an $n \times (n - k)$ matrix E_{n-k}. Clearly E_k has rank k and E_{n-k} has rank $(n - k)$. E_k may be regarded as a basis of a k-dimensional subspace and E_{n-k} as a basis for an $(n - k)$-dimensional subspace. Let the ¯rst subspace be denoted L and the second as $L^{?}$. It is clear that any vector in L is orthogonal to any vector in $L^{?}$ because

$$E_k^{>} E_{n-k} = 0.$$

This implies that for all $\mu \in R^k$ and all $\lambda \in R^{n-k}$:

$$(E_k\mu)^\top E_{n-k}\lambda = \mu^\top E_k^\top E_{n-k}\lambda$$
$$= 0.$$

Thus any vector in L, written $E_k\mu$, is orthogonal to all vectors in $L^\perp$, written $E_{n-k}\lambda$, indeed, the superscript $\perp$ denotes orthogonal complement since $L + L^\perp = R^n$, by definition, and both L and $L^\perp$ intersect, by construction, only at the origin $= 0$. In this particular case, there is only intersection at the origin; it is then usual to write $L \oplus L^\perp = R^n$ where $\oplus$ denotes direct sum; however, it is not necessary for the subspaces to be orthogonal. Thus two subspaces L and M might not be orthogonal yet $L \cap M = 0$ in which case we could write $L \oplus M = V$, say, and M would be the complement of L in V.

Definition 12 *Let L be a subspace of an n-dimensional Euclidean space R^n. A vector in R^n is orthogonal to L if it is orthogonal to every vector in L.* ■

Two orthogonal subspaces of great interest in econometrics may now be introduced. Consider the linear model introduced earlier in equation (1:4)

$$y = X\beta + \varepsilon,$$

where X has k linearly independent columns and n rows. The first subspace in R^n of interest is the range (or span) of X which is the set of all vectors that are linear

combinations of the columns of X:

$$R[X] = \{x \in R^n : x = X\beta; \beta \in R^k\}.$$

Clearly, R[X] is a subspace. With $x_1 = X\beta_1$ and $x_2 = X\beta_2$, $x_1 + x_2 = X(\beta_1 + \beta_2) = X\beta$, $\beta = \beta_1 + \beta_2$. Hence, $X\beta = x \in R[X]$. Also if α is a scalar and $x \in R[X]$, $x\alpha = X\beta\alpha = X\mu$; $\mu \in R^k$, hence $\beta x \in R[X]$. Since X has rank k, it is always possible to form k linearly independent linear combinations of its columns such that

$$XM = W$$

where M is a $(k \times k)$, non-singular matrix Then the rank of W = the rank of X and $R[W] = \{W \in R^n : w = W\mu; \mu \in R^k\}$. But $W\mu = XM\mu = X\beta$; say, and hence $R[X] = R[W]$. Clearly the dimension of R[X] is k.

The other subspace of interest is the null-space or kernel of $X^\top$:

$$N(X^\top) = \{z \in R^n : X^\top z = 0\}.$$

Clearly, $N(X^\top)$ is a sub space, for if z_1 and $z_2 \in N(X^\top)$ then $X^\top(z_1 + z_2) = X^\top z = 0$. $z_1 + z_2 = z$, whereupon $z \in N(X^\top)$. Moreover, if $X^\top z = 0$, then $X^\top(\alpha z) = \alpha X^\top z = 0$, $\alpha \in R$ and $\alpha z \in N(X^\top)$.

Theorem 3 For the $(n \times k)$ real matrix X of rank k, $N(X^\top) = R[X]^\perp$.

Proof. Let $z \in N(X^\top)$, then $X^\top z = 0$ which implies, for all $\beta \in R^k$, $\beta^\top X^\top z = 0$ or $(X\beta)^\top z = 0$ and $z \in R[X]^\perp$. Let $z \in R[X]^\perp$: Then $X^\top z = 0$ and $z \in N(X^\top)$. ■

The dimension of $R[X]$ = rank X = k. The dimension of $N\left[X^{>}\right]$ = the number of columns of $X^{>}$ ¡ rank $X^{>}$ = $(n ¡ k)$. The dimension of $N[A]$, where A is any matrix, is called the nullity of A. Moreover, in view of Theorem 3, it is clearly the case that, if A has order $(n £ k)$ and rank p,

$$\dim R[A] + \dim N\left[A^{>}\right] = p + (n ¡ p) = n$$

and this is equivalent to saying

$$\text{rank } A + \text{nullity of } A^{>} = n.$$

Of course, the nullity of $X = k$ ¡ rank $X = k$ ¡ $k = 0$ i.e. $N[X]$ contains only the zero vector. i.e. $X^{>}z = 0$ has no solution save $z = 0$.

Every vector subspace may be represented as the range or the null-space of a matrix. Let V_1 and V_2 be subspaces of R^n of dimension k_1 and k_2. Let X_1 and X_2 be $(n £ k_1)$ and $(n £ k_2)$ matrices of rank k_1 and k_2. Then V_1 could be de¯ned as $R[X_1]$ and V_2 as $R[X_2]$.

Theorem 4 If V_1 and V_2 are subspaces of R^n, then $(V_1 + V_2)^{?} = V_1^{?} \setminus V_2^{?}$.

Proof. Let $V_1 = R[X_1]$ with X_1 an $(n \times k_1)$ matrix of full rank. Let $V_2 = R[X_2]$ with X_2 an $(n \times k_2)$ matrix of rank k_2 Then

$$\begin{aligned}
(V_1 + V_2)^{\perp} &= [R[X_1] + R[X_2]]^{\perp} \\
&= R\left[X_1 : X_2\right]^{\perp} \\
&= N\begin{bmatrix} X_1' \\ X_2' \end{bmatrix} \\
&= \left\{\mu \in R^n : \begin{bmatrix} X_1' \\ X_2' \end{bmatrix}\mu = 0\right\} \\
&= V_1^{\perp} \cap V_2^{\perp}
\end{aligned}$$

since $N(X_1') = V_1^{\perp}$ and $N(X_2') = V_2^{\perp}$. ■

Theorem 5 $(V_1 \cap V_2)^{\perp} = V_1^{\perp} + V_2^{\perp}$.

Proof. Let $V_1 = N(X_1')$ and $V_2 = N(X_2')$. Where X_1 and X_2 are appropriate $(n \times k_1)$ and $(n \times k_2)$ matrices of rank k_1 and k_2 respectively. Then

$$\copyright N(X_1') \cap N(X_2')^{\perp} = \left\{\mu \in R^n : \begin{bmatrix} X_1' \\ X_2' \end{bmatrix}\mu = 0\right\}.$$

It follows that

$$\begin{aligned}
\copyright N(X_1') \cap N(X_2')^{\perp} &= R\left[X_1 : X_2\right] \\
&= R[X_1] + R[X_2] \\
&= V_1^{\perp} + V_2^{\perp}. \blacksquare
\end{aligned}$$

3 Projection and Least Squares

3.1 Projection Matrices

The idea of orthogonal projection arises in its simplest form when there are two arbitrary vectors in the plane R^2, separated by an acute angle μ, as indicated in Figure 3.1. A perpendicular is dropped

from the vector y onto the line on which the vector x lies, yielding another vector which will be labelled $m = Py$. Here P is the matrix that transforms y into m, and the vector difference $(y - m) = (I_2 - P)y$ is labelled r. The problem is to determine the matrix P. Clearly, since y, x, m and r all lie in the plane R^2, P must be a (2×2) matrix.

Notice first that, since x and m are co-linear, m must be obtainable from x by scalar multiplication. If this is written $m = x\lambda$, then λ is a positive fraction because m is in the same direction as x but shorter. Indeed, λ is nothing more than the proportion of the length of x by which the vector x must be reduced to have the

length of m. Thus

$$\lambda = \frac{\|m\|}{\|x\|},$$

where length is determined by the natural scalar product which, for x and y, will be written $(x; y) = x^{>}y$: It follows that

$$m = Py = x\lambda = x\left(\frac{\|m\|}{\|x\|}\right). \tag{3.1}$$

Now it was established in chapter 2 that $(x; y) = \|x\|\|y\|\cos\mu$ and $\cos\mu = \frac{\|m\|}{\|y\|}$. Thus

$$x^{>}y = \|x\|\|y\|\left(\frac{\|m\|}{\|y\|}\right) = \|x\|\|m\|,$$

or

$$\|m\| = \frac{x^{>}y}{\|x\|}. \tag{3.2}$$

Substituting (3:2) into (3:1),

$$m = Py = x\left(\frac{x^{>}y}{\|x\|^2}\right). \tag{3.3}$$

But $\|x\|^2 = x^{>}x$, whereupon

$$m = Py = x(x^{>}x)^{-1}x^{>}y, \tag{3.4}$$

implying that $P = x(x^{>}x)^{-1}x^{>}$, a (2×2) matrix as required. Notice that P is symmetric and idempotent, $P = P^{>} = P^2$; this may be checked by applying the rules

for transposition and matrix multiplication. This symmetry and idempotence of P is no coincidence, but rather a consequence of the orthogonality of m and (y ¡ m) = r, for $m^{>} r = 0$ implies

$$(Py)^{>} (I_2 ¡ P) y = 0$$

or $y^{>} P^{>} (I_2 ¡ P) y = 0$ for all y, which cannot be unless $P^{>} = P^{>}P$. Upon transposing $P^{>}$ in the last expression,

$$P = P^{>} = P^2.$$

This is quite a general rule: if the line x is replaced by the k-dimensional linear subspace L, while y in R^2 is replaced by y in R^n, k < n, the required (n £ n) projection matrix, P, that takes y in R^n orthogonally into Py in L, will have the properties of symmetry $P = P^{>}$ and idempotence ($P = P^2$). These properties are a consequence of the orthogonality condition arising from the natural scalar product.

Consider again the linear model (1:4) of chapter 1 and let R[X] = L, L being a subspace of R^n on which the natural scalar product is de¯ned as before. The vector y is an element of R^n and the vector X¯ is set equal to ¹. Then (1:4) becomes

$$y = ¹ + " \qquad (3.5)$$

where y 2 R^n and ¹ 2 L ½ R^n: The coe± cient vector ¯ in equation (1:4) is unknown and so the location of ¹ in L is unknown; ¹ must therefore be estimated and this

may be done by setting $\mu = \hat{\mu}$ where $\hat{\mu}$ is the orthogonal projection of y on L, that is, $\hat{\mu} = Py$ in which P is the orthogonal projection matrix from R^n onto L. This leaves ε to be estimated as $\hat{\varepsilon} = y - \hat{\mu} = (I_n - P)y$, $\hat{\varepsilon}$ being orthogonal to $\hat{\mu}$, i.e. $\hat{\varepsilon} \in L^{?}$. Since orthogonality is defined according to (:; :) and L has basis matrix X, then $\hat{\mu} = Py = X\hat{\beta}$, say; and $\hat{\varepsilon} = y - X\hat{\beta} = (I_n - P)y$ is required to be orthogonal to every vector x in L, that is, setting $x = X\beta$; for all $\beta \in R^k$,

$$(x; y - \hat{\mu}) = (X\beta)^{>}(y - Py) = \beta^{>}X^{>}\left(y - X\hat{\beta}\right) = 0.$$

Thus

$$\beta^{>}X^{>}y - \beta^{>}X^{>}X\hat{\beta} = 0$$

for all $\beta \in R^k$ and since the last equality must hold identically in β

$$\hat{\beta} = \left(X^{>}X\right)^{-1}X^{>}y$$

and

$$X\hat{\beta} = Py = X\left(X^{>}X\right)^{-1}X^{>}y.$$

This reveals that $P = X\left(X^{>}X\right)^{-1}X^{>}$ and again $P = P^{>} = P^2$.

Another property of P that is important is that it is unique for L. For suppose that $W = XM$ is another basis matrix of L, M being non-singular of order k. By corresponding arguments to those used above, the orthogonal projection matrix from

R^n to L relative to the basis matrix W is $W(W'W)^{-1}W'$. But

$$W(W'W)^{-1}W' = XM(M'X'XM)^{-1}M'X'.$$

By the rule for inversion of a matrix product and the non-singularity of M and $(X'X)$,

$$\begin{aligned} XM(M'X'XM)^{-1}M'X' &= XMM^{-1}(X'X)^{-1}(M')^{-1}M'X' \\ &= X(X'X)^{-1}X' \\ &= P. \end{aligned}$$

Thus the matrix P is invariant to the basis used to represent L. P is, of course, nnd and, not surprisingly since P depends only on X, R[P] = R[X]. Moreover, the rank of P = trP = $trX(X'X)^{-1}X' = tr(X'X)^{-1}X'X = k$.

Finally note that for all $x \in L$,

$$(y - x) = (y - \hat{y}) + (\hat{y} - x). \tag{3.6}$$

Now $(\hat{y} - x) \in L$ and $(y - \hat{y})$, by definition, is orthogonal to every vector in L, including $(\hat{y} - x)$. Thus $(y - \hat{y})'(\hat{y} - x) = (\hat{y} - x)'(y - \hat{y}) = 0$ and

$$\|y - x\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - x\|^2 \tag{3.7}$$

which is a representation of Pythagoras' Theorem. Since x is any vector in L and $\hat{y}$ is unique (because y is given and P is unique), x may be chosen so as to make $\|y - x\|^2$ as small as possible. Equation (3:7) indicates that

$$\min_{x \in L} \|y - x\|^2 = \|y - \hat{y}\|^2 \tag{3.8}$$

i.e. that by choosing $x = \hat{} = Py$, $(y - x)^{>}(y - x)$ is minimized for given y and all x in L. This is the principle of least squares: the sum of squares $(y - x)^{>}(y - x)$ is made least by setting $x = Py = \hat{}$. An alternative way of saying the same thing is that the shortest distance between y and the hyperplane L is $k(I_n - P)yk$.

For any $x \in L$, x may be represented uniquely by $x = X_{¸}$ for some $¸ \in R^k$; because X is a basis of L. Thus if P is the orthogonal projection from R^n onto L, then

$$Px = X\left(X^{>}X\right)^{-1}X^{>}X_{¸} = X_{¸} = x.$$

Since x is any vector in L, then $Px = x \; 8x \in L$. In the orthogonal compliment $L^{?}$ of L in R^n, recall that $L^{?} = R[X]^{?} = N\left(X^{>}\right)$, of dimension $(n - k)$. Let $z \in L^{?}$. Then $Pz = X\left(X^{>}X\right)^{-1}X^{>}z$ and since $z \in N\left(X^{>}\right)$, $X^{>}z = 0$, a condition that holds for every $z \in L^{?}$. Thus, given that P is the orthogonal projection onto L, then $Px = x$ for every $x \in L$ and $Pz = 0$ for every $z \in L^{?}$. Indeed, for any vector y in R^n, there exists a unique decomposition

$$y = x + z \tag{3.9}$$

with $x \in L$ and $z \in L^{?}$. Moreover

$$Py = Px + Pz = x + 0 = x$$

and

$$(I_n - P)y = (I_n - P)x + (I_n - P)z = 0 + z = z.$$

Since P is the unique orthogonal projection matrix onto L, then $(I_n - P)$ is the unique orthogonal projection matrix onto $L^{?}$:

Suppose now that L_0 is a subspace that lies entirely in L: $L_0 \subset L \subset R^n$. L_0 may be formulated as L subject to a set of r linear restrictions, $r < k$, like $A^{>}x = 0$, $x \in L$, A being a given $(n \times r)$ matrix of rank r. Then $L_0 = L \cap N\left(A^{>}\right) = \left\{x \in L : A^{>}x = 0\right\}$. Alternatively, $X_0 = XM$, where M is a known $[k \times (k - r)]$ matrix of rank $(k - r)$, and X_0 is a basis of L_0. The unique orthogonal projection matrix from R^n onto L_0, is then given by $P_0 = X_0\left(X_0^{>}X_0\right)^{-1}X_0^{>} = XM\left(M^{>}X^{>}XM\right)^{-1}M^{>}X^{>}$. The last expression cannot be simplified because M is not invertible. Let x_0 be any vector in L_0. Since X_0 is the $[n \times (k - r)]$ basis for L_0, x_0 has unique representation $x_0 = X_0\gamma$, for some $\gamma \in R^{k-r}$. Thus $x_0 = X_0\gamma = XM\gamma$. Notice that $P_0x = x_0$ and that $Px_0 = PXM\gamma = XM\gamma = X_0\gamma = x_0$. It follows that $PP_0 = PXM\left(M^{>}X^{>}XM\right)^{-1}M^{>}X^{>} = P_0$. Moreover $P_0P = P_0$. Indeed, if P and Q are orthogonal projection matrices and $PQ = Q$, $QP = Q$ then $R[Q] \subset R[P]$.

3.2 Means and Decompositions

The equiangular vector in R^n, which was introduced in chapter 2, plays an important role in statistics as the following special case will reveal. Let x be a vector of n observations x_i $(i = 1; 2; \cdots; n)$; x will be regarded as an element in n-dimensional Euclidean space R^n on which the natural scalar product is defined. The equiangular

vector in R^n is de¯ned by $e^> = [1; 1; ¢¢¢; 1]$. The orthogonal projection from R^n onto the equiangular line $E = R[e]$ is

$$p = e\,(e^>e)^{¡1}e^>.$$

Notice that $e^>x = x^>e = \sum x_i$ and that $e^>e = n$. It follows that $(e^>e)^{¡1}e^>x = \frac{1}{n}\sum x_i = \bar{x}$, the mean of the observations $x_1; x_2; ¢¢¢; x_n$. Thus

$$px = \begin{bmatrix} \bar{x} \\ \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix} = e\bar{x}; \quad (I_n ¡ p)x = \begin{bmatrix} x_1 ¡ \bar{x} \\ x_2 ¡ \bar{x} \\ \vdots \\ x_n ¡ \bar{x} \end{bmatrix},$$

and it follows immediately that $x^>(I_n ¡ p)x = \sum (x_i ¡ \bar{x})^2$ because $(I_n ¡ p)$ is symmetric and idempotent.

Suppose now that the n observations $x_1; x_2; ¢¢¢; x_n$ represent a randomly selected sample from a population which has mean $¹$ and variance $¾^2$. Then

$$(x ¡ e¹) = p(x ¡ e¹) + (I_n ¡ p)(x ¡ e¹) \tag{3.10}$$

is the unique decomposition of the vector $(x ¡ e¹) \in R^n$ into $p(x ¡ e¹)$ on E and $(I_n ¡ p)(x ¡ e¹)$ on $E^?$. Notice carefully that the dimension of E (= rank p)is dim $E = 1$ while dim $E^?$ = rank $(I_n ¡ p) = (n ¡ 1)$; also, since $e \in E$, $(I_n ¡ p)e = 0$. The decomposition (3:10) thus corresponds to the decomposition (3:9), where $N(e^>)$ =

$E^?$. Now the right-hand side (RHS) of (3:10) comprises two orthogonal vectors and hence their scalar product is zero. Consequently,

$$(x ¡ e^1)^> (x ¡ e^1) = (x ¡ e^1)^> p (x ¡ e^1) + x^> (I_n ¡ p) x \qquad (3.11)$$

and, in more familiar statistical notation, (3:11) may be written

$$\sum (x_i ¡ {}^1)^2 = n(\bar{x} ¡ {}^1)^2 + \sum (x_i ¡ \bar{x})^2. \qquad (3.12)$$

The decomposition (3:11) or (3:12)) is important in the development of the F- and t-distributions, on the assumption that the population from which x is drawn is normal.

Note ¯rst that the rank of the left-hand side (LHS) of (3:11), and hence of (3:12), is n, the ¯rst term on the RHS has rank 1 and the second term has rank $(n ¡ 1)$. Thus the ranks on the RHS sum to the rank of the LHS. Secondly, if the population from which x is drawn is normal, then each

$$\frac{x_i ¡ {}^1}{¾} » NID(0; 1); \; i = 1; 2; ¢¢¢; n.$$

Hence, dividing (3:12) throughout by $¾^2$, the LHS is distributed as the sum of squares of n independent $N(0; 1)$ variates, that is, as a central $Â^2(n)$ distribution. Thirdly, as a consequence of the sum of ranks on the RHS being equal to the rank of the LHS and the known distribution of $\sum (x_i ¡ {}^1)^2$, $\left\{ \frac{n(\bar{x} ¡ {}^1)^2}{¾^2} \right\}$ and $\left\{ \frac{\sum (x_i ¡ \bar{x})^2}{¾^2} \right\}$ are independently distributed as $Â^2(1)$ and $Â^2(n ¡ 1)$. Hence their ratio adjusted by rank (that is

degrees of freedom) is

$$\left[\frac{n(\bar{x}_{\cdot 1})^2}{\sigma^2} \Big/ \frac{\sum (x_{i1} - \bar{x})^2}{(n-1)\sigma^2}\right] \sim F(1; n-1).$$

Hence, eliminating σ^2, taking the square root and re-arranging

$$\frac{(\bar{x}_{\cdot 1} - \mu_1)}{s/\sqrt{n}} \sim t(n-1) \tag{3.13}$$

where $s^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2$. The expression (3:13) should be familiar from elementary statistics.

Turning back now to the model (1:4) of chapter 1, let the first column of X be e, so that the model may be written

$$y_i = \beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + \cdots + x_{ik}\beta_k + \varepsilon_i \tag{3.14}$$

for $i = 1; 2; 3; \cdots; n$. In this case

$$X = [e; x_{\cdot 2}; x_{\cdot 3}; \cdots; x_{\cdot k}]. \tag{3.15}$$

Notice that for $L = R[X]$, $e \in L$ and hence for the orthogonal projection matrix P onto L,

$$Pe = e; \quad Pp = p.$$

Similarly, if $X_1 = [e; x_{\cdot 2}; x_{\cdot 3}; \cdots; x_{\cdot j}]$ and $X_2 = [x_{\cdot j+1}; x_{\cdot j+2}; \cdots; x_{\cdot k}]$ then $PX_1 = X_1$ and $PX_2 = X_2$. With this background, consider, $y'(I - P)y = \sum (y_i - \bar{y})^2$ where

$\bar{y}$ is the mean of the elements of y. A decomposition, corresponding to (3:11), that applies to (3:14) is

$$y^{>}(I_n - p)y = y^{>}(P - p)y + y^{>}(I_n - P)y. \tag{3.16}$$

Such a decomposition is said to be `adjusted for the mean' since, instead of decomposing $y^{>}y$ into $y^{>}Py$ and $y^{>}(I_n - P)y$, the projection p onto the equiangular line is inserted into the LHS and the first term on the RHS and $py = e\bar{y}$; this is precisely the adjustment for the mean.

The following nomenclature is used:

$y^{>}(I_n - p)y = S_T^2$, or the total sum of squares;

$y^{>}(P - p)y = S_E^2$, or the explained sum of squares;

$y^{>}(I_n - P)y = S_U^2$, or the unexplained sum of squares.

S_T^2 has rank $(n - 1)$, S_E^2 has rank $(k - 1)$ and S_U^2 has rank $(n - k)$ because $tr(I_n - p) = (n - 1)$, $tr(P - p) = (k - 1)$ and $tr(I_n - P) = (n - k)$. Clearly $(n - 1) = (k - 1) + (n - k)$ and , re-arranging (3:16),

$$S_U^2 = S_T^2 \left(1 - \frac{S_E^2}{S_T^2}\right).$$

If $\frac{S_E^2}{S_T^2} = R^2$ then

$$S_U^2 = S_T^2 \left(1 - R^2\right). \tag{3.17}$$

Equation (3:17), in particular R^2, has a nice interpretation arising from the Cauchy-Schwartz inequality. This comes about as follows.

$$\frac{S_E^2}{S_T^2} = \frac{y^{>}(P - p)y}{y^{>}(I_n - p)y} \cdot \frac{y^{>}(P - p)y}{y^{>}(P_a - p)y} = \frac{\left(y^{>}(P - p)y\right)^2}{\|(I_n - p)y\|^2 \, \|(P - p)y\|^2}.$$

Now $y^{>}(P - p)y$ is equal to $y^{>}(P - p)(I_n - p)y$ because

$$(P - p)(I_n - p) = (P - p - p + p) = (P - p).$$

Hence

$$R^2 = \frac{\left(y^{>}(P - p)(I_n - p)y\right)^2}{\|(I_n - p)y\|^2 \, \|(P - p)y\|^2}$$

It follows from the Cauchy-Schwartz inequality that

$$0 \cdot \frac{\left(y^{>}(P - p)(I_n - p)y\right)^2}{\|(I_n - p)y\|^2 \, \|(P - p)y\|^2} \cdot 1. \tag{3.18}$$

Let μ be the angle between $(P - p)y$ and $(I_n - p)y$. Then

$$R^2 = \cos^2 \mu \tag{3.19}$$

and from equation (3:18)

$$-1 \cdot R \cdot 1. \tag{3.20}$$

From (3:17) and (3:18)), $(1 - R^2)$ is seen as the proportion of S_T^2 that is unexplained (S_U^2) and hence R^2 is the proportion of S_T^2 that is explained. Since $(P - p)y =$

$X\hat{\beta} - e\bar{y}$, writing $\hat{y}$ for $X\hat{\beta}$,

$$y'(P - p)y = \sum (\hat{y}_i - \bar{y})^2$$

which is seen as that part of S_T^2 which the least squares line $X\hat{\beta} = \hat{y} = \hat{\alpha}$, adjusted by the mean of y ($e\bar{y}$), explains. Finally, from (3:18),

$$R^2 = \frac{\{\sum (\hat{y}_i - \bar{y})(y_i - \bar{y})\}^2}{\sum (\hat{y}_i - \bar{y})^2 \sum (y_i - \bar{y})^2}$$

the square of the product moment correlation coefficient between y_i and $\hat{y}_i$ from elementary statistics. Of course, $py = e\bar{y}$ and $py = p[Py + (I_n - P)y]$; but $E \perp L$ and so $pP = p$, $p(I - P) = 0$. Thus $py = e\bar{y} = pPy = pX\hat{\beta}$. It follows that the mean of $y_1, y_2, \cdots, y_n$ is $\bar{y}$ and the mean of $\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n$ is also $\bar{y}$.

3.3 Orthogonality and Least Squares

There are various different ways to formulate the method of least squares using the natural scalar product. The one that emphasizes minimizing the sum of squared errors, from which the term least squares comes, may be written

$$\arg\min_{\theta \in L} \|y - \theta\|^2 = \hat{\alpha} = Py. \tag{3.21}$$

In equation (3:21), the expression $\|k\|^2$ is squared length defined by the natural scalar product $(\cdot, \cdot)$. A corresponding expression emphasizes the orthogonality of $(y - \hat{\alpha})$ with L. Since Py is unique for given y, then $\hat{\alpha}$ is unique, being defined by the condition

$$(y - \hat{\alpha}, x) = 0 \quad \forall x \in L. \tag{3.22}$$

Expression (3:22) is important and it plays an especially important role in the discussion of invariance below (section 3.6).

Least squares which de¯nes orthogonality relative to the natural scalar product (:; :) is usually referred to as ordinary least squares or OLS. A more general expression for the scalar product is $h:; :i = (:; A:)$ for some $(n £ n)$, nnd matrix A. To be a proper scalar product, A must be pd but, under certain conditions, A may be nnd whereupon $h:; :i$ is said to be a quasi-scalar product. If in fact $A (6= I_n)$ is pd, then least square relative to $h:; :i$ is referred to as generalized least squares (GLS); where A is nnd then least squares relative to $h:; :i$ belongs to the class of generalized instrumental variables estimators (GIVE). GIVE covers the following methods: instrumental variables (IV), two-stage least squares (2SLS), three-stage least squares (3SLS), generalized 2SLS and generalized 3SLS (G2SLS and G3SLS) and the generalized method of moments (GMM). At this stage, it is not convenient or helpful to enter into a discussion of these methods. Rather, attention will directed toward the general case of least squares relative to the quasi-scalar product $h:; :i$.

Model (1:4) will take the form of (3:14) and X the form of (3:15). The matrix A may be pd or nnd and its rank will be K, $k · K · n$, so that $X^>AX$ is invertible; alternatively $L^? \setminus R[A] = Á$. In this setting, the least squares criterion will be (3:22), relative to the scalar product $h:; :i$. Thus $¹ = X^-$ is selected at $^ = X^^$ such that

$$hy ¡ ^; xi = (y ¡ ^; Ax) = 0 \quad 8x 2 L. \tag{3.23}$$

Writing $x = X\beta$ for any $\beta \in R^k$, then

$$\beta^\top X^\top A y - \beta^\top X^\top A X \bar{\beta} = 0 \tag{3.24}$$

identically in β. Thus since $X^\top A X$ is invertible,

$$X\hat{\beta} = \hat{x} = X\left(X^\top A X\right)^{-1} X^\top A y. \tag{3.25}$$

If $X\left(X^\top A X\right)^{-1} X^\top A = F$, then $F \neq F^\top$ but $F = F^2$. For this reason F is often referred to as an oblique projection matrix. More precisely, F is the projection matrix on L, orthogonal relative to $\langle\cdot;\cdot\rangle$. In the same language, $P = X(X^\top X)^{-1}X^\top$ is the projection matrix on L, orthogonal relative to the natural scalar product $(\cdot;\cdot)$.

Recall that in section 3.3, relative to the scalar product $(\cdot;\cdot)$, $p = e\left(e^\top e\right)^{-1} e^\top$ and

$$\begin{aligned} S_T^2 &= y^\top (I_n - p) y \\ &= (y; (I_n - p) y) \\ &= (y - py; y - py) \\ &= \|(I_n - p) y\|^2 \end{aligned}$$

because $(I_n - p)$ is symmetric and idempotent. Similarly,

$$\begin{aligned} S_E^2 &= y^\top (P - p) y = \|(P - p))y\|^2, \\ S_U^2 &= y^\top (I_n - P) y = \|(I_n - P) y\|^2. \end{aligned}$$

What happens to these quantities when the scalar product is altered from $(\cdot;\cdot)$ to $\langle\cdot;\cdot\rangle = (\cdot; A\cdot)$?

[2] It is easily seen that p has to be replaced by $f = e^{\mathbf{i}} e^{>} A e^{\phi_{i-1}} e^{>} A$ following the form of F using the equiangular vector e. But neither F nor f is symmetric (though each is idempotent). On the other hand, for two vectors z and w, $\langle z; w\rangle = (z; Aw)$. Thus, for example,

$$\begin{aligned}\langle y; (I_n - f)y\rangle &= (y; A(I_n - f)y) \\ &= \left(y; A - Ae^{\mathbf{i}} e^{>} A e^{\phi_{i-1}} e^{>} A\right).\end{aligned}$$

Now

$$\begin{aligned}\langle (I_n - f)y; (I_n - f))y\rangle &= (y - e^{\mathbf{i}} e^{>} A e^{\phi_{i-1}} e^{>} Ay; \\ &\quad Ay - Ae^{\mathbf{i}} e^{>} A e^{\phi_{i-1}} e^{>} Ay) \\ &= y^{>}Ay - y^{>}Ae^{\mathbf{i}} e^{>} A e^{\phi_{i-1}} e^{>} Ay \\ &\quad - y^{>}Ae^{\mathbf{i}} e^{>} A e^{\phi_{i-1}} e^{>} Ay \\ &\quad + y^{>}Ae^{\mathbf{i}} e^{>} A e^{\phi_{i-1}} e^{>} Ae^{\mathbf{i}} e^{>} A e^{\phi_{i-1}} e^{>} Ay \\ &= y^{>}Ay - y^{>}Afy - y^{>}Afy + y^{>}Afy,\end{aligned}$$

because $Af = f^{>}A$, that is Af is symmetric; and $f^{>}Af = Af$. Hence

$$\begin{aligned}\langle y; (I_n - f)y\rangle &= \langle (I_n - f)y; (I_n - f)y\rangle \\ &= \|(I_n - f)y\|^2_{\langle\cdot\rangle}\end{aligned}$$

where $k\cdot k_{hi}^2$ denotes length relative to the scalar product $h\cdot;\ \cdot i$.

It follows from this discussion that $(I_n - F)$, $(F - f)$ and $(I_n - f)$ are symmetric relative to the scalar product $h\cdot;\ \cdot i$; using $(F - f)$, then

$$A(F - f) = (F - f)^{>} A = (F - f)^{>} A(F - f)$$

or for arbitrary vectors w and z in R^n

$$hz;\ (F - f)wi = h(F - f)z;\ wi = h(F - f)z;\ (F - f)wi.$$

Notice, incidentally, that just as P obeys, relative to $(\cdot;\ \cdot)$,

$$(\cdot;\ P\cdot) = (P\cdot;\ \cdot) = (P\cdot;\ P\cdot) = \left(\cdot;\ P^2\cdot\right),$$

so F obeys, relative to $h\cdot;\ \cdot i$

$$h\cdot;\ F\cdot i = hF\cdot;\ \cdot i = hF\cdot;\ F\cdot i = \left(\cdot;\ F^2\cdot\right).$$

The same holds for f, $(I_n - f)$; $(F - f)$ and $(I_n - F)$, indeed any projection matrix which is orthogonal relative to $h\cdot;\ \cdot i$.

Turning back to S_T^2, S_E^2 and S_U^2, relative to the scalar product $h\cdot;\ \cdot i$; these become

$$S_T^2 = k(I_n - f)yk_{hi}^2,$$

$$S_E^2 = k(F - f)yk_{hi}^2,$$

$$S_U^2 = k(I_n - F)yk_{hi}^2.$$

It is natural to de¯ne R^n, relative $h:;:i$, by analogy with R^2 relative to $(:;:)$ in (3:18). Thus from (3:18)

$$R^2 = \frac{(fP ¡ pg; fI_n ¡ pgy)^2}{k(P ¡ p)yk^2 k(I_n ¡ p)yk^2}$$

which is $\cos^2 \mu$ relative to the natural scalar product. When $h:;:i$ replaces $(:;:)$, (3:18) is replaced by

$$R^2_{hi} = \frac{h(F ¡ f)y; (I_n ¡ f)yi^2}{k(F ¡ f)yk^2_{hi} \; k(I_n ¡ f)yk^2_{hi}} \tag{3.26}$$

which is $\cos^2 \mu$ relative to $h:;:i$.

One way to look upon the scalar product $h:;:i$ is as the natural scalar product in transformed vectors. If, for example, A is pd, then there exists a non-singular R of order n such that $R^>R = A$. Thus $hw; zi = (w; Az) = (w; R^>Rz) = (Rw; Rz)$. Thus when w and z are transformed into $Rw = r$ and $Rz = s$, then $hw; zi = (r; s)$ and any analysis of vectors relative to $h:;:i$ may be recast, by a transformation matrix R, into a corresponding analysis of transformed vectors relative to the natural scalar product $(:;:)$.

If the matrix A is an orthogonal projection matrix of rank K (and hence is nnd), then there exists an $(n £ K)$ matrix B such that $A = BB^>$ and $B^>B = I_K$. In this case, $hw; zi = (w; Az) = \left(B^>w; B^>z\right) = (p; q)$ with $B^>w = p; B^>z = q$. Note carefully that p and q will be K-tuples. The transformation $B^>$ is an example of a partial isometric matrix or partial isometry. Suppose that A is the orthogonal projection

matrix on a subspace M of dimension K (along $M^{?}$ of dimension $(n - K)$), then for every $x \in M$, $Ax = x$ and $\|B^{>}x\|^2 = x^{>}BB^{>}x = x^{>}Ax = x^{>}x = \|x\|^2$. Thus $\|B^{>}x\| = \|x\|$; for every $z \in M^{?}$, $B^{>}z = 0$. For example consider $P = X\left(X^{>}X\right)^{-1}X^{>}$ onto L along $L^{?}$. Since X has full rank, $X^{>}X$ is pd and hence there exists a $(k \times k)$ non-singular Q such that $Q^{>}Q = \left(X^{>}X\right)^{-1}$. Then $P = XQ^{>}QX^{>} = BB^{>}$ with $B = XQ^{>}$ an $(n \times k)$ matrix. For any $x \in L$, $\|B^{>}x\| = \sqrt{x^{>}BB^{>}x} = \sqrt{x^{>}x} = \|x\|$ as required. For any $z \in L^{?}$, $B^{>}z = QX^{>}z = 0$ since $L^{?} = N\left(X^{>}\right)$. Partial isometries are important in econometrics since they are transformations which hold lengths of vectors in a subspace fixed while annihilating vectors lying orthogonal to that subspace.

3.4 Restricted Least Squares

By restricted least squares is meant the model (1:4) subject to $r < k$ linear restrictions on the estimation of β; corresponding restrictions may be placed on the estimation of γ. Let these restrictions be defined in terms of a known $(k \times r)$ matrix B of rank r and a corresponding $(n \times r)$ matrix A:

$$B^{>}\beta = 0 \;\Rightarrow\; A^{>}\gamma = 0.$$

For given B and arbitrary $(n \times n)$ pd matrix Q,

$$B^{>}\beta = 0 \;\Rightarrow\; B^{>}\left(X^{>}QX\right)^{-1}X^{>}QX\beta = 0$$

yielding for $^{1} = X^{-}$ and $A = QX\left[X^{>}QX\right]^{¢_{i\,1}}B$, $A^{>1} = 0$. A is not unique for given B, but $PA = X\left[X^{>}X\right]^{¢_{i\,1}}B$ is. $B^{>-} = 0$ and $A^{>1} = 0$ are called restraint equations. Alternatively, if there are r linear restraints on k coe± cients $-$, then there must be $(k ¡ r)$ `free' coe± cients to estimate; let these be the $[(k ¡ r) £ 1]$ vectors ®. Then $-$ and ® must be related by

$$- = M®$$

where M is a $[k £ (k ¡ r)]$ matrix of elements determined by $B^{>-} = 0$.

Example 5 Let $k = 4$ and $r = 2$, so that

$$- = \begin{bmatrix} -_1 \\ -_2 \\ -_3 \\ -_4 \end{bmatrix}; \quad B^{>-} = \begin{bmatrix} 2 & ¡1 & 0 & 0 \\ 0 & 0 & 1 & ¡1 \end{bmatrix}\begin{bmatrix} -_1 \\ -_2 \\ -_3 \\ -_4 \end{bmatrix} = 0.$$

The second equations imply $2-_1 = -_2$ and $-_3 = -_4$: Thus $(k ¡ r) = 4 ¡ 2 = 2$) ® is $(2 £ 1)$. Let $®_1 = -_1$ and $®_2 = -_3$. Then

$$\begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} ®_1 \\ ®_2 \end{bmatrix} = M® = \begin{bmatrix} ®_1 \\ 2®_1 \\ ®_2 \\ ®_2 \end{bmatrix} = \begin{bmatrix} -_1 \\ -_2 \\ -_3 \\ -_4 \end{bmatrix}.$$

The equations $M® = -$ are called freedom equations. ■

The model (1:4) may now be expressed as

$$y = X\bar{} + " \tag{3.27}$$

subject to $B^{>}\bar{} = 0$ or as

$$y = {}^{1} + " \tag{3.28}$$

subject to $A^{>}{}^{1} = 0; {}^{1} 2 L$, in restraint equation form; or as

$$y = XM® + ", \tag{3.29}$$

$M® = \bar{}$; in freedom equation form. From (3:28) and (3:29) the subspace $L = R[X]$ of dimension k is now restricted to $L_0 = R[XM]$ of dimension $(k ¡ r)$. If the entire linear space R^n is considered, then this may be decomposed into orthogonal subspaces as follows

$$R^n = L_0 © (L_0^? \setminus L) © L^?. \tag{3.30}$$

The corresponding decomposition of y relative to (:; :) is

$$y = P_0 y + (P ¡ P_0) y + (I_n ¡ P) y \tag{3.31}$$

or adjusting for the mean,

$$(I_n ¡ p) y = (P_0 ¡ p) y + (P ¡ P_0) y + (I_n ¡ P) y \tag{3.32}$$

in which $P_0 = XM \left[M^{>}X^{>}XM\right]^{¢_{i\,1}} M^{>}X^{>}$. For later reference, it will be helpful to ¯nd an expression for $(P ¡ P_0)$ in terms of the matrix A in equation (3:28) and P.

Notice that in (3:28), the restricted subspace L_0 must be

$$L_0 = L \setminus N\left(A^{\top}\right).$$

The subspace of interest, from (3:30) is $L_0^{\perp} \setminus L$:

$$\begin{aligned} L_0^{\perp} \setminus L &= \left(L \setminus N\left(A^{\top}\right)\right)^{\perp} \setminus L \\ &= \left(L^{\perp} + R[A]\right) \setminus L \\ &= \left(L^{\perp} + PR[A] + (I_n - P)R[A]\right) \setminus L \\ &= \left(L^{\perp} + R[PA]\right) \setminus L \\ &= R[PA]. \end{aligned}$$

The dimension of $R[PA]$ is the rank of PA and this will be the same as the rank of A if and only if (i®) $R[A] \setminus L^{\perp} = 0$. This last condition is reasonable since $A = QX\left(X^{\top}QX\right)^{-1}B$ for arbitrary pd Q and $PA = X\left(X^{\top}X\right)^{-1}B$ then has the rank of B, namely r. Although A is arbitrary up to the choice of Q, PA is unique; indeed $PA = A$ where $Q = I_n$ which is admissable.

Given $L_0^{\perp} \setminus L = R[PA]$ and $R[A] \setminus L^{\perp} = 0$, the unique orthogonal projection on $L_0^{\perp} \setminus L$ is from (3:30) and (3:31),

$$\begin{aligned} P - P_0 &= PA\left(A^{\top}PA\right)^{-1}A^{\top}P \\ &= X\left(X^{\top}X\right)^{-1}B\left[B^{\top}\left(X^{\top}X\right)^{-1}B\right]^{-1}B^{\top}\left(X^{\top}X\right)^{-1}X^{\top}. \end{aligned} \tag{3.33}$$

Thus the decomposition (3:31) yields

$$(P - P_0)y = X\left[X'X\right]^{-1}B\left[B'\left[X'X\right]^{-1}B\right]^{-1}B'\left[X'X\right]^{-1}X'y. \qquad (3.34)$$

Setting $X\hat{\beta} = Py$ and $X\hat{\beta}_0 = P_0y$, where $\hat{\beta}$ is the (unrestricted) estimator of β in (1:4) and $\hat{\beta}_0$ is the corresponding restricted estimator in (3:27), (3:34) may be recast as

$$X\hat{\beta} - X\hat{\beta}_0 = X\left[X'X\right]^{-1}B\left[B'\left[X'X\right]^{-1}B\right]^{-1}B'\hat{\beta}$$

leading to

$$\hat{\beta}_0 = \hat{\beta} - \left[X'X\right]^{-1}B\left[B'\left[X'X\right]^{-1}B\right]^{-1}B'\hat{\beta}, \qquad (3.35)$$

a familiar expression in least squares theory.

When $A'\lambda = 0$ in (3:28) is replaced by the (a± ne) restrictions $A'\lambda = \mu$, $\mu \neq 0$, $\lambda \in L$, then there must exist a solution λ^* to $A'\lambda = \mu$, say $A'\lambda^* = \mu$, whereupon $A'(\lambda - \lambda^*) = 0$, $\lambda \in L$. Then (3:28) is replaced by

$$y - \lambda^* = (\lambda - \lambda^*) + \varepsilon;\ A'(\lambda - \lambda^*) = 0;\ \lambda \in L,$$

λ^* being regarded as a given constant vector. Repeating the same analysis, (3:35) is replaced by

$$\hat{\beta}_0 = \hat{\beta} - \left[X'X\right]^{-1}B\left[B'\left[X'X\right]^{-1}B\right]^{-1}\left(B'\hat{\beta} - \mu\right). \qquad (3.36)$$

Thus the formulation (3:28) is seen as being quite general in the sense that its least squares solution can accommodate the more general $A'\lambda = \mu$.

The relation between the freedom equation $\beta = M\theta$ and the restraint on equation $B'\beta = 0$ is straightforward to establish since $0 = B'\beta = B'M\theta$. Since θ is chosen without restriction, $B'M = 0$, i.e. the vectors comprising M are selected from the null of B'. $N(B')$ has dimension $(k - r)$ and M has $(k - r)$ columns. Hence given B, M is unique.

Example 6 From example 5,

$$M = \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}; \quad B = \begin{bmatrix} 2 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}.$$

Hence

$$B'M = \begin{bmatrix} 2 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 2 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

as is required by the relation $B'M = 0$. The nullity of B' is $4 - 2$ in this case and M has two columns. Thus the solution is unique. ■

3.5 The Frisch-Waugh Theorem

It often happens in econometric work that the raw data available need to be adjusted to make them applicable to the problem under examination. If the application in-

volves a linear least squares regression and if the adjustments required are linear, the question arises: is it appropriate to adjust the data before applying them to OLS regression, or should the raw data be applied directly to the regression, augmented to make allowances for the needed adjustments? This question is answered in the Frisch-Waugh Theorem which was ¯rst applied in relation to linear adjustment for seasonality.

Consider the model (1:4) re-written as

$$y = X_1 \bar{}_1 + X_2 \bar{}_2 + " \tag{3.37}$$

in which X_1 is the ¯rst k_1 columns of X and X_2 the remaining $(k ¡ k_1) = k_2$ columns. X_1 is to be thought of as the `observations' on each of the variables entering into the linear adjustments to be made, while $X_2$ represents n observations on each of the $k_2$ variables which are of economic signi¯cance in `explaining' y: Thus equation (3:37) is the regression augmented to include the required adjustment. For example, in terms of seasonality, y is the vector of raw data, that is, data that have not been de-seasonalized, X_1 represents the seasonal e®ects and X_2 the economic variables that theory predicts will explain the behavior of y. These latter observations are also in the form of raw data.

Now suppose y and X_2 are each adjusted for seasonality by ¯tting the following

arti¯cial regression by OLS:

$$y = X_1\mu_1 + u_1 \tag{3.38}$$

$$X_2 = X_1\pounds_2 + U_2. \tag{3.39}$$

In (3:38), μ_1 is $(k_1 \pounds 1)$ and u_1 is $(n \pounds 1)$; in (3:39) $\pounds_2$ is a $(k_1 \pounds k_2)$ matrix of coe± - cients and U_2 is $(n \pounds k_2)$. The OLS regressions of (3:38) and (3:39) are, respectively,

$$y = P_1 y + (I_n \text{ ¡ } P_1)y = X_1\hat{\mu}_1 + \hat{u}_1 \tag{3.40}$$

$$X_2 = P_1 X_2 + (I_n \text{ ¡ } P_1)X_2 = X_1\hat{\pounds}_2 + \hat{U}_2. \tag{3.41}$$

Here $P_i = X_i \left(X_i^{>} X_i\right)^{\text{¡ } 1} X_i^{>}$ $i = 1; 2$; $(I_n \text{ ¡ } P_i)$ is denoted M_i. Having adjusted the variables for seasons, the true economic e®ects, devoid of seasonal in°uences, may be determined via OLS regression obtained by premultiplying (3:37) by M_1:

$$M_1 y = M_1 X_2\bar{}_2 + v_1 \tag{3.42}$$

whereupon $\hat{X_2\bar{}_2} = \left(X_2^{>} M_1 X_2\right)^{\text{¡ } 1} X_2^{>} M_1 y$. This is precisely the estimate of $X_2\bar{}_2$ obtained by applying OLS directly to (3:37). Of course $X_2\bar{}_2 = F_2 y$, as indicated earlier. This simple result indicates that de-seasonalizing the variables linearly and then calculating the appropriate linear regression produces precisely the same estimate of $\bar{}_2$ as would have been produced by ¯tting (3:37) directly.

The Frisch-Waugh Theorem in fact provides the same answer via another route. The intuition here is to de-seasonalize y by ¯tting (3:38), yielding (3:40), and then

¯tting

$$M_1 y = X_{1,1} + X_{2,2} + v_2. \tag{3.43}$$

The inclusion of X_2 in (3:43) is justi¯ed on the ground that the matrix X_2 has not yet been de-seasonalized. Consequently $\hat{,}_2$ should re°ect only that part of X_2 which is orthogonal to $R[X_1]$ that is $M_1 X_2$. When estimated,

$$X_2 \hat{,}_2 = X_2 \left(X_2' M_1 X_2\right)^{-1} X_2' M_1 y$$

and thus $X_2 \hat{\beta}_2$ from (3:42) = $X_2 \hat{,}_2 = X_2 \hat{\beta}_2$ from (3:37) = $F_2 y$. Notice that in ¯tting (3:43),

$$\begin{aligned} X_1 \hat{,}_1 &= X_1 \left(X_1' M_2 X_1\right)^{-1} X_1' M_2 M_1 y \\ &= F_1 y - X_1 \left(X_1' M_2 X_1\right)^{-1} X_1' M_2 X_1 \left(X_1' X_1\right)^{-1} X_1' y \\ &= F_1 y - P_1 y \\ &= X_1 \left[\hat{\beta}_1 - \hat{\mu}_1\right]. \end{aligned}$$

Thus the ¯tted regression equation (3:43) is

$$y - X_1 \hat{\mu}_1 = X_1 \left[\hat{\beta}_1 - \hat{\mu}_1\right] + X_2 \hat{\beta}_2 + \hat{v}_2 \tag{3.44}$$

or using the projection matrices P_1, F_2 and P,

$$y - P_1 y = (F_1 - P_1) y + F_2 y + (I_n - P) y. \tag{3.45}$$

In (3:45), $\hat{v}_2$ of (3:44) has been written as $(I_n - P)y$. This may be seen from the following.

$$\begin{aligned} \hat{v}_2 &= (I_n - P)M_1 y \\ &= (I_n - P_1 - P + P_1)y \\ &= (I_n - P)y \\ &= \hat{u} \end{aligned}$$

$\hat{u}$ coming from the fitting of equation (3:37)), as indicated in (3:45).

The Frisch-Waugh Theorem (which will not be stated as a proper theorem) will be summarized as follows. If it is desired to estimate β_2 by applying OLS to equation (3:37), then the following alternative methods are equivalent.

1. Direct estimation of (3:37) by OLS.

2. Linear adjustment by OLS of y and X_2 as in equations (3:38) and (3:39), followed by OLS estimation of the linear regression of the adjusted y on the adjusted X_2 as given in (3:40) and (3:41).

3. Linear adjustment of y as in (3:42), followed by the OLS estimation of β_2 as in (3:44) or the estimation of γ_2 as in (3:43).

These calculations also reveal that:

a. in 3. above the estimate of β_1 in (3:43) is $\hat{\beta}_1$ from (3:37) as indicated in 1. above less $\hat{\mu}_1$ from the linear adjustment to y as given in 2. or 3. above;

b. the estimated residual from fitting (3:37) is precisely the estimated residual from fitting (3:42) and (3:43).

Thus the Frisch-Waugh Theorem reveals the convenience of projection matrices in making alternative OLS calculations.

3.6 Invariance

When OLS is applied, orthogonality is defined by the natural scalar product $(\cdot;\ \cdot)$. When GLS or GIVE is applied, the scalar product defining orthogonality is $(\cdot;\ A\cdot)$, A being a pd or nnd matrix. In terms of the model (1:4), OLS yields for $\mu = X\beta$ the estimate

$$\hat{\mu} = X\hat{\beta} = X\left(X^{\top}X\right)^{-1}X^{\top}y = Py \tag{3.46}$$

and P is orthogonal onto the range of X relative to $(\cdot;\ \cdot)$: The corresponding estimate for GLS is

$$\hat{\mu} = X\tilde{\beta} = X\left(X^{\top}AX\right)^{-1}X^{\top}Ay = Fy \tag{3.47}$$

and F is orthogonal onto the range of X relative to $\langle\cdot;\ \cdot\rangle$; moreover, A is such that the matrix $X^{\top}AX$ is invertible. The question arises: when are the two estimators

the same? The answer is provided by a theorem which has come to be known in econometrics as Kruskal's Theorem. This will be proved when A is pd. The theorem is stated as:

Theorem 6 Let $y = \beta + \varepsilon$ be a model in which y and ε lie in R^n and $\beta \in L$, a subspace of dimension $k < n$. On R^n is defined the natural scalar product $(\cdot;\cdot)$ and an alternative $\langle \cdot;\cdot \rangle = (\cdot; A\cdot)$, A being pd. The OLS estimate of β in L is the same as the GLS estimate of β in L if and only if the subspace L is invariant under A.

Proof. If L is invariant under A, then for every x in L, $Ax \in L$ and hence $AL \subseteq L$. Let L be invariant under A. Then for OLS, $(y - \hat{\beta}; Ax) = 0$ for every $x \in L$. But for GLS

$$0 = \langle y - \tilde{\beta}; x \rangle = (y - \tilde{\beta}; Ax) = (y - \tilde{\beta}; w)$$

for every $w \in L$. Since orthogonality ensures uniqueness of $\hat{\beta}$ and $\tilde{\beta}$ and each is such that

$$(y - \hat{\beta}; x) = (y - \tilde{\beta}; x) = 0$$

for all $x \in L$, $\hat{\beta} = \tilde{\beta}$.

Now let $\hat{\beta} = \tilde{\beta}$. Then

$$(y - \hat{\beta}; x) = (y - \tilde{\beta}; Ax) = 0$$

for all $x \in L$. Since $\hat{\beta} = \tilde{\beta}$, by uniqueness, $Ax \in L$ and L is invariant under A. ■

Example 7 Let $X = [e; x_2; x_3; ¢¢¢; x_k]$ and $A = \begin{pmatrix} I_n(1 - ½) + ee^{>}½ \end{pmatrix}$. Show that $L = R[X]$ is invariant under A. Notice that if $AX\mu$, $\mu \in R^k$ may always be written as $XM\mu$ for some non-singular matrix M of order k, then L must be invariant under A. For any $x \in L$, $x = X\mu$, $\mu \in R^k$. Thus $Ax = AX\mu = XM\mu = X_¸ \in L$. Now AX is given by

$$\begin{aligned} & \begin{pmatrix} I_n(1 - ½) + ee^{>}½ \end{pmatrix} [e; x_2; x_3; ¢¢¢; x_k] \\ = & \quad e\,f(1 - ½) + n½g : X(1 - ½) + [en\text{¾}_2½; en\text{¾}_3½; ¢¢¢; en\text{¾}_k½] \end{aligned}$$

where $X = [x_2; x_3; ¢¢¢; x_k]$. But the RHS of the last equality may be re-written

$$\begin{aligned} AX = & \quad e : X \begin{bmatrix} 1 + (n - 1)½ & n\text{¾}_2½; n\text{¾}_3½; ¢¢¢; n\text{¾}_k½ \\ 0 & I_{k-1}(1 - ½) \end{bmatrix} \\ = & \quad XM. \end{aligned}$$

Hence, $AX\mu = XM\mu = X_¸$ and $L = R[X]$ is invariant under A. ■

A useful theorem for proving invariance is

Theorem 7 If $L ½ R^n$ is invariant under the $(n \times n)$ matrix A, then $PAP = AP$ for every projection P on L. If $PAP = AP$ for some projection P on L, then L is invariant under A. ■

Example 8 From the previous example, note that $e \in L$. Hence $Pe = e$. In addition

$$\begin{aligned} PAP &= P\begin{bmatrix} I_n & (1-\alpha) + ee^{>} \end{bmatrix}\alpha P \\ &= \begin{bmatrix} P & (1-\alpha) + ee^{>} \end{bmatrix}\alpha P \\ &= \begin{bmatrix} I & (1-\alpha) + ee^{>} \end{bmatrix}\alpha P, \end{aligned}$$

which establishes invariance without the pain of manipulations. ■

When A is nnd, Kruskal's Theorem may be extended to include this condition. In Theorem 6, the condition of invariance is stated $AL \subseteq L$, meaning that AL implies L. If A is non-singular, then $AL = L$ and hence L is also invariant under A^{-1} i.e. $L = A^{-1}L$. When A is singular and L is invariant under A, then AL will lie in L but will not cover it, and so it is still correct to write $AL \subseteq L$.

4 The Multivariate Normal and Related Distributions

4.1 Probability and Random Variables

There is some mathematical difficulty in defining the concepts of probability and random variable. The conventional point of departure is a random experiment.

Definition 13 A process giving rise to observable outcomes is called a random experiment, denoted by E, if the following conditions are upheld.

1. All outcomes of E are known a priori.
2. In any particular performance of E, called a trial, the outcome cannot be known a priori.
3. E may be repeated under the same conditions. ■

Conditions 1–3 in Definition 13 may not suit every situation to which the concept of probability may apply. However, the definition of a random experiment is a starting point from which an expansion of ideas may take place.

The set of all possible outcomes of E forms a sample space denoted by S. The elements of S are called sample points or, more usually, elementary events. On the basis of the elementary events, more broadly based events comprising subsets of S may be formed. These subsets are called events in S and the way in which they are constructed is represented as an algebra called a σ-algebra.

Definition 14 Let F denote the class of subsets of S which have the following two properties.

1. $A \in F \Rightarrow A^c \in F$, A^c being the complement of A in S; that is, F is closed under complementation.

2. $A_i \in F$, $i = 1, 2, \ldots \Rightarrow \bigcup_{i=1}^{\infty} A_i \in F$; that is, F is closed under countable unions. The class F is known as a σ-algebra or a σ-field, and (S, F) is called a measurable space. $A \in F$ is said to be an F-measurable set. ■

One further element is required for a probability space. This is provided by:

Definition 15 Probability $P(\cdot) : F \to R$ is a set function defined on F satisfying:

1. $P(A) \geq 0$ for every A in F:

2. $P(S) = 1$;

3. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ for every sequence of disjoint events in F, that is for events having the property $A_i \cap A_j = \emptyset$, $i \neq j$.

The function $P(\cdot)$ is referred to as a probability measure. ■

Putting definitions 13–15 together permits:

Definition 16 Let a random experiment have sample space S and let S be endowed with a σ-field F and a probability measure $P(\cdot)$. The triple $(S, F, P(\cdot))$ constitutes a probability space. ■

The concept of a random variable may be developed from a probability space in the following way. A random variable $X(:)$ is an F-measurable function from S to the real line R, that is to say, for every $s \in S$, $X(s) = x \in R$. This raises a problem: while the events of the probability space are determined by the σ-field F defined on S, the mapping $X : S \to R$ takes points in S directly onto the real line, thereby 'avoiding' F. Clearly, the mapping $X : S \to R$ must be consistent with the definition of the σ-field F. Briefly, there must be defined on R a σ-algebra that is consistent with F. The σ-algebra on the real line is referred to as a Borel field and is denoted by B. The elements of B are Borel sets denoted by $B \in B$. This arrangement leads to the following definition.

Definition 17 A random variable $X(:)$ is an F-measurable function from S to R such that, for every Borel set $B \in B$ on R the set $X^{-1}(B) = \{s : X(s) \in B;\ s \in S\}$ satisfies $X^{-1}(B) \in F$. ■

In Definition 17, suppose that $S_0 \subset S$ is a subset of S such that $X(S_0) = B$. Since $S_0 \in F$ by definition, the statement $X^{-1}(B) \in F$ is merely a requirement that the inverse image of B forms a proper subset S_0 of S, as defined by the σ-field F.

Three points should be noted about the definition of a random variable. First, a random variable is defined for a specific σ-algebra, F. Second, if $X(:) : S \to R$ is to be a random variable, then the order of argument is from the Borel field B to the σ-field F, not the other way round. Finally, a random variable is, strictly

speaking, neither random nor a variable. $X(:)$ is a real-valued function and the terms random and variable do not enter its de¯nition. Randomness (or probability) enters the picture only after the de¯nition of the random variable $X(:)$ has been stated, as a means of completing the mathematical model induced by the de¯nition of X.

In discussing random variables, it is often necessary to discuss pairs of random variables, or triplets or n-tuples. It is then standard to refer to a random vector or a vector-valued random variable. Before a random vector can be de¯ned, it is necessary to introduce the concept of Borel sets of a ¯nite-dimensional Euclidean space V on which the scalar product $(:; :)$ is de¯ned. Let $kxk = (x; x)^{\frac{1}{2}}$. The open ball of radius r about the point z in V is de¯ned as $fx : kx ¡ zk < rg$. The open ball is used to de¯ne a Borel ¯eld on V, to stand in place of the Borel ¯eld on R used in de¯ning a random variable.

De¯nition 18 The Borel ¯eld of V, denoted by B_V, is the smallest ¾-algebra that contains all of the open balls. $(V; B_V)$ is a measurable space. ■

Note that B_V is independent of the scalar product of V. If two scalar products are de¯ned on V, these will generate the same Borel ¯eld B_V.

In de¯ning a random vector, it is necessary to have an appropriate probability space which, as before, may be regarded as the triple $(S; F; P(:))$. With this in place, the de¯nition of a random vector is stated as:

Definition 19 A random vector $X(:)$ in V is a function mapping S into V such that, for each Borel set $B \in B_V$, $X^{-1}(B) \in F$, $X^{-1}(B) = \{s : X(s) \in B; s \in S\}$. ■

In Definition 19, let $V = R^n$. Then the Borel set B may be regarded as comprising n subsets $B_1; B_2; ¢¢¢; B_n$ and $B_V = B \times B \times ¢¢¢ \times B = B^n$. Then $X^{-1}(B) = \{s : X_1(s) \in B_1; X_2(s) \in B_2; ¢¢¢; X_n(s) = B_n; s \in S\}$, where $X_1; X_2; ¢¢¢; X_n$ are the co-ordinates of X corresponding to the sample point s.

The space on which a random vector is defined is seldom of much interest and so the argument of a random vector is ordinarily suppressed. What is of interest is the distribution of X on V. Let X be defined on S to V where $(S; F; P(:))$ is the probability space. For each set $B \in B_V$, let $Q(B) = P(X^{-1}(B))$. Obviously $Q : B_V \to R$ is a probability measure on B_V, and $(V; B_V; Q)$ is also a probability space. Q is called the induced distribution of X, meaning that Q is induced by X and P. In summary, given the probability space $(S; F; P(:))$, a random vector defined by $X : S \to V$ such that $X^{-1}(B) = \{s : X(s) \in B; B \in B_V; s \in S\}$, then $Q = P(X^{-1}(B))$ on B_V is the induced distribution of X on B_V. It is also possible to proceed the other way round. Suppose it is given that Q is a probability measure on B_V (V being a finite dimensional vector space on which a scalar product is defined). Then there will exist a random variable X and a probability space $(S; F; P(:))$ such that Q is the induced distribution of X. The proof of this claim is achieved essentially by substitution. Setting $V = S$, $B_V = F$ and $P = Q$, let X be a random vector such that $X(v) = v$ for

$v \in V$. Then $P(X^{-1}(B)) = Q$ for each $B \in \mathcal{B}_V$ and Q is the induced distribution on $\mathcal{B}_V = F$.

Although Q is, properly speaking, the distribution induced by X and P on $\mathcal{B}_V$, it is unusual to use this phraseology. Rather, it is usual to say: "X is a random vector which ranges over V with distribution Q." This form will be used throughout the sequel.

A function $f : V \to W$ is called Borel measurable if the inverse image of each set $B \in \mathcal{B}_W$ is in $\mathcal{B}_V$, i.e. if $f^{-1}(B) = \{x : f(x) \in B;\ x \in V\} \in \mathcal{B}_V$, in which W is a metric space, $\mathcal{B}_W$ is a Borel field on W and $(W; \mathcal{B}_W)$ is a Borel measurable space. If f is Borel measurable, then $f(x)$ is a random vector in W. When f is continuous, it is Borel measurable. If $W = R$ and f is Borel measurable, then $f(X)$ is a real-valued random variable.

4.2 Expectations

Definition 20 *Let X be a random vector which ranges over V with distribution Q and let f be a real-valued, Borel-measurable function defined on V. The expectation of $f(X)$, assumed to be finite, is $Ef(X) = \int_V f(x)\,dQ$ the integral being a Lebesgue integral.* ■

Let $V = R^n$ on which is defined the natural scalar product. Let $Q(dx)$ denote the standard Lebesgue measure on R^n. If q is a non-negative function on R^n such that

the Riemann integral $\int_{R^n} q(x)\,dx = 1$, then q is called a density function. $Q(B) = \int_B q(x)\,dx$ is a probability measure on R^n implying that Q is the distribution of some random variable, say X. Let $e_1; e_2; ¢¢¢; e_n$ denote the orthonormal basis in R^n (whose sum is the equiangular vector). Then $(e_i; X)$ represents X_i the i'th co-ordinate of X. Assuming ¯nite expectations, $EX_i = \int_{R^n} (e_i; x)\, q(x)\,dx = {}^1{}_i$, the mean of X_i. The vector ^{1} comprises co-ordinates ${}^1{}_1; {}^1{}_2; ¢¢¢; {}^1{}_n$ and is the mean vector of X. Notice that for any non-random vector $z \in R^n$, $E(z; X) = E[\sum_i (z_i e_i; X)] = \sum_i [z_i E(e_i; X)] = \sum_i z_i {}^1{}_i = (z; {}^1)$ for all $z \in R^n$, implying that ^{1} is unique. This leads to the appealing equation $E(z; x) = (z; Ex)$ which is certainly valid in the co-ordinate case.

In referring to Euclidean subspaces, it is convenient to write the Euclidean vector space V with scalar product $(:; :)$ as $fV; (:; :)g$. This notation is used in:

Proposition 8 Let $X \in fV; (:; :)g$ with mean ^{1}. Let $Y \in fW; h:; :ig$, $! \in W$ and A be the matrix of a linear transformation $V \to W$. The random variable $Y = AX + !$ has expectation $A{}^1 + !$.

Proof. Let $z \in W$.

$$\begin{aligned} E\langle z, Y\rangle &= E\langle z, AX + \omega\rangle \\ &= E\langle z, AX\rangle + \langle z, \omega\rangle \\ &= E\left\langle A^{\top}z, X\right\rangle + \langle z, \omega\rangle \\ &= \left\langle A^{\top}z, \mu\right\rangle + \langle z, \omega\rangle \\ &= \langle z, A\mu\rangle + \langle z, \omega\rangle \\ &= \langle z, A\mu + \omega\rangle . \blacksquare \end{aligned}$$

In addition to the mean, the most important moment is the variance-covariance matrix of a random vector. Consider the random vector X in $f\mathbb{R}^n(\cdot, \cdot)g$. Using the basis matrix $I_n = [e_1, e_2, \cdots, e_n]$, then $(e_i, X) = X_i$. Let $E\left[(X_i - \mu_i)(X_j - \mu_j)\right] = \sigma_{ij}$, $i, j = 1, 2, \cdots, n$, and $\Sigma = [\sigma_{ij}]$. Let z and w lie in $\mathbb{R}^n$. Then

$$\begin{aligned} \mathrm{Cov}\{(z, X), (w, X)\} &= \mathrm{Cov}\left(\sum_i z_i X_i, \sum_j w_j X_j\right) \\ &= \sum_i \sum_j z_i w_j \,\mathrm{Cov}(X_i, X_j) \\ &= \sum_i \sum_j z_i w_j \sigma_{ij} \\ &= (z, \Sigma w). \end{aligned}$$

Alternatively, let $E\left[(X - \mu_1)(X - \mu_1)^\top\right] = \Sigma$. Then

$$\begin{aligned}
Cov\{(z; X); (w; X)\} &= Cov\{(z; X - \mu_1); (w; X - \mu_1)\} \\
&= E\{(z; X - \mu_1) \cdot (X - \mu_1; w)\} \\
&= E\left[z^\top (X - \mu_1)(X - \mu_1)^\top w\right] \\
&= z^\top \Sigma w \\
&= (z; \Sigma w).
\end{aligned}$$

The matrix Σ (which is nnd or pd) is called the variance-covariance matrix of X, or the dispersion of X relative to $(\cdot; \cdot)$. In place of $Cov\{(z; X); (w; X)\}$ could be written $D\{(z; X); (w; X)\}$. The notation $D(X) = \Sigma$ is also used.

Definition 21 The unique non-negative definite matrix Σ on $\{V; (\cdot; \cdot)\}$ to $\{V; (\cdot; \cdot)\}$ that satisfies

$$D\{(z; X); w; X\} = (z; \Sigma w)$$

is called the dispersion of X relative to $(\cdot; \cdot)$. ■

It should be emphasized that the dispersion depends on the scalar product specified. The next result shows how the dispersion changes as a function of the scalar product.

Proposition 9 *Let X be a random vector which ranges over $\{V; (\cdot;\cdot)\}$ with dispersion Σ. Let $\langle\cdot;\cdot\rangle$ be another scalar product given by $\langle\cdot;\cdot\rangle = (\cdot; A\cdot)$, A being a positive-definite matrix. The dispersion of X in $(V; \langle\cdot;\cdot\rangle)$ is ΣA.*

Proof. It is necessary to establish that

$$D\{\langle z; X\rangle; \langle w; X\rangle\} = \langle z; \Sigma A w\rangle$$

for all $z, w \in V$. Now

$$\begin{aligned} D\{\langle z; X\rangle; \langle w; X\rangle\} &= D\{(z; AX); (w; AX)\} \\ &= D\{(Az; X); (Aw; X)\} \\ &= (Az; \Sigma Aw) \\ &= (z; A\Sigma Aw) \\ &= \langle z; \Sigma A w\rangle. \blacksquare \end{aligned}$$

There are two consequences of Proposition 9. First, if $D(X)$ exists in one scalar product, it exists in all scalar products. Second, if $D(X) = \Sigma$ in $\{V; (\cdot;\cdot)\}$, then the dispersion relative to $\langle z; w\rangle = (z; \Sigma^{-1} w)$ is the identity.

Another result which is a generalization of Proposition 9 is:

Proposition 10 *Let X be a random vector which ranges over $\{V; (\cdot;\cdot)\}$. If A is the matrix of a linear transformation $V \to W$, where $\{W; \langle\cdot;\cdot\rangle\}$ is a Euclidean subspace, then $D(AX + \beta) = A\Sigma A^{\top}$ for all $\beta \in W$.*

Proof.

$$
\begin{aligned}
& D\{\langle z; AX + !\rangle; \langle w; AX + !\rangle\} \\
= \; & D\{\langle z; AX\rangle + \langle z; !\rangle; \langle w; AX\rangle + \langle w; !\rangle\} \\
= \; & D\{\langle z; AX\rangle; \langle w; AX\rangle\} \\
= \; & D\left(\left\langle A^{>}z; X\right\rangle; \left\langle A^{>}w; X\right\rangle\right) \\
= \; & \left\langle A^{>}z; \S A^{>}w\right\rangle \\
= \; & \left\langle z; A\S A^{>}w\right\rangle
\end{aligned}
$$

Thus

$$D\{AX + !\} = A\S A^{>}. \blacksquare$$

4.3 Special Covariance Structures

Relative to the Euclidean space $\{V; \langle \cdot; \cdot\rangle\}$, the group of orthogonal transformations $V \to V$ will be denoted by $O(V)$.

Definition 22 A random vector X which ranges over $\{V; \langle \cdot; \cdot\rangle\}$ with distribution Q has an orthogonally invariant, or spherical, distribution if its distribution and the distribution of MX are identical for all $B \in \mathcal{B}_V$ and $M \in O(V)$. $\blacksquare$

An implication of a spherical distribution is that the corresponding dispersion is $I\sigma^2$ for some $\sigma^2 > 0$. Thus if X is a random vector ranging over $\{V; \langle \cdot; \cdot\rangle\}$ with

spherical distribution, then $D(X) = I\sigma^2$ for some $\sigma^2 > 0$ and $D(MX) = D(X)$ for all $M \in O(V)$.

Definition 23 If $X \in \{V; (\cdot;\cdot)\}$ and

$$D(X) = I\sigma^2$$

for some $\sigma^2 > 0$, then X is said to have a weakly spherical distribution. This is equivalent to the condition $D(X) = D(MX)$ for all $M \in O(V)$. ■

The difference between a spherical and a weakly spherical distribution is that, in the former case, the whole distribution is invariant under orthogonal transformation whereas in the latter case, the dispersion is a positive multiple of the identity and this dispersion is invariant under orthogonal transformation as a consequence of Proposition 10.

4.4 The Multivariate Normal Distribution

In classical terminology, a probability distribution induced by a random variable or random vector is normally written as a distribution function (or cumulative density function). This was referred to as Q in section 4.1. If X is the random variable of interest with induced continuous distribution on R then $P(X \cdot x) = Q(x)$. The corresponding probability density function (pdf) in this case is $q(x)$ where $dQ(x) =$

$q(x)\,dx$. The interpretation of this last expression is $P\left[x - \frac{1}{2}dx \le X \le x + \frac{1}{2}dx\right] = q(x)\,dx$. Clearly $\int_{-\infty}^{x} q(t)\,dt = Q(x)$.

The multivariate normal distribution occupies a central place in the distribution theory now to be introduced. In elementary statistics, the univariate normal distribution is introduced as $X \sim N(\mu; \sigma^2)$, meaning that X is distributed normally with mean μ and variance σ^2. In this case, the pdf corresponding to $X \sim N(\mu; \sigma^2)$is

$$f(x) = \left[2\pi\sigma^2\right]^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{4.1}$$

Here the integral

$$\int_{-\infty}^{\infty} g(x)\,dx = \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \left[2\pi\sigma^2\right]^{\frac{1}{2}}$$

and hence, setting $f(x) = (2\pi\sigma^2)^{-\frac{1}{2}} g(x)$, the integral $\int_{-\infty}^{\infty} f(x)\,dx = 1$, as required. If a random sample of n is drawn from $N(\mu; \sigma^2)$, then the sample may be written as an n-tuple of mutually independent random variables $X_1; X_2; \cdots; X_n$. If A_i represents the event $\left\{x_i - \frac{1}{2}dx_i \le X_i \le x_i + \frac{1}{2}dx_i\right\}$, $i = 1; 2; \cdots; n$; then $P(A_1 \cap A_2 \cap \cdots \cap A_n)$ is given by

$$\left[2\pi\sigma^2\right]^{-\frac{n}{2}} \exp\left(-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}\right) dx_1\,dx_2\cdots dx_n. \tag{4.2}$$

If $I_n\sigma^2 = \Sigma; x^{\top} = [x_1; x_2; \cdots; x_n]$, then the pdf (4:2) may be expressed more compactly as

$$f(x)\,dx = (2\pi)^{-\frac{n}{2}} \det{}^{-\frac{1}{2}}\Sigma \exp\left(-\frac{1}{2}(x-e\mu)^{\top}\Sigma^{-1}(x-e\mu)\right) dx \tag{4.3}$$

where $f(x)$ now stands for $f(x_1)\,f(x_2)\cdots f(x_n)$ of (4:1) and dx for $dx_1;\, dx_2\cdots dx_n$. If $X_1; X_2; \cdots; X_n$ form a dependent set of normal random variables, such that for $i; j = 1; 2; \cdots; n$

$$\int_{R^n} x_i f(x)\,dx = EX_i = \mu_i$$

$$\int_{R^n} (x_i - \mu_i)(x_j - \mu_j) f(x)\,dx = E\left[(X_i - \mu_i)(X_j - \mu_j)\right] = \sigma_{ij},$$

then Σ in (4:3) becomes $\Sigma = [\sigma_{ij}]$ and e^{μ}, μ being scalar in (4:3), is replaced by the vector μ, defined by $\mu' = [\mu_1; \mu_2; \cdots; \mu_n]$. Then the pdf for n dependent X_i's is:

$$f(x) = (2\pi)^{-\frac{n}{2}} \det{}^{-\frac{1}{2}}\Sigma \exp -\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu). \tag{4.4}$$

In equation (4:4) Σ is presumed to be pd whereupon $\det\Sigma > 0$ and Σ^{-1} exists. In this case the random vector X has pdf according to (4:4) and this is indicated by writing $X \sim N(\mu; \Sigma)$.

Notice carefully that, in view of the positive definiteness of Σ; there will exist a non-singular $(n \times n)$ matrix A such that $A\Sigma A' = I_n$, $A'A = \Sigma^{-1}$. Let

$$Z = A(X - \mu) \tag{4.5}$$

or

$$X = A^{-1}Z + \mu. \tag{4.6}$$

The Jacobian of the transformation (4:6) is

$$\begin{aligned} |J| &= \text{mod det } \frac{\partial (X)}{\partial (Z)} \\ &= \text{mod det}^{-1} A. \end{aligned}$$

Thus, corresponding to (4:4), the pdf of the vector Z is

$$g(z) = (2\pi)^{-\frac{n}{2}} \det{}^{-\frac{1}{2}} \Sigma \exp\left[-\frac{1}{2} z^{\top} z\right] |J|.$$

But $\det{}^{-1} \Sigma = \det\left[A^{\top} A\right] = \det A^2 \Rightarrow \det{}^{-\frac{1}{2}} \Sigma = \text{mod det } A$. Thus

$$\det{}^{-\frac{1}{2}} \Sigma \, |J| = \text{mod det } A \; \text{mod det}^{-1} A = 1.$$

Hence

$$g(z) = 2\pi^{-\frac{n}{2}} \exp\left[-\frac{1}{2} z^{\top} z\right]. \tag{4.7}$$

Moreover from (4:5) $EZ = 0$ and $DZ = A \Sigma A^{\top} = I_n$. Thus Z in (4:7) obeys $Z \sim N(0; I_n)$ and (4:5) represents the standardization of X; that is, the transformation of $X \sim N(\mu; \Sigma)$ to $Z \sim N(0; I_n)$. Notice that, for the normal distribution, independence and zero covariance are synonymous, Usually, independence implies zero covariance but not vice-versa. Also notice that the affine transformation (4:5) or (4:6) (meaning that the hyperplane in which X lies has a different origin from the hyperplane in which Z lies because $\mu \neq 0$) allows inheritance of normality. This is true of any finite, affine transformation of a normal vector: an affine transformation of a normal variate is normal.

The standardized multivariate normal distribution of $Z \sim N(0; I_n)$ implies a set of n independent standardized normal variates $Z_i \sim N(0; 1)$ with $\text{cov}(Z_i; Z_j) = 0$ for all $i \neq j$. This is commonly written $Z_i \sim NID(0; 1)$ $i = 1; 2; \cdots; n$, NID meaning normally and independently distributed. The distribution of Z plays an important role when the dispersion of X is singular, that is Σ is nnd. In this case, $\det \Sigma = 0$ and Σ^{-1} does not exist. When this is the case, the distribution of X is said to be degenerate. Moreover, (4:4) is ill de¯ned. What can be made of a degenerate distribution? The secret to understanding degeneracy lies with the standardized normal distribution and a generalization of (4:6).

Consider m non-homogeneous linear functions of Z where $Z^{\top} = [Z_1; Z_2; \cdots; Z_n]$:

$$X = A^{\top}Z + \mu. \tag{4.8}$$

In equation (4:8) $A^{\top}$ is an $(m \times n)$ constant matrix, μ is an $(m \times 1)$ vector and X is an $(m \times 1)$ vector comprising elements $X_1; X_2; \cdots; X_m$. Since $Z \sim N(0; I_n)$, X has mean μ and dispersion $A^{\top}A = \Sigma$. If A has rank m, then $m < n$, and $A^{\top}A = \Sigma$ is $(m \times m)$ and positive de¯nite. In this case it is clearly in order to write $X \sim N(\mu; \Sigma)$. Now consider the case when the rank of A is $q < m$. In this case it is possible to show that there will exist a constant matrix $B^{\top}$ of order $(q \times m)$ such that $B^{\top}A^{\top}AB = I_q$. Setting $W = B^{\top}(x - \mu)$, W will have the $N(0; I_q)$ distribution. This implies that, if X has a degenerate distribution, in the sense that it has mean μ and dispersion Σ of rank $q < m$ (m being the number of random variables comprising X), then X is

distributed on a q-dimensional subspace of the m-dimensional subspace on which X is de¯ned.

Another degenerate case arises when $m > n$ and A has rank n. It is clear from (4:8) that $\S = A^{>}A$ with rank $n < m$ and order $(m \pounds m)$. Thus X has a degenerate distribution. Nevertheless $AA^{>}$ will be $(n \pounds n)$ and pd whereupon

$$\mathbf{i}_{AA^{>}}{}^{\mathbf{¢}_{i\ 1}} A(X \mathbf{¡}\ ^{1}) = Z \mathbf{»}\ N(0; I_n)$$

and so the underlying distribution upon which X is based lies on an n-dimensional hyperplane, whereas X itself, being an m-tuple with $m > n$, appears to lie in R^m.

Commonly with a linear hypothesis on R^n, $y \mathbf{»}\ N(^{1}; I_n¾^2)$. Here the dispersion is non-singular and $^{1} \mathbf{2}\ L$, a k-dimensional subspace of R^n. The estimate of ^{1} is $\mathbf{\hat{}}^{\mathbf{\wedge}} = Py$ where P is the orthogonal projection on L of rank $k < n$. How is Py distributed? Evidently like $N(^{1}; P¾^2)$, except that P is nnd and hence is singular. Thus Py has a degenerate distribution | not surprisingly, because $Py \mathbf{2}\ L$ and L has dimension k less than the number n of rows and columns of P. It is surely reasonable to expect that Py is distributed on L, even though it is expressed as if it is distributed in R^n (and is thereby degenerately distributed). In this case it is known that $P = BB^{>}$ where $B^{>}$ is the $(k \pounds n)$ partial isometry such that $B^{>}B = I_k$. Using $B^{>}$,

$$B^{>}(Py \mathbf{¡}\ ^{1}) = B^{>}(y \mathbf{¡}\ ^{1})$$

which has distribution $N(0; I_k¾^2)$. Thus Py has a degenerate distribution on R^n, but

a non-degenerate distribution on R^k as anticipated.

The notation $X \in R^n$ and $X \sim N(\mu; \Sigma)$ is taken to mean that there exists a $Z \in R^q$ where q is the rank of Σ, and an $(n \times q)$ matrix A of rank q, such that

$$X = AZ + \mu$$

with $Z \sim N(0; I_q)$. This implies that $\Sigma = AA^{\top}$ which is nnd. Under any linear transformation B, B being an $(m \times n)$ matrix of constants, if $X \sim N(\mu; \Sigma)$, $BX \sim N\left(B\mu; B\Sigma B^{\top}\right)$.

4.5 The Non-central Chi-square Distribution

Definition 24 If $X \in R^m$ has the $N(\mu; I_m)$ distribution, the random variable $X^{\top}X$ has a non-central chi-square distribution with m degrees of freedom and non-centrality parameter $\lambda = \mu^{\top}\mu$, written $\chi^2(m; \lambda)$. ■

The ordinary, or central, chi-square distribution is the special case of the non-central chi-square distribution when the non-centrality parameter is zero. If the distribution is said simply to be chi-square with m degrees of freedom, without specifying the non-centrality parameter, then it is understood that $\lambda = 0$. Notice that λ is a simple quadratic form in the mean of the present normal distribution i.e. $\lambda = \mu^{\top}\mu$. Some authors define the non-centrality parameter as $\sqrt{\mu^{\top}\mu}$ and others as $\left(\frac{\mu^{\top}\mu}{2}\right)$. The important point is that if $\mu \neq 0$, $\lambda \neq 0$.

The mean and variance of $X \sim \chi^2(m; \pm)$ are given by

$$EX = m + \pm$$

$$VX = 2m + 4\pm.$$

Thus if $\pm = 0$, the mean is m and the variance is $2m$.

From definition 24, it follows that if $Q_1 \sim \chi^2(m_1; \pm_1)$ independently of $Q_2 \sim \chi^2(m_2; \pm_2)$, then $Q_1 + Q_2 \sim \chi^2(m_1 + m_2; \pm_1 + \pm_2)$.

Theorem 11 If X ranges over R^m according to the $N(\mu; \Sigma)$ distribution, Σ being pd, then

$$(X - \mu)^\top \Sigma^{-1} (X - \mu) \sim \chi^2(m).$$

Proof. There always exists an $(m \times m)$ non-singular matrix Q such that $Q \Sigma Q^\top = I_m$, $Q^\top Q = \Sigma^{-1}$. Let $X \sim N(\mu; \Sigma)$ as indicated. Then $Z = Q(X - \mu) \sim N(0; I_m)$ implying that $Z^\top Z = (X - \mu)^\top Q^\top Q (X - \mu) \sim \chi^2(m)$. ■

Theorem 12 Let X range over R^m according to $N(0; I_m)$ and let A be a fixed, symmetric matrix of order $(m \times m)$ and rank q. A necessary and sufficient condition for $X^\top A X \sim \chi^2(q)$ is that $A = A^\top = A^2$.

Proof. Let $x \in R^m$ be a realization of X: Clearly $x^\top x \sim \chi^2(m)$ and

$$x^\top x = x^\top A x + x^\top (I_m - A) x.$$

Let M be an orthogonal matrix such that $MAM^{>} = D$, D being diagonal with q non-zero and $(m ¡ q)$ zero elements. Since $M^{i\,1} = M^{>}$, $MM^{>} = I_n$ and

$$x^{>}M^{>}AMx + x^{>}M^{>}(I_m ¡ A)Mx = x^{>}x$$

or

$$x^{>}Dx + x^{>}(I_m ¡ D)x = x^{>}x.$$

D has $(m ¡ q)$ zero diagonal elements, so the corresponding elements of $(I ¡ D)$ must each be unity. Hence the non-zero elements of D must also be unity. Since a necessary and su± cient condition for A to be symmetric and idempotent is that its characteristic values are unity or zero, A must be symmetric and idempotent, and hence so must $(I_m ¡ A)$.

On the other hand, assume $A = A^{>} = A^2$. Then there exists an orthogonal matrix M such that

$$x^{>}M^{>}AMx = x^{>}Dx = x_1^{>}x_1$$

in which x_1 is a $(q £ 1)$ subvector of x. But $x_1 » N(0; I_q)$. Hence $x_1^{>}x_1 » Â^2(q)$. ■

Theorem 13 Let $x » N(0; I_m)$. A necessary and su± cient condition for $x^{>}A_1x$ and $x^{>}A_2x$; of rank r_1 and r_2 respectively, to be independent $Â^2(r_1)$ and $Â^2(r_2)$ variates is that $A_1A_2 = 0$. ■

4.6 Non-central F-distribution

Definition 25 If Q_1 and Q_2 are independent random variables and $Q_1 \sim \chi^2(m_1; \lambda)$, $Q_2 \sim \chi^2(m_2)$, the distribution of the quotient $\frac{Q_1/m_1}{Q_2/m_2}$ is called the non-central F-distribution with m_1 and m_2 degrees of freedom and non-centrality parameter λ, written $F(m_1; m_2; \lambda)$. $F(m_1; m_2; 0)$ is the central F-distribution with m_1 and m_2 degrees of freedom, written $F(m_1; m_2)$. Occasionally, one encounters the doubly non-central F-distribution in which $Q_1 \sim \chi^2(m_1; \lambda_1)$ and $Q_2 \sim \chi^2(m_2; \lambda_2)$ independently. This is written $F(m_1; m_2; \lambda_1; \lambda_2)$. ■

4.7 Non-central t-distribution

Definition 26 If $Q_1 \sim \chi^2(1; \lambda_1)$ independently of $Q_2 \sim \chi^2(m_2)$ then $\left\{\frac{Q_1}{Q_2/m_2}\right\} \sim F(1; m_2; \lambda_1)$ and its square root $\left\{\frac{Q_1}{Q_2/m_2}\right\}^{\frac{1}{2}}$ has the non-central t-distribution with m_2 degrees of freedom and non-centrality parameter $\sqrt{\lambda_1}$. Clearly $t^2(m_2; \sqrt{\lambda_1}) = F(1; m_2; \lambda_1)$. ■

4.8 Cochran's Theorem

Theorem 14 Let $X \sim N(0; I_m)$ and let $X^\top X = \sum_{i=1}^{i=k} X^\top A_i X$, the A_i being fixed matrices of rank r_i. The following conditions are equivalent (i.e. any one implies the

other two):

$$\sum_{i=1}^{i=k} r_i = m;$$

Each $X^{>} A_i X \,\hat{\text{A}}\sim\, \hat{\text{A}}^2(r_i)$;

The $X^{>} A_i X$ are mutually independent. ■

Cochran's theorem is stated without proof. It is the central distribution theorem for the linear hypothesis.

5 Gauss-Markov Estimation and the Linear Hypothesis

5.1 The Gauss-Markov Theorem

The model to be considered is of a random vector y which ranges over R^n on which the natural scalar product (:; :) has been de¯ned. For ¯xed w and z in R^n,

$$E(w; y) = (w; \eta) \qquad \eta \in L \subset R^n; \dim L = k < n; \tag{5.1}$$

$$D[(w; y); (z; y)] = (w; z)\sigma^2. \tag{5.2}$$

Thus $E\mathbf{y} = \eta \in L$, a k-dimensional subspace of R^n, and $D\mathbf{y} = I_n\sigma^2$. A basis for L is available in the form of an $(n \times k)$ matrix X of rank k, and so η may be expressed as $\eta = \eta(\beta) = X\beta$, where β is a $(k \times 1)$ vector which lies in some subset of R^k. This subset of R^k is easily shown to be $R(X')$. For suppose the rank of X is $r \leq k$, then in R^k there are two orthogonal subspaces: $R(X')$ of dimension r and $R(X')^{\perp} = N[X]$ of dimension $(k - r)$. Thus β may be uniquely decomposed as

$$\beta = \beta_0 + (\beta - \beta_0)$$

with $\beta_0 \in R(X')$ and $(\beta - \beta_0) \in N[X]$. Now

$$X\beta = X\beta_0 + 0.$$

because $N[X] = \{\mu : X\mu = 0; \mu \in R^k\}$. Thus $X\beta = X\beta_0$. Moreover, if $r = k$, $N[X] = 0$; whereupon $\beta = \beta_0 \in R(X')$. In the model described above, $r = k$ and hence

$\bar{} \in R\left(X^{>}\right)$. Note also that

$$R\left(X^{>}\right) = R\left(X^{>}X\right) = R\left(\left(X^{>}X\right)^{i\ 1}\right).$$

The model $y \sim (^1; I_n\frac{3}{4}^2)$; $^1 \in L$, may also be written as

$$y = X\bar{} + "; \qquad E" = 0; D" = I_n\frac{3}{4}^2, \tag{5.3}$$

with the X matrix considered as a $\bar{}$xed basis of L. In particular, for any $x \in L$, $E(x; ") = E(x; y \ ¡ \ ^1) = (x; 0) = 0$.

It is convenient and useful to consider the estimation of ^{1} in terms of the parameter vector $\bar{}$. Estimation of $\bar{}$ is considered in terms of linear combinations of its elements.

De$\bar{}$nition 27 A parametric function is a linear function of unknown parameters. If the unknown parameters are represented by the vector $\bar{}$ in (5:3), then for a $\bar{}$xed vector $c \in R^k$; $^a = (c; \bar{})$ is a parametric function. ■

The estimability of ^{a} is de$\bar{}$ned in terms of unbiased linear estimators.

De$\bar{}$nition 28 The parametric function $^a = (c; \bar{})$ is said to be estimable if there exists a $\bar{}$xed vector $a \in R^n$ such that the linear estimator $(a; y)$ is unbiased for ^{a}; that is if $E(a; y) = (c; \bar{})$. ■

Theorem 15 The parameter function ^{a} is estimable if and only if there exists a $\bar{}$xed vector $a \in R^n$ such that $c^{>} = a^{>}X$.

Proof. If a is estimable, then

$$E(a; y) = (c; \beta). \tag{5.4}$$

But $E(a; y) = (a; \theta)$ and $\theta = X\beta$. Hence

$$a^{>} X\beta = c^{>}\beta \tag{5.5}$$

must hold identically in β implying $a^{>}X = c^{>}$. If $c^{>} = a^{>}X$, then (5:5) holds, which may be written as (5:4). Thus a is estimable if and only if $c^{>} = a^{>}X$. ■

Lemma 16 If a is estimable, then there exists a unique, unbiased linear estimator $(a^{¤}; y)$ with $a^{¤} \in L$.

Proof. If a is estimable, there exists a vector a such that $E(a; y) = (c; \beta)$ and $c^{>} = a^{>}X$. Let $a = a^{¤} + (a - a^{¤})$ with $a^{¤} \in L$; $(a - a^{¤}) \in L^{?}$. Then

$$(a; y) = (a^{¤}; y) + (a - a^{¤}; y).$$

Taking expectations,

$$(a; \theta) = (a^{¤}; \theta) + 0$$

because $(a - a^{¤}) \perp L$ and $\theta \in L$. Thus $E(a; \theta) = E(a^{¤}; \theta) = (c; \beta)$.

To prove uniqueness, let the vector $® \in L$ satisfy $E(®; y) = (c; \beta)$. Then $0 = E(a^{¤}; y) - E(®; y) = (a^{¤} - ®; \theta)$ implying, since $\theta \in L$, that $(a^{¤} - ®) \in L^{?}$. But $a^{¤}$ and $®$ both lie in L by construction. Thus $(a^{¤} - ®)$ lies in both L and $L^{?}$, implying $(a^{¤} - ®) = 0$, that is $a^{¤} = ®$ and $a^{¤}$ is shown to be unique. ■

Theorem 17 (The Gauss-Markov Theorem) Under the assumptions of model (5:1), every estimable function a has a unique linear estimator which has minimum variance in the class of unbiased linear estimators. The unique estimator is $\langle a; \hat{\beta} \rangle = \langle c; \hat{\mu} \rangle$ where $\hat{\beta}$ is the least squares estimator of β in (5:3) and $\hat{\mu} = X\hat{\beta}$.

Proof. Let $(a; y)$ be an unbiased linear estimator of a, then $a = a^{\circ} + (a - a^{\circ})$; $a^{\circ} \in L$; $(a - a^{\circ}) \in L^{\perp}$. Now $V(a; y) = \|a\|^2\sigma^2$ and

$$\|a\|^2\sigma^2 = \|a^{\circ}\|^2\sigma^2 + \|a - a^{\circ}\|^2\sigma^2.$$

Since $\|a - a^{\circ}\|^2\sigma^2 \geq 0$ and $V(a^{\circ}; y) = \|a^{\circ}\|^2\sigma^2$,

$$V(a; y) \geq V(a^{\circ}; y).$$

But $(a; y)$ is any unbiased linear estimator a while $(a^{\circ}; y)$ is unique. Moreover, if P is the orthogonal projection matrix on L, $(a^{\circ}; y) = (Pa; y) = (a; Py) = \langle a; \hat{\mu} \rangle = \langle X'a; \hat{\beta} \rangle = \langle c; \hat{\beta} \rangle$, in view of theorem 15. This proves the Theorem. ■

A minimum variance unbiased linear estimator is often referred to as a best linear unbiased estimator or a BLUE.

5.2 The Linear Hypothesis

5.2.1 The Setting

To test a linear hypothesis, a distributional assumption is required. This is done by augmenting (5:3) with the maintained hypothesis

$$H : y \sim N\left(\mu(\theta); I_n \sigma^2\right); \tag{5.6}$$

$\mu(\theta) \in L$ as is required under (5:3). The null hypothesis to be tested may be written as r linearly independent equations

$$H_0 : (a_i; \mu(\theta)) = 0 \qquad i = 1; 2; \ldots; r; \; r < k; \; \mu(\theta) \in L. \tag{5.7}$$

Let A be the matrix $[a_1; a_2; \ldots; a_r]$ which, by the linear independence of the a_i, is of rank r. Thus (5:7) may be re-written

$$H_0 : A^{\top}\mu(\theta) = 0; \qquad \mu(\theta) \in L, \tag{5.8}$$

whereupon it is clear that, if $L_0 = L \setminus N\left(A^{\top}\right)$, then

$$H_0 : \mu \in L_0 \subset L, \tag{5.9}$$

L_0 being of dimension $(k - r)$.

The statement that $\mu \in L$ is a linear hypothesis because L is a linear subspace. Similarly H_0 is a linear hypothesis because $L_0 \subset L$ is a linear subspace.

In (5:7){(5:9), H_0 is formulated as an hypothesis concerning μ as a function of the parameter vector θ. The same hypothesis may also be expressed simply in terms

of β. As has been demonstrated in chapter 3, the a_i and hence the matrix A are not unique. What is unique is the matrix $PA = X\left(X^{\top}X\right)^{-1}B$, where B is a $(k \times r)$ matrix of rank r such that

$$H_0 : B^{\top}\beta = 0 \qquad \beta \in R\left(X^{\top}\right) \tag{5.10}$$

is consistent with (5:7){(5:9).

The formulations of H_0 in (5:7){(5:9) are in restraint form. In (5:10), the implied subspace in which β must lie on H_0 in $R\left(X^{\top}\right) \cap N\left(B^{\top}\right)$, having dimension $(k - r)$. Moreover, $\dim N\left(B^{\top}\right) = (k - r)$. Let a basis of $N\left(B^{\top}\right)$ in $R\left(X^{\top}\right)$ be M a $(k \times (k - r))$ matrix satisfying

$$B^{\top}M = 0.$$

It follows that if θ is a $((k - r) \times 1)$ vector satisfying

$$\beta = M\theta \tag{5.11}$$

and $\beta \in R\left(X^{\top}\right) \cap N\left(B^{\top}\right)$; then (5:11) represents the freedom equation form of H_0 which appears in restraint form in (5:10).

The setting deβned by (5:6) and one of (5:7){(5:11) covers all of the known linear tests in econometrics, including the a± ne hypothesis H_0^{α} :

$$H_0^{\alpha} : B^{\top}\beta = \mu \qquad \mu \neq 0; \beta \in R\left(X^{\top}\right). \tag{5.12}$$

Since $B^{>}$ has rank $r < k$, with r rows and k columns, it is clear that (5:12) has more than one solution. If any one of these is $¸$ then $B^{>}¸ = µ$ and

$$B^{>}(¯ ¡ ¸) = 0. \tag{5.13}$$

The maintained hypothesis in (5:6) may now be re-formulated, using $y^{¤} = y ¡ X¸$ and $¹^{¤} = X(¯ ¡ ¸)$, as

$$H^{¤} : y^{¤} » N\left(¹^{¤}; I_n¾^2\right) \qquad ¹^{¤} 2 L. \tag{5.14}$$

If (5:14) replaces (5:6) and H_0 in (5:10) is replaced by

$$H_0^{¤} : B^{>}¯^{¤} = 0 \tag{5.15}$$

where $¯^{¤} = (¯ ¡ ¸)$, then (5:6) and the a± ne hypothesis (5:12) has been accommodated into the linear hypothesis framework, (5:14) corresponding to (5:6), (5:12) corresponding to (5:10).

5.2.2 Decomposition

From chapter 3, the natural decomposition of R^n is

$$R^n = L_0 © \left(L_0^? \setminus L\right) © L^? \tag{5.16}$$

with $L = L_0 © \left(L_0^? \setminus L\right)$. The orthogonal projection from R^n on L_0 is P_0 and $PP_0 = P_0P = P_0$; also $(I_n ¡ P)(I_n ¡ P_0) = (I_n ¡ P_0)(I_n ¡ P) = (I_n ¡ P); P(I_n ¡ P) =$

$(I_n - P)P = 0$, $P_0(I_n - P_0) = (I_n - P_0)P_0 = 0$, and $P(I_n - P_0) = (I_n - P_0)P = P - P_0$. Corresponding to the decomposition (5:16),

$$I_n = P_0 + (P - P_0) + (I_n - P), \tag{5.17}$$

$(P - P_0) = PA\left(A^{>}PA\right)^{-1}A^{>}P = X\left(X^{>}X\right)^{-1}B\left[B^{>}X^{>}X\left(X^{>}X\right)^{-1}B\right]^{-1}B^{>}X^{>}X\left(X^{>}X\right)^{-1}X^{>}$ from chapter 3, with $P_0 = XM\left(M^{>}X^{>}XM\right)^{-1}M^{>}X^{>}$, M satisfying $B^{>}M = 0$ for the matrix M of order $(k \times (k - r))$, from equation (5:11). It follows from (5:6) and (5:17) that

$$\frac{\|y\|^2}{\sigma^2} = \frac{\|P_0 y\|^2}{\sigma^2} + \frac{\|(P - P_0)y\|^2}{\sigma^2} + \frac{\|(I_n - P)y\|^2}{\sigma^2}. \tag{5.18}$$

Moreover $\frac{\|y\|^2}{\sigma^2} \sim \chi^2(n; \delta)$; $\delta = \frac{\beta^{>}\beta}{\sigma^2}$. Since rank $P_0 = (k - r)$, rank $(P - P_0) = r$ and rank $(I_n - P) = (n - k)$, the sum of the right hand side ranks equals the left hand side rank. Hence, by Cochran's theorem, the three right-hand side elements of (5:18) are independent χ^2-variates.

Consider the last two terms on the right-hand side of (5:18) in regard to testing H_0 against the alternative hypothesis $H_a : A^{>}\beta \neq 0;\ \beta \in L$, or $B^{>}\gamma \neq 0;\ \gamma \in R\left[X^{>}\right]$, taking H as established. On H_a

$$\frac{\|(P - P_0)y\|^2}{\sigma^2} \sim \chi^2(r; \delta); \qquad \delta = \frac{\beta^{>}(P - P_0)\beta}{\sigma^2}.$$

On H_0, the same element has the same distribution except that $\delta = \frac{\beta^{>}(P - P_0)\beta}{\sigma^2} = 0$ since $\beta \in L_0 \subset L$. With regard to the last component of (5:18),

$$\frac{\|(I_n - P)y\|^2}{\sigma^2} \sim \chi^2(n - k; 0)$$

since $(I_n - P)\mu = 0$ on H and on H_0. It follows from this discussion that

$$F = \frac{k(P - P_0)yk^2}{k(I_n - P)yk^2} : \frac{n - k}{r} \tag{5.19}$$

has the $F(r; n - k)$-distribution, central on H_0, non-central on H_a. Notice that both H_0 and H_a are consistent with H. In (5:19), the evidence that y has a `large' component in $\mathcal{L}_0^? \setminus \mathcal{L}^{\perp}$ relative to its component in $\mathcal{L}^?$ is taken as counter to $H_0$, whereupon $F$ will be `large'; and the evidence that y in $\mathcal{L}_0^? \setminus \mathcal{L}^{\perp}$ is `small' relative to its component in $\mathcal{L}^?$ is taken as counter to $H_a$ and hence not counter to $H_0$, whereupon $F$ will be `small' and H_0 `acceptable' as a working hypothesis.

5.2.3 Alternative Forms of the F-test

Various alternative expressions for F in (5:19) may be developed.

$$\begin{aligned} F &= \frac{y^{\top}(P - P_0)y}{y^{\top}(I_n - P)y} : \frac{(n - k)}{r} \\ &= \frac{y^{\top}(I_n - P_0)y - y^{\top}(I_n - P)y}{y^{\top}(I_n - P)y} : \frac{(n - k)}{r} \end{aligned} \tag{5.20}$$

whereupon, if $\hat{\varepsilon} = (I_n - P)y$ and $\tilde{\varepsilon} = (I_n - P_0)y$, the unrestricted and restricted estimators of ε respectively, then

$$F = \frac{\tilde{\varepsilon}^{\top}\tilde{\varepsilon} - \hat{\varepsilon}^{\top}\hat{\varepsilon}}{\hat{\varepsilon}^{\top}\hat{\varepsilon}} : \frac{(n - k)}{r}. \tag{5.21}$$

Equation (5:21) is easily calculated directly from a computer output by noting the restricted and unrestricted sums of squared errors.

The next expression for F is developed from the known result

$$\begin{aligned}(P - P_0) &= PA\left[A^{>}PA\right]^{-1}A^{>}P \\ &= X\left[X^{>}X\right]^{-1}B\left\{B^{>}\left[X^{>}X\right]^{-1}B\right\}^{-1}B^{>}\left[X^{>}X\right]^{-1}X^{>}.\end{aligned}$$

Substituting the last into (5:20) and noting that $\hat{\sigma}^2 = \frac{y^{>}(I_n - P)y}{(n-k)}$ is an unbiased estimator of σ^2 because $E\hat{\sigma}^2 = \frac{E\,tr[y^{>}(I_n - P)y]}{(n-k)} = \frac{E\,tr[\varepsilon^{>}(I_n - P)\varepsilon]}{(n-k)} = \frac{E\,tr[(I_n - P)\varepsilon\varepsilon^{>}]}{(n-k)} = \frac{tr(I_n - P)I_n\sigma^2}{(n-k)} = \frac{(n-k)\sigma^2}{(n-k)} = \sigma^2$, then

$$\begin{aligned}rF &= \left[B^{>}A'\right]\left\{B^{>}\left[X^{>}X\right]^{-1}\sigma^2 B\right\}^{-1}B^{>}A' \\ &= \left[B^{>}\sqrt{n}A'\right]\left\{B^{>}\left(\frac{X^{>}X}{n}\right)^{-1}B\right\}\left[B^{>}\sqrt{n}A'\right].\end{aligned} \tag{5.22}$$

This last expression is the large-sample Wald statistic for H_0 against H_a in circumstances that do not require X to be fixed; merely that $\frac{X^{>}X}{n}$ converges in probability to a positive-definite matrix of finite elements, say M_{XX}, that $\frac{X^{>}\varepsilon}{n} \xrightarrow{p} 0$ and that, on H_0, $B^{>}\sqrt{n}A'$ converges in distribution to $N\left[0; B^{>}M_{XX}^{-1}B\sigma^2\right]$. As will be determined later, such conditions are not unreasonable for some economic observations when n is `large'. But there is no need to call upon large-sample theory in the case of model (5:6) since

$$F = r^{-1}\left[B^{>}A'\right]\left\{\hat{D}\left[B^{>}A'\right]\right\}^{-1}B^{>}A' \tag{5.23}$$

where $\hat{D}\left[B^{>}A'\right]$ is the estimated dispersion of $B^{>}A'$ on H_0, that is $B^{>}\left[X^{>}X\right]^{-1}\hat{\sigma}^2 B$. In view of the connection to (5:22), (5:23) may be regarded as a small-sample special-

ization of the Wald statistic. In fact, the test is based on the Wald Principle which is a principle for developing statistics to test restrictions on a parameter vector.

The Wald Principle is quite general. Let μ be a vector-valued parameter and $\hat{\mu}_n$ a consistent unrestricted estimator of it, based on a random sample of n observations. As is commonly the case for sufficiently large n, $\sqrt{n}\left(\hat{\mu}_n - \mu\right) \sim N(0, \Sigma)$, Σ being a positive definite-dispersion. The null hypothesis, $h(\mu) = 0$, is an $(r \times 1)$ vector of functions $h_i(\mu)$; $i = 1, 2, \ldots, r$. When n is large and there is sufficient regularity, $\sqrt{n}\left[h\left(\hat{\mu}_n\right) - h(\mu)\right] \sim N\left[0, H_\mu' \Sigma H_\mu\right]$ where $H_\mu = \frac{\partial h'(\mu)}{\partial \mu}$. Then

$$\sqrt{n}h'\left(\hat{\mu}_n\right)\left[H_\mu' \Sigma H_\mu\right]^{-1} h\left(\hat{\mu}_n\right)\sqrt{n} \sim \chi^2(r, \lambda)$$

with $\lambda = 0$ on the null hypothesis: $h(\mu) = 0$. The null hypothesis may be tested using the statistic

$$W = nh'\left(\hat{\mu}_n\right)\left[H_{\hat{\mu}_n}' \hat{\Sigma} H_{\hat{\mu}_n}\right]^{-1} h\left(\hat{\mu}_n\right) \sim \chi^2(r, \lambda)$$

because $H_{\hat{\mu}_n}$ and $\hat{\Sigma}$ will be `close' to H_μ and Σ, for large n. In the case of the linear hypothesis examined above, (5:21) gives the form of W with $\sqrt{n}h\left(\hat{\mu}_n\right) = \sqrt{n}B'\hat{\Lambda}$ and $B'\left[\frac{X'X}{n}\right]^{-1} B = H_{\hat{\mu}_n}' \hat{\Sigma} H_{\hat{\mu}_n}$. In all cases, only unrestricted estimates are applied to determine W. Thus the Wald Principle is to be seen as a principle for developing tests of general hypotheses, of the kind $h(\mu) = 0$, on the basis of unrestricted, asymptotically normal estimators $\hat{\mu}_n$ of μ, using a standardized quadratic form in $h\left(\hat{\mu}_n\right)$.

There are now three forms of the F-test of (5:18) for a linear hypothesis: (5:20), (5:21) and (5:23). The form (5:20) is a difference-in-regression formulation: $\frac{y^{>}(P - P_0)y}{r\hat{\sigma}^2}$; (5:21) is a difference-in-residual-sum-of-squares formulation: $\frac{\hat{u}^{>}\hat{u} - \hat{\hat{u}}^{>}\hat{\hat{u}}}{r\hat{\sigma}^2}$; and (5:23) is the Wald formulation based on unrestricted estimation of β and σ^2. All three forms are identical. A fourth form is an equivalent test involving a different distribution. This comes through the following result.

Theorem 18 Let y be a random vector ranging over R^n according to (5:6). Then a test for H_0 in (5:8) is $F = \frac{y^{>}(P - P_0)y}{r\hat{\sigma}^2}$ where $\hat{\sigma}^2 = \frac{y^{>}(I_n - P)y}{(n - k)}$, P being the orthogonal projection on L and P_0 being the orthogonal projection $L_0 \subset L \subset R^n$, L and L_0 having dimensions k and $(k - r)$. A test equivalent to F is $M = \frac{y^{>}(P - P_0)y}{y^{>}(P - P_0)y + y^{>}(I_n - P_0)y}$ or

$$M = \frac{y^{>}(P - P_0)y}{y^{>}(I_n - P_0)y}. \tag{5.24}$$

This has the $\beta\left(\frac{r}{2}; \frac{n-k}{2}\right)$ distribution, central on H_0, non-central otherwise. ■

Theorem 18 is stated without proof. Both the F- and M-tests are exact, small-sample tests; one is just another form of the other. Thus F and M will never yield conflicting results. As will now be demonstrated, M is a test based on the Lagrange Multiplier Principle of testing restrictions.

5.2.4 The Lagrange Multiplier Principle

Consider the Lagrange expression

$$L(\theta; \lambda) = (y - \theta)'(y - \theta) + 2\theta' A\lambda; \qquad \theta \in \Theta$$

in which 2λ is the Lagrange Multiplier enforcing the restriction $A'\theta = 0$. The first-order conditions for minimizing $(y - \theta)'(y - \theta)$ subject to $A'\theta = 0; \theta \in \Theta$, are the stationary conditions for $L(\theta; \lambda)$:

$$-y + \theta + A\tilde{\lambda} = 0$$

$$A'\theta = 0$$

where $\tilde{\lambda}$ and θ are the restricted estimates of θ and λ. Since $\theta \in \Theta$ and $\theta \in \Theta_0 \subset \Theta$

$$Py - \theta = PA\tilde{\lambda} \tag{5.25}$$

$$A'P\theta = (PA)'\theta = 0. \tag{5.26}$$

Premultiplying equation (5:25) by A' and noting that $A'\theta = A'P\theta = 0$, then $A'Py = A'PA\tilde{\lambda}$ and

$$\tilde{\lambda} = [A'PA]^{-1} A'Py. \tag{5.27}$$

Also note that, since we know that $\theta = P_0 y$, (5:25) may also be written

$$(P - P_0)y = PA\tilde{\lambda}$$

implying, from (5:27), that

$$(P - P_0)y = PA\left(A^{>}PA\right)^{-1}A^{>}Py$$

as is to be expected. From (5:27)

$$\tilde{\gamma} \sim N\left[\left(A^{>}PA\right)^{-1}A^{>}\iota;\ \left(A^{>}PA\right)^{-1}\sigma^2\right]$$

which on H_0 is $N\left[0;\ \left(A^{>}PA\right)^{-1}\sigma^2\right]$ since then $A^{>}\iota = 0$. Hence

$$\frac{y^{>}PA\left(A^{>}PA\right)^{-1}A^{>}PA\left(A^{>}PA\right)^{-1}A^{>}Py}{\sigma^2} \sim \chi^2(r;\pm).$$

Substituting in the estimator of σ^2 on H_0

$$\sigma^2 = \frac{y^{>}(I_n - P_0)y}{(n - k + r)}$$

there emerges

$$\begin{aligned} LM &= \tilde{\gamma}^{>}\left[D\left(\tilde{\gamma}\right)\right]^{-1}\tilde{\gamma} \\ &= \frac{y^{>}(P - P_0)y}{\frac{y^{>}(I_n - P_0)y}{(n-k+r)}} \end{aligned}$$

which is the large-sample Lagrange Multiplier statistic for testing H_0. Moreover

$$\frac{LM}{(n - k + r)} = M$$

and M is seen to be a small-sample specialization of the Lagrange Multiplier statistic.

5.3 Applications of the F-test

In considering various examples of the F-test, it is worth noting that, of the various expressions (5:20), (5:21) and (5:23), it is usually (5:23) that is easiest to apply to ¯nd speci¯c theoretical results, while (5:21) is the most practical form. Sometimes it is bene¯cial to adopt the freedom-equation approach.

5.3.1 Testing the Di®erence Between Two Means

The situation to be examined is that of two independent samples of n observations, in the $(n \pounds 1)$ vector y_1, and m observations in the $(m \pounds 1)$ vector y_2. The maintained hypothesis is written in terms if $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ an $((n + m) \pounds 1)$ vector. The common population mean of the ¯rst sample is $\bar{}_1$ and its variance is $¾^2$; the common population mean and variance for the second sample are $\bar{}_2$ and $¾^2$. Thus the variances are the same. Under H

$$y \gg N\left(\iota_1; I_{n+m}¾^2\right)$$

where $\iota \; 2 \; L$ and L has basis

$$E = \begin{bmatrix} e_n & 0 \\ 0 & e_m \end{bmatrix}$$

e_n being the equiangular vector on R^n and e_m being the equiangular vector on R^m. Thus L = R [E] and

$$\iota = E\bar{}$$

where $\beta^{>} = [\beta_1 \quad \beta_2]$; β being $(2 \pounds 1)$. In this formulation, then, the X-matrix is replaced by E. It is desired to test

$$H_0 : \beta_1 = \beta_2 \;) \; [1 \quad \text{¡} \; 1]\beta = 0.$$

Thus $B^{>} = [1 \quad \text{¡} \; 1]$ and $r = 1$. On H_0, $\beta_1 = \beta_2 = \beta$ and

$$\mu = e_{n+m}\beta.$$

Let $P_n = e_n \left(e_n^{>} e_n\right)^{\text{¡}\,1} e_n^{>}$ with P_m the same using e_m; using e_{n+m}, the orthogonal projection onto its range is P_0. Thus

$$P = \begin{bmatrix} P_n & 0 \\ 0 & P_m \end{bmatrix}$$

and

$$F = \frac{y^{>}(P \text{¡} P_0)y}{y^{>}(I_n \text{¡} P)y} : \frac{n+m \text{¡} 2}{1}.$$

$$y^{>}(P \text{¡} P_0)y = n(\bar{y}_1 \text{¡} \bar{y})^2 + m(\bar{y}_2 \text{¡} \bar{y})^2$$

$$= n\bar{y}_1^2 + m\bar{y}_2^2 \text{¡} (n+m)\bar{y}^2$$

where $\bar{y}_1, \bar{y}_2$ are the sample means and $(n+m)\bar{y} = n\bar{y}_1 + m\bar{y}_2$. Hence

$$(n+m)\bar{y}^2 = \frac{n^2\bar{y}_1^2 + m^2\bar{y}_2^2 + 2nm\bar{y}_1\bar{y}_2}{n+m}$$

and

$$y^{>}(P \text{¡} P_0)y = \frac{nm(\bar{y}_1 \text{¡} \bar{y}_2)^2}{n+m}.$$

Now

$$\frac{y^{>}(I_{n+m} - P)y}{n+m-2} = \frac{(n-1)\hat{\sigma}_1^2 + (m-1)\hat{\sigma}_2^2}{(n+m-2)}$$
$$= \hat{\sigma}^2.$$

It follows that

$$F = \frac{nm(\bar{y}_1 - \bar{y}_2)^2}{\frac{n+m}{\sigma^2}}$$
$$= \frac{(\bar{y}_1 - \bar{y}_2)^2}{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}.$$

$F \sim F(1; n+m-2)$ centrally on H_0, non-centrally otherwise. Hence

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}$$

which has the $t(n+m-2)$-distribution, centrally on H_0 non-centrally otherwise.

5.3.2 Testing a Regression Coefficient

In the linear regression (5:3) and (5:6), the null hypothesis is $H_0 : \beta_i = 0$, β_i being the i'th element of β. In this case $B^{>}\beta = \beta_i = 0$ and

$$B^{>} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 \end{bmatrix},$$

with unity in the i'th position. Clearly $B^{>}(X^{>}X)^{-1}X^{>}y = \hat{\beta}_i$ and $B^{>}(X^{>}X)^{-1}B = m^{ii}$ the (i; i)'th element of $(X^{>}X)^{-1}$. Since r = 1, applying (5:23)

$$F = \frac{\hat{\beta}_i^2}{m^{ii}\hat{\sigma}^2} \sim F(1; n-k).$$

Hence

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

where $SE(\hat{\beta}_i) = \hat{\sigma}\sqrt{m^{ii}}$ and $t \sim t(n - k)$, centrally on H_0, non-centrally otherwise.

When the null is $\beta_i = \beta_i^*$, then the t-test becomes

$$t = \frac{\hat{\beta}_i - \beta_i^*}{SE(\hat{\beta}_i)}.$$

5.3.3 Testing a Linear Combination of Regression Coefficients

As an example, consider, in the context of (5:3) with (5:6), $H_0 : 2\beta_2 + 3\beta_3 = 1$ as a null hypothesis to test. In this case, $B'\beta = 1$ with $B' = [0 \quad 2 \quad 3 \quad 0 \quad 0 \quad \cdots \quad 0]$. Let γ be a solution to $B'\gamma = 1$, e.g. $\gamma' = [0 \quad 1/2 \quad 0 \quad 0 \quad 0 \quad \cdots \quad 0]$. Then $B'(\beta - \gamma) = 0$; $r = 1$ and

$$\hat{\sigma}^2 F = (B'\hat{\beta} - \gamma)'\left[2\left(2m^{22} + 3m^{32}\right) + 3\left(2m^{23} + 3m^{33}\right)\right]^{-1}(B'\hat{\beta} - \gamma)$$

where m^{ij} is the (i, j)'th element of $(X'X)^{-1}$. F will follow the $F(1, n - k)$ distribution and hence

$$t = \frac{2\hat{\beta}_2 + 3\hat{\beta}_3 - 1}{\hat{\sigma}\sqrt{4m^{22} + 12m^{23} + 9m^{33}}} \sim t(n - 1)$$

since $m^{23} = m^{32}$ centrally on H_0, non-centrally otherwise. The denominator of t is $SE(2\hat{\beta}_2 + 3\hat{\beta}_3)$.

5.3.4 Testing a Block of Regression Coefficients

Again using (5:3) and (5:6) as the maintained hypothesis, consider the special case

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

in which X_1 has k_1 columns and X_2 has k_2 columns. The null hypothesis is $H_0 : \beta_2 = 0$. In this case, $L = R[X]$ where $X = [X_1 : X_2]$, $L_0 = R[X_1]$ and hence $P_0 = X_1\left(X_1'X_1\right)^{-1}X_1' = P_1$. Now

$$\begin{aligned} P &= F_1 + F_2 \\ &= X_1\left(X_1'M_2X_1\right)^{-1}X_1'M_2 + X_2\left(X_2'M_1X_2\right)^{-1}X_2'M_1. \end{aligned}$$

Then

$$\begin{aligned} P - P_0 &= (I - P_0)P \\ &= (I - P_1)P \\ &= M_1P \\ &= M_1F_2 \\ &= M_1X_2\left(X_2'M_1X_2\right)^{-1}X_2'M_1. \end{aligned}$$

Thus

$$
\begin{aligned}
F &= \frac{y^{>}(P - P_0)y}{y^{>}(I_n - P)y} \cdot \frac{n - k}{r} \\
&= \frac{y^{>}(P - P_1)y}{y^{>}(I_n - P)y} \cdot \frac{n - k}{k_2} \\
&= \frac{y^{>}M_1X_2\left(X_2^{>}M_1X_2\right)^{-1}X_2^{>}M_1y}{k_2\hat{\sigma}^2}.
\end{aligned}
$$

Moreover

$$
k_2F = \frac{y^{>}M_1X_2\left(X_2^{>}M_1X_2\right)^{-1}X_2^{>}M_1y}{\hat{\sigma}^2}.
$$

F here is distributed as $F(k_2; n - k)$ and k_2F is approximately $\chi^2(k_2)$ for `large' n.

5.3.5 Durbin-Hausman Testing

Referring to section 5.3.5, where $\gamma_2 = 0$, the maintained model is

$$y = X_1\gamma_1 + X_2\gamma_2 + \varepsilon \tag{5.28}$$

and, on H_0, the model is

$$y = X_1\gamma_1 + \varepsilon. \tag{5.29}$$

It will be convenient to write $\gamma_1 = \beta_1$ in (5:29) and to test

$$y = X_1\beta_1 + \varepsilon \tag{5.30}$$

as a model related to (5:28) when $\gamma_1 = \beta_1$. This permits a re-interpretation of H_0 from $H_0 : \gamma_2 = 0$ to $H_0^{*} : \gamma_1 - \beta_1 = 0$ where γ_1 is taken to be from (5:28). A natural

way to test H_0^{α} is to consider a standardized quadratic form based upon the difference

$$\begin{aligned}
\tilde{\alpha}_1 - \alpha_1 &= \left(X_1'M_2X_1\right)^{-1}X_1'M_2y - \left(X_1'X_1\right)^{-1}X_1'y \\
&= \left(X_1'M_2X_1\right)^{-1}\left[X_1'M_2 - \left(X_1'M_2X_1\right)\left(X_1'X\right)\left(X_1'X_1\right)^{-1}\right]y \\
&= \left(X_1'M_2X_1\right)^{-1}X_1'M_2(I - P_1)y \\
&= -\left(X_1'M_2X_1\right)^{-1}X_1'P_2M_1y. \qquad (5.31)
\end{aligned}$$

Clearly, (5:31) is zero whenever L_1 and L_2 are orthogonal; to proceed it is presumed L_1 is not orthogonal to L_2. Let $A = X_1'M_2X_1$ and $P_2X_1 = X_2Q$, $Q = \left(X_2'X_2\right)^{-1}X_2'X_1$. A is always non-singular. When $k_1 = k_2$, Q is square and non-singular and

$$\tilde{\alpha}_1 - \alpha_1 = -A^{-1}Q'X_2'M_1y \sim N\left(-A^{-1}Q'X_2'M_1\iota;\ A^{-1}Q'X_2'M_1X_2QA^{-1}\sigma^2\right) \qquad (5.32)$$

with $-A^{-1}Q'X_2'M_1\iota = 0$ when $\iota \in L_1$, i.e. when H_0 and H_0^{α} coincide. Thus, given L_1 and L_2 are not orthogonal,

$$\begin{aligned}
&\left(\tilde{\alpha}_1 - \alpha_1\right)'\left[A^{-1}Q'X_2'M_1X_2QA^{-1}\sigma^2\right]^{-1}\left(\tilde{\alpha}_1 - \alpha_1\right) \qquad (5.33) \\
&= \frac{y'M_1X_2QA^{-1}\left[AQ^{-1}\left(X_2'M_1X_2\right)^{-1}Q'^{-1}A\right]AQ'X_2'M_1y}{\sigma^2} \\
&= \frac{y'M_1X_2\left[X_2'M_1X_2\right]^{-1}X_2'M_1y}{\sigma^2} \sim \chi^2(k_2;\pm)
\end{aligned}$$

with $\pm = 0$ when $\iota \in L_1$. Moreover,

$$\frac{y'M_1X_2\left[X_2'M_1X_2\right]^{-1}X_2'M_1y}{k\hat{\sigma}^2} \sim F(k;\ n-k) \qquad (5.34)$$

where σ^2 is calculated from (5:28). Thus the testing of $H_0^{\alpha}: \beta_1 - \gamma_1 = 0$, when $k_1 = k_2$ and L_1 is not orthogonal to L_2, is seen to lead to precisely the same test as $H_0: \gamma_2 = 0$. Although $k_1 = k_2$ seems to be a very special case, in fact it is quite usual in specification testing using instrumental variables (IV's).

When $k_2 < k_1$, (5:32) still holds except that now the dispersion is singular, being of order $(k_1 \times k_2)$ and rank k_2. The easiest way then to proceed is to standardize the distribution of $\hat{\alpha}_1 - \beta_1$ on L_2. Note that Q is $(k_2 \times k_1)$ of rank k_2, whereupon QQ' has a unique inverse. Also let C be a non-singular matrix of order $(k_2 \times k_2)$ such that $CX_2'M_1X_2C' = I_{k_2}$. Then

$$C(QQ')^{-1}QA(\hat{\alpha}_1 - \beta_1) = -CX_2'M_1y \sim N(0, I_{k_2}\sigma^2)$$

when $\pi \in L_1$ or when H_0^{α} holds and $L_1$is not orthogonal to L_2. It follows that

$$\frac{(\hat{\alpha}_1 - \beta_1)'A'Q'(QQ')^{-1}C'C(QQ')^{-1}QA(\hat{\alpha}_1 - \beta_1)}{\sigma^2} \sim \chi^2(k_2; \pm)$$

or

$$\frac{y'M_1X_2QA^{-1}A'Q'(QQ')^{-1}(QQ')^{-1}X_2'M_1X_2(QQ')^{-1}(QQ')^{-1}QAA^{-1}Q'X_2'M_1y}{\sigma^2}$$

$$= \frac{y'M_1X_2(X_2'M_1X_2)^{-1}X_2M_1y}{\sigma^2} \sim \chi^2(k_2; \pm)$$

i.e. (5:33), and hence (5:34). Thus when $k_1 \geq k_2$ and L_1 lies strictly oblique to L_2, then an F-test for H_0 is equivalent to an F-test for H_0^{α}. In Durbin-Hausman testing, the difference in estimators is not between unrestricted and restricted estimators, but

rather between an IV estimator of the form $\left(X^{\top}AX\right)^{-1}X^{\top}Ay = \hat{\beta}_A$, where A is an orthogonal projection matrix, and the least squares estimator of the same coefficient $\left(X^{\top}X\right)^{-1}X^{\top}y = \hat{\beta}$. This

$$\hat{\beta}_A - \beta = \left(X^{\top}AX\right)^{-1}X^{\top}AMy$$

$M = I_n - P$, with dispersion

$$D\left(\hat{\beta}_A - \beta\right) = \left(X^{\top}AX\right)^{-1}X^{\top}AMAX\left(X^{\top}AX\right)\sigma^2.$$

The standardized quadratic form is

$$\frac{y^{\top}MAX\left(X^{\top}AX\right)^{-1}\left[X^{\top}AX\left(X^{\top}AMAX\right)^{-1}X^{\top}AX\right]\left(X^{\top}AX\right)^{-1}X^{\top}AMy}{\sigma^2}$$

$$= \frac{y^{\top}MAX\left(X^{\top}AMAX\right)^{-1}X^{\top}AMy}{\sigma^2}$$

with corresponding F-statistic given by

$$F = \frac{y^{\top}MAX\left(X^{\top}AMAX\right)^{-1}X^{\top}AMy}{k\hat{\sigma}^2}$$

where $\hat{\sigma}^2$ has been calculated from the equation

$$y = X\gamma + AX\beta + \varepsilon.$$

Under appropriate conditions, $F \sim F(k; n - 2k)$. Thus the testing of $H_0: \gamma - \beta = 0$ is the same as testing $\beta = 0$ or $AX\beta = \delta_2 = 0$; δ_2 otherwise belonging to a k-dimensional subspace of R[A]. Notice that AX is of the same order and rank as X.

This is equivalent to the case mentioned above when $k_1 = k_2$. $R[A]$ is presumed oblique to L, and this corresponds to the case above when L_1 is not orthogonal to L_2.

5.3.6 Testing a Set of r Restrictions

This may be done using formulation (5:10) or (5:12). All that is necessary is to de¯ne B. Let k = 6 and the restrictions be

$$\bar{}_1 + \bar{}_2 + \bar{}_3 = 1$$

$$2\bar{}_4 + \bar{}_6 = 2.$$

Then

$$\begin{bmatrix} & & & & & & & & \\ 1 & 1 & 1 & 0 & 0 & 0 & \bar{} & = & 1 \\ & & & & & & & & \\ 0 & 0 & 0 & 2 & 0 & 1 & & & 2 \end{bmatrix}$$

or $B^\top \beta = \theta \neq 0$, as in (5:12). Consider a solution to $B^\top \lambda = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$,

$$\lambda = \begin{bmatrix} 1 \\ -1 \\ 1 \\ \frac{1}{2} \\ 6 \\ 1 \end{bmatrix}.$$

Then $B^\top \lambda = \theta$ and it is straightforward to write

$$y - X\lambda = X(\beta - \lambda) + \varepsilon$$

leading to

$$y^* = X\beta^* + \varepsilon$$

subject to $B^\top \beta^* = 0$. The F-test is then

$$\begin{aligned} F &= \frac{\left(B^\top \hat{\beta}^*\right)^\top \left[B^\top \left(X^\top X\right)^{-1} B\right]^{-1} B^\top \hat{\beta}^*}{2\sigma^2} \\ &= \frac{\left(B^\top \hat{\beta} - \theta\right)^\top \left[B^\top \left(X^\top X\right)^{-1} B\right]^{-1} \left(B^\top \hat{\beta} - \theta\right)}{2\sigma^2} \end{aligned}$$

which has the $F(2; n-k)$ distribution, $\hat{\beta}$ being $\left(X^\top X\right)^{-1} X^\top y$ and $X\lambda = \theta$.

5.3.7 Testing for Structural Change

Consider two time periods comprising n_1 and n_2 observations which follow the linear regression

$$y_1 = X_1\beta_1 + \varepsilon_1 \qquad (5.35)$$
$$y_2 = X_2\beta_2 + \varepsilon_2$$

or

$$y = X\beta + \varepsilon \qquad (5.36)$$

where $y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$, $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$ and

$$X = \begin{bmatrix} X_1 & 0 \\ 0 & X_2 \end{bmatrix}.$$

In this situation β_1 and β_2 are presumed to be different. The null hypothesis is $H_0 : \beta_1 = \beta_2 = \beta_0$ and it is assumed that both n_1 and n_2 are greater than k. The model under H_0 is

$$y = X_0\beta_0 + \varepsilon \qquad (5.37)$$

where

$$X_0 = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$

$\bar{}_0$ having half the number of elements as $\bar{}$. Notice that H_0 may be written

$$\begin{bmatrix} I_k & -I_k \end{bmatrix} \bar{} = 0$$

or

$$\bar{} = \begin{bmatrix} I_k \\ I_k \end{bmatrix} \bar{}_0 = M\bar{}_0. \tag{5.38}$$

Thus using (5:38) in (5:36)

$$\begin{aligned} y &= XM\bar{}_0 + " \\ &= X_0\bar{}_0 + " \end{aligned}$$

where $X_0 = XM$ in (5:37).

The orthogonal projection matrix onto R [X] is

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$$

where $P_i = X_i \left(X_i^{>} X_i\right)^{-1} X_i^{>}$; $i = 1; 2$; $P_0 = X_0 \left(X_0^{>} X_0\right)^{-1} X_0^{>}$. Then the F-statistic is

$$F = \frac{y^{>}(P - P_0)\, y}{k\hat{\sigma}^2}$$

where $\hat{\sigma}^2 = \frac{y^{>}(I_n - P)y}{(n_1 + n_2 - 2k)} = \frac{\hat{"}_1^{>}\hat{"}_1 + \hat{"}_2^{>}\hat{"}_2}{(n_1 + n_2 - 2k)}$. It is an assumption of equation (5:36) that $"_1 \sim N(0; I_{n_1}\sigma^2)$ and $"_2 \sim N(0; I_{n_2}\sigma^2)$, so that $"_1$ and $"_2$ have a common mean 0 and a common variance σ^2.

If the two regressions contain constant terms, it may be desirable to consider an unrestricted model (5:36) in which every coe± cient may di®er from one period to the next, and a null in which corresponding slope coe± cients are the same but the constants di®er. More generally, it may be desirable to consider restrictions on only a subset of coe± cients. For example, let the two regressions from (5:36) be

$$y = \begin{bmatrix} X_{11} & X_{12} & 0 & 0 \\ 0 & 0 & X_{21} & X_{22} \end{bmatrix} \begin{bmatrix} \beta_{11} \\ \beta_{12} \\ \beta_{21} \\ \beta_{22} \end{bmatrix} + \varepsilon \qquad (5.39)$$

$$= X_0 \beta_0 + \varepsilon,$$

the columns of X_{11} and X_{21} referring to the same variables, the columns of X_{12} and X_{22} referring to the same or possibly di®erent variables. The null hypothesis is $H_0 : \beta_{11} = \beta_{21} = \beta_1$ and the null model is then

$$y = \begin{bmatrix} X_{11} & X_{12} & 0 \\ X_{21} & 0 & X_{22} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_{12} \\ \beta_{22} \end{bmatrix} + \varepsilon \qquad (5.40)$$

$$= X_0 \beta_0 + \varepsilon$$

in an obvious notation. P is the orthogonal projection for (5:39) and

$$P = \begin{bmatrix} P_1 & 0 \\ 0 & P_2 \end{bmatrix}$$

where P_1 refers to $[X_{11} : X_{12}]$ and P_2 to $[X_{21} : X_{22}]$, it being presumed that the k_1 columns of $[X_{11} : X_{12}]$ are less than n_1 and similarly $k_2 < n_2$ where k_2 is the number of columns in $[X_{21} : X_{22}]$: In equation (5:39), the projection is

$$P_0 = \begin{bmatrix} \vdots & P_{12} & 0 \\ P_{\alpha} & \vdots & \cdots \quad \cdots \\ & \vdots & p \quad P_{22} \end{bmatrix},$$

P_{α} being determined by $X_{11}^{>} \quad X_{21}^{>}$, P_{12} and P_{22} referring to X_{12} and X_{22}. The F-statistic takes on the same form as before:

$$F = \frac{y^{>}(P - P_0)y}{k_{11}\hat{\sigma}^2}$$

where k_{11} is the number of elements in β_{11}, and so also in β_{21}, and $\hat{\sigma}^2 = \frac{y^{>}(I_n - P)y}{(n_1+n_2 - k_1 - k_2)} = \frac{\hat{\varepsilon}_1^{>}\hat{\varepsilon}_1 + \hat{\varepsilon}_2^{>}\hat{\varepsilon}_2}{(n_1+n_2 - k_1 - k_2)}$.

An important requirement of the structural change test is that the two variances, of ε_1 and ε_2, are the same. Returning to the model (5:35), (5:36), an estimate of the error variance in the first period is $\hat{\sigma}_1^2 = \frac{y_1^{>}(I_n - P_1)y_1}{(n_1 - k_1)}$ and for the second period is $\hat{\sigma}_2^2 = \frac{y_2^{>}(I_n - P_2)y_2}{(n_2 - k_2)}$. In the decomposition

$$y^{>}y = y^{>}P_0y + y^{>}(P - P_0)y + y^{>}(I_n - P)y,$$

$y^{>}(I_n - P)y$ may itself be decomposed into

$$y^{>}(I_n - P)y = y^{>}\begin{bmatrix} (I_{n_1} - P_1) & 0 \\ 0 & 0 \end{bmatrix}y + y^{>}\begin{bmatrix} 0 & 0 \\ 0 & (I_{n_1} - P_2) \end{bmatrix}y$$

having ranks of, on the left-hand side, $(n - k_1 - k_2)$ and, on the right-hand side, $(n_1 - k_1)$ and $(n_2 - k_2)$; $n = n_1 + n_2$. Thus $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, each divided by σ^2, are independent χ^2-variates, and

$$F = \frac{y_1'(I_{n_1} - P_1)y_1}{y_2'(I_{n_2} - P_2)y_2} \cdot \frac{n_2 - k_2}{n_1 - k_1},$$

and $F \sim F(n_1 - k_1; n_2 - k_2)$. F is suitable for testing the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$. This should be done prior to any test for structural change since $\sigma_1^2 = \sigma_2^2$ is a condition for application of the F-distribution.

5.3.8 Recursive Residuals

The aim with so-called recursive residuals is to calculate residuals sequentially to facilitate the testing of the constancy of a regression. Thus a regression may be fitted to the first k observations, then to $k + 1$; $k + 2$, and so on, until all n observations are used. At each stage, the additional observation is tested to determine whether it is significantly different from the previous regression. To formulate the various tests, the following notation is used.

X represents the $(n \times k)$ matrix of n observations on k explanatory variables, the first k observations being represented by X_k. The orthogonal projection onto $R[X_k] = X_k(X_k'X_k)^{-1}X_k' = P_k = I_k$ since X_k is $(k \times k)$ of rank k. If r observations are added to X_k there results X_{k+r} and hence $P_{k+r} = X_{k+r}(X_{k+r}'X_{k+r})^{-1}X_{k+r}'$. Using

this notation, let

$$\hat{P}_r = \begin{bmatrix} P_{k+r} & 0 \\ 0 & I_{n-k-r} \end{bmatrix} \qquad r = 0; 1; \cdots; (n-k),$$

and note that $\hat{P}_r$ has rank $k + n - k - r = n - r$. Consequently, as r increases, so the rank declines and the implicit number of restrictions increases (since as observations are added to the last regression, so these are restricted to have the same regression coefficient vector as all the observations before them). In particular, $\hat{P}_0$ has rank n and there are no restrictions (indeed $\hat{P}_0 = I_n$); $\hat{P}_1$ has rank $(n-1)$ and one restriction; $\hat{P}_2$ has rank $(n-2)$ and 2 restrictions; and so on until $\hat{P}_{n-k}$ which has rank $k + n - k - (n-k) = k$ representing $(n-k)$ restrictions, and $\hat{P}_{n-k} = P = X(X^{\top}X)^{-1}X^{\top}$.

In geometrical terms, the rank of a matrix represents the dimension of the range of that matrix; and the smaller the dimension of a subspace, the greater the number of restrictions on R^n to determine that subspace. Thus $R[\hat{P}_0] \supset R[\hat{P}_1] \supset R[\hat{P}_2] \supset \cdots \supset R[\hat{P}_{n-k}]$. Consider, then, the decomposition

$$\begin{aligned} y^{\top}y = {} & y^{\top}\hat{P}_{n-k}y + y^{\top}(\hat{P}_{n-k+1} - \hat{P}_{n-k})y + y^{\top}(\hat{P}_{n-k+2} - \hat{P}_{n-k+1})y + \cdots \\ & \cdots + y^{\top}(\hat{P}_1 - \hat{P}_2)y + y^{\top}(I_n - \hat{P}_1)y. \end{aligned}$$

Three observations may be made about this decomposition.

1. For any two distinct values of r, say m and q; m < q, $P_q P_m = P_q$ and hence

$$\begin{aligned}(P_q - P_{q+1})(P_m - P_{m-1}) &= P_q - P_q - P_{q+1} + P_{q+1} \\ &= 0.\end{aligned}$$

Thus all the quadratic forms are naturally independent and the ranks of the right-hand side quadratic forms sum to $k + (k+1) - k + (k+2) - (k+1) + \cdots + (n-1) - (n-2) + n - (n-1) = k + 1 + 1 + \cdots + 1 + 1 = k + (n - k) = n =$ the rank of the left-hand side quadratic form. Division by σ^2 ensures that the left-hand side is $\chi^2(n; \delta)$ while each of the terms on the right-hand side is an independent $\chi^2(1; \delta_r)$ except the first which is $\chi^2(k; \delta_r)$, $r = 1; 2; \cdots; (n - k)$ and $\S \delta_r = \delta$.

2. Let $y'(P_q - P_{q+1})y = y'(I_n - P_{q+1})y - y'(I_n - P_q)y = S^2_{q+1} - S^2_q = U^2_q$; U_q is called the q'th recursive residual.

3. Moving forward from the k'th observation,

$$\frac{y'(P_1 - P_2)y}{y'(I_n - P_2)y} \sim \beta\left(\frac{1}{2}; 1\right),$$

$$\frac{y'(P_2 - P_3)y}{y'(I_n - P_3)y} \sim \beta\left(\frac{1}{2}; \frac{3}{2}\right),$$

$$\frac{y'(P_3 - P_4)y}{y'(I_n - P_4)y} \sim \beta\left(\frac{1}{2}; 2\right),$$

and so on; these statistics represent ways of testing whether the $(k + r)$'th observation is signi¯cantly di®erent from the preceding $(k + r ¡ 1)$ observations, in respect of the linear regression taking the general form of $y = X¯ + "$. These tests are based on the Lagrange Multiplier Principle. Alternatively an F-test may be used at each stage, beginning with

$$\frac{y^{>} \overset{3}{P}_1 ¡ \overset{´}{P}_2 y}{y^{>} \overset{3}{I}_n ¡ \overset{´}{P}_1 y} » F(1; 1).$$

If this test does not reject $H_0 : ¯_{k+1} ¡ ¯_{k+2} = 0$ (where $¯_{k+1}$ refers to the ¯rst $k + 1$ observations and $¯_{k+2}$ to the ¯rst $k + 2$ observations) then the numerator and denominator are added to yield $y^{>} \overset{3}{I}_n ¡ \overset{´}{P}_2 y$ and the next test is

$$\frac{y^{>} \overset{3}{P}_2 ¡ \overset{´}{P}_3 y}{y^{>} \overset{3}{I}_n ¡ \overset{´}{P}_3 y} » F(1; 2).$$

Again if this test does not reject $H_0 : ¯_{k+2} ¡ ¯_{k+3} = 0$ then we proceed to add $y^{>} \overset{3}{P}_2 ¡ \overset{´}{P}_3 y + y^{>} \overset{3}{I}_n ¡ \overset{´}{P}_2 y = y^{>} \overset{3}{I}_n ¡ \overset{´}{P}_3 y$ and use

$$\frac{y^{>} \overset{3}{P}_3 ¡ \overset{´}{P}_4 y}{y^{>} \overset{3}{I}_n ¡ \overset{´}{P}_3 y} » F(1; 3)$$

and so on. The method is referred to as \adding insigni¯cant sums of squares" (as is done at each stage).

A full discussion of the tests and associated tests is found in R. L. Brown, J. Durbin and J. M. Evans, \Techniques for testing the constancy of relationships over

time" (with discussion), Journal of the Royal Statistical Society, Series B, 37(1975), 149-92.

6 Limits, Continuity and Convergence

6.1 Introduction

The discussion in this chapter is decidedly terse. The aim is to provide a minimum of understandable material, sufficient to tackle the econometric problems that arise later on. Many of these problems are provided with a solution that rests on reasonable assumptions about the data being analyzed. To understand the assumptions it is necessary to understand certain definitions. The definitions may appear forbidding, while in fact being natural to the task at hand. To ease the way forward, each definition is followed by an example. A good reference in the material covered (and many extensions) is White (1984).

6.2 Real Numbers

Definition 29 Let $\{b_n\}$ be a sequence of real numbers, $n = 1; 2; 3; ¢¢¢$. If for every real $\pm > 0$ there exists an integer $N(\pm)$ and a real number b such that

$$|b_n - b| < \pm \qquad \text{for all } n \geq N(\pm)$$

then b is said to be the limit of $\{b_n\}$ and $\{b_n\}$ is said to converge to b as $n \to 1$; this is written $b_n \to b$ as $n \to 1$ or $\lim_{n \to 1} \{b_n\} = b$. ■

Example 9 Let $b_n = b + \frac{1}{n}$; $n = 1; 2; 3; ¢¢¢$. Then $b_1 = b + 1$; $b_2 = b + \frac{1}{2}$; $b_3 = b + \frac{1}{3}$; $¢¢¢$. Clearly, $fb_n g \to b$ as $n \to 1$.

Example 10 Consider the expansion of $\left(1 + \frac{a}{n}\right)^{¢_n}$:

$$\left(1 + \frac{a}{n}\right)^n = 1 + n\left(\frac{a}{n}\right) + \frac{n(n-1)}{2!}\left(\frac{a}{n}\right)^2 + ¢¢¢ + \frac{n(n-1)¢¢¢(n-r+1)}{r!}\left(\frac{a}{n}\right)^r + ¢¢¢.$$

As $n \to 1$

$$\left(1 + \frac{a}{n}\right)^n = 1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + ¢¢¢ = e^a.$$

Hence $\lim_{n \to 1}\left(1 + \frac{a}{n}\right)^{¢_n} = e^a$.

Turning to vectors of real elements, let b_n be a $(k \times 1)$ vector with typical element b_n^i $(i = 1; 2; ¢¢¢; k)$. Let $b_n^i \to b^i$ as $n \to 1$ for each i. Then $fb_n g \to b$ as $n \to 1$ where b has elements b^i. Matrices are treated similarly, except it is usually necessary to make an assumption about the properties of the limiting matrix. For example, if X is an $(n \times k)$ real matrix, then $X^{>}X = \left[\sum_q x_{iq} x_{jq}\right]$ $i; j = 1; 2; ¢¢¢; k$; $q = 1; 2; ¢¢¢; n$. If X has rank k, then $X^{>}X$ has rank k and is positive de¯nite. Let

$$\frac{1}{n}X^{>}X \to M_{XX} \quad \text{as } n \to 1$$

where $\frac{1}{n}\sum_q x_{iq} x_{jq} \to m_{ij}$. Commonly it will be assumed that the m_{ij} are ¯nite and M_{XX} is a positive-de¯nite matrix.

6.3 Continuity

Definition 30 Let $g(\cdot)$ be a vector-valued function $R^p \to R^k$ and let $b \in R^p$. The function $g(\cdot)$ is continuous at b if for any sequence $\{b_n\}$ whose limit as $n \to \infty$ is b,

$$\lim_{n \to \infty} g(b_n) = g(b).$$

Equivalently, $g(\cdot)$ is continuous at b if, for every $\varepsilon > 0$, there exists a $\delta(\varepsilon) > 0$ such that for $a \in R^p$ and $|a_i - b_i| < \delta(\varepsilon)$; $i = 1, 2, \ldots, p$; $|g_j(a) - g_j(b)| < \varepsilon$; $j = 1, 2, \ldots, k$. If $B \subset R^p$, then g is continuous on B if it is continuous for every point of B. ■

Example 11 If $a_n \to a$ and $b_n \to b$; $a_n + b_n \to a + b$ and $a_n b_n \to ab$.

Example 12 The matrix inverse function is continuous at every point that represents a non-singular matrix. Thus if $\frac{1}{n}X'X \to M_{XX}$, then $\left(\frac{1}{n}X'X\right)^{-1} \to M_{XX}^{-1}$, so long as M_{XX} is finite and non-singular.

6.4 The Order of a Sequence

Definition 31 The sequence $\{b_n\}$ is at most of order n^λ, written $O(n^\lambda)$ if, for some finite real number c there exists an integer N such that $|n^{-\lambda} b_n| < c$ for all $n \geq N$. ■

Definition 32 The sequence $\{b_n\}$ is of order smaller that n^λ, written $o(n^\lambda)$, if, for every $\delta > 0$, there exists an $N(\delta)$ such that $|n^{-\lambda} b_n| < \delta$ for all $n \geq N(\delta)$. ■

Clearly $\{b_n\} = O(n^{\delta})$ if $\left(\frac{n}{n^{\delta}}\right)^{\alpha} b_n$ is eventually bounded; and $b_n = o(n^{\delta})$ if $\left(\frac{n}{n^{\delta}}\right)^{\alpha} b_n \to 0$. If $\{b_n\}$ is $o(n^{\delta})$, then it must be $O(n^{\delta})$. Also, if $\{b_n\} = O(n^{\delta})$, then for every $\varepsilon > 0$, $\{b_n\} = o(n^{\delta+\varepsilon})$. If $\{b_n\} = O(n^0) = O(1)$, then it is eventually bounded and it may or may not have a limit. If $\{b_n\} = o(1)$, then $b_n \to 0$.

Example 13 Let $\sum x_i = O(n)$. Then $\frac{1}{n}\sum x_i = O(1)$ and this implies that, for all $\varepsilon > 0$; $\sum x_i = o(n^{1+\varepsilon})$.

Example 14 Let $b_n = 4 + 3n + 6n^2$. Then $\{b_n\} = O(n^2)$ and $\{b_n\} = o(n^{2+\varepsilon})$ for all $\varepsilon > 0$. ■

In regard to matrices and vectors, these are of $O(n^{\delta})$ or $o(n^{\delta})$ if each element is $O(n^{\delta})$ or $o(n^{\delta})$.

6.5 Almost Sure Convergence

It is often of interest in statistics generally, and hence in econometrics, to consider averages of sequences of random variables $\{Y_i\}$. Let ω represent the entire random sequence $\{Y_i\}$. Let $b_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Definition 33 Let $\{b_n(\omega)\}$ represent a sequence of random variables. $b_n(\omega)$ converges almost surely to b, written $b_n(\omega) \xrightarrow{a.s.} b$ if there exists a real number b such that $P_\omega[b_n(\omega) \to b] = 1$ where the probability measure P_ω refers to the distribution of ω

which determines the joint distribution function for the entire sequence of random variables. ■

What $P_! [b_n(!) \; ! \; 1] = 1$ means is that some sequences in $fb_n(!)g$ may not converge, but these have probability measure zero.

If $fb_n(!)g \overset{a.s.}{!} b$ and $g(:)$ is continuous at b, then $g(b_n) \overset{a.s.}{!} g(b)$.

Example 15 If $y = X^- + "$ is the familiar least squares model in n observations and k variables, and these k variables are random variables, then let $\frac{1}{n}X^>X \overset{a.s.}{!} M_{XX}$ a $^-$nite pd matrix and $n^{i\ 1}X^>" \overset{a.s.}{!} 0$. It follows that $^{\mathbf{i}}X^>X^{\mathbf{¢}_{i\ 1}}X^>y = \hat{}$ may be written as $^{\mathbf{i}}\frac{1}{n}X^>X^{\mathbf{¢}_{i\ 1}}\frac{1}{n}X^>y = {}^{\mathbf{i}}\frac{1}{n}X^>X^{\mathbf{¢}_{i\ 1}}\frac{1}{n}X^>(X^- + ")$. Thus $\hat{} = ^- + {}^{\mathbf{i}}\frac{1}{n}X^>X^{\mathbf{¢}_{i\ 1}}\frac{1}{n}X^>" \overset{a.s.}{!} ^- + M_{XX}¢0 = ^-$.

Example 16 Consider now $\approx = {}^{\mathbf{i}}X^>A_nX^{\mathbf{¢}_{i\ 1}}X^>A_ny$. If it is assumed that $\frac{1}{n}X^>A_nX \overset{a.s.}{!} M_A$ which is $^-$nite and positive de$^-$nite and $\frac{1}{n}X^>A_n" \overset{a.s.}{!} 0$, then $\hat{} \overset{a.s.}{!} ^-$ by the same reasoning as above. ■

6.6 Convergence in Probability

De$^-$nition 34 If there exists a real number b such that, for every $" > 0$,

$$P_! [jb_n(!) ¡ bj < "] = 1 \qquad \text{as } n \; ! \; 1,$$

then $b_n(!)$ is said to converge in probability to b, which is written $b_n(!) \overset{p}{!} b$ or $p\lim_{n\,!\,1}[b_n(!)] = b$. If $b_n(!) \overset{p}{!} b$ and $g(:)$ is continuous at b, then $g(b_n(!)) \overset{p}{!}$

$g(b)$. Let $g(:)$ be a vector-valued function $R^p \to R^k$ and $b_n(!)$ and b be $(p \pounds 1)$ vectors. If $g(:)$ is continuous at b, then $b_n^i(!) \xrightarrow{p} b^i$ implies that $g_j(b_n(!)) \xrightarrow{p} g_j(b)$ $i = 1; 2; ¢¢¢; p$; $j = 1; 2; ¢¢¢; k$. ■

6.7 Convergence in Distribution

From elementary statistics it is known that if $X \sim (1; ¾^2); 0 < ¾^2 < 1$, and $x_1; x_2; ¢¢¢; x_n$ is a sequence of random selections from this distribution then, by the central limit theorem, for sufficiently large n,

$$\frac{\bar{x} - 1}{¾ = \sqrt{n}} \sim N(0; 1)$$

or $\sqrt{n}(\bar{x} - 1) \sim N(0; ¾^2)$.

Definition 35 Let $\{b_n\}$ be a sequence of random variables with joint distribution function $\{F_n(:)\}$. If $F_n(y) \to F(y)$ as $n \to 1$ for every point y, where $F(y)$ is the distribution function of the random variable Y, then $\{b_n\}$ converges in distribution to the distribution of the random variable Y which is written $b_n \xrightarrow{d} Y$. In this definition vector may be interchanged with variable. ■

Let y range over R^p according to a distribution with mean 1, a $(p \pounds 1)$ vector, and dispersion $§$, a $(p \pounds p)$ pd matrix. This is written $y \sim (1; §)$. Let $\{y_n\}$ be a random sequence of n from $(1; §)$. Then $y_1; y_2; ¢¢¢; y_n$ are n $(p \pounds 1)$ vectors each

with mean μ and as $n \to \infty$

$$\sqrt{n}\left[\frac{1}{n}\sum (y_i - \mu)\right] \xrightarrow{d} N(0, \Sigma)$$

or

$$\sqrt{n}(\bar{y} - \mu) \xrightarrow{d} N(0; \Sigma)$$

where $\bar{y} = \frac{1}{n}\sum y_i$ is a $(p \times 1)$ vector of sample means. By an abuse of notation, it is common to write

$$\bar{y} \overset{a}{\sim} N\left(\mu; \frac{1}{n}\Sigma\right).$$

Let μ be a parameter vector in R^k and $\hat{\mu}_n$ an estimate of μ based upon a random sample of n such that $\hat{\mu} \overset{a}{\sim} N\left(\mu; \frac{1}{n}\Sigma\right)$. Let $h(\mu)$ be a set of $r < k$ functions of μ which are real and continuous at μ. Then

$$h(\hat{\mu}) \overset{a}{\sim} N\left(h(\mu); n^{-1} H_\mu^\top \Sigma H_\mu\right),$$

where $H_\mu = \left[\frac{\partial h(\mu)^\top}{\partial \mu}\right]$ evaluated at μ. Now $\sqrt{n}(\hat{\mu} - \mu) \overset{a}{\sim} N(0; \Sigma)$ and hence $\operatorname{plim} \hat{\mu} = \mu$. Moreover, by Taylor's theorem,

$$h(\hat{\mu}) = h(\mu) + \frac{\partial h(\mu)}{\partial \mu^\top}(\hat{\mu} - \mu) + R$$

where $R \to 0$ as $\hat{\mu}_n \to \mu$. Hence

$$\sqrt{n}\left[h(\hat{\mu}) - h(\mu)\right] = \sqrt{n} H^\top (\hat{\mu} - \mu) + o_p(1) \xrightarrow{d} N\left(0; H_\mu^\top \Sigma H_\mu\right).$$

On the null hypothesis $h(\mu) = 0$,

$$h\left(\hat{\mu}_n\right) \overset{a}{\sim} N\left[h(\mu);\, n^{-1} H_{\mu}^{\top} \Sigma H_{\mu}\right].$$

Assuming that $\hat{\mu} \xrightarrow{p} \mu$; $H_{\hat{\mu}} \xrightarrow{p} H_{\mu}$; and $\hat{\Sigma} \xrightarrow{p} \Sigma$ then

$$H_{\hat{\mu}}^{\top} \hat{\Sigma} H_{\hat{\mu}} \xrightarrow{p} H_{\mu}^{\top} \Sigma H_{\mu}$$

whereupon, if $h(\mu) = 0$, then

$$W = \sqrt{n} h^{\top}\left(\hat{\mu}\right) \left[H_{\hat{\mu}}^{\top} \hat{\Sigma} H_{\hat{\mu}}\right]^{-1} h\left(\hat{\mu}\right) \sqrt{n} \overset{a}{\sim} \chi^2(r; 0).$$

7 Maximum-likelihood Estimation Procedures and Associated Tests of Signi¯cance

7.1 The General Problem

There are n independent observations $x_1; x_2; ¢¢¢; x_n$ from a known probability density function $f(x;\mu)$ and it is also known that the hypothesis $H : \mu = \mu_0 \; 2 \; -$ applies, $-$ being some subset (or all) of some k-dimensional Euclidean space R^k. It is desired to test whether the true value of μ, denoted by μ_0, is an element of $!$, a $(k ¡ r)$-dimensional subspace of $-$. There are two ways of expressing $!$ when $-$ is known: in the form of restraint equations, $H_0 : h(\mu) = 0$, where h represents r independent functions, or in the form of freedom equations, written $H_0 : \mu = \mu(®)$, in which $®$ is a set of $(k ¡ r)$ independent parameters: $® \; 2 \; A ½ R^{k¡r}$. It is sometimes the case that a combination of restraint and freedom equations is speci¯ed. There is always a freedom speci¯cation corresponding to a restraint equation speci¯cation and vice-versa; but the corresponding relationship is often di± cult to derive in practice.

Using the restraint form, $! = \{\mu : h(\mu) = 0; \mu \; 2 \; - ½ R^k\}$. Using the freedom equation speci¯cation, $! = \{\mu : \mu = \mu(®); ® \; 2 \; A ½ R^{k¡r}\}$.

7.2 General Justi¯cation

There are essentially two general justi¯cations for using maximum-likelihood methods. The ¯rst comes from Bayes' Theorem which states that

$$P(\mu \in \Omega \mid x_1, x_2, \cdots, x_n) \propto P(x_1, x_2, \cdots, x_n \mid \mu \in \Omega) P(\mu \in \Omega). \tag{7.1}$$

The probability that $\mu \in \Omega$, given that the sample $x_1, x_2, \cdots, x_n$ has been observed, is proportional to the probability that $x_1, x_2, \cdots, x_n$ will be observed, given that μ is indeed an element of Ω, times the probability that μ lies in Ω. $P(x_1, x_2, \cdots, x_n \mid \mu \in \Omega)$ is in principle a probability but, since the x_i's are observed and so are to be treated as parameters, and $\mu \in \Omega$ is to be treated as variable, $P(x_1, x_2, \cdots, x_n \mid \mu \in \Omega)$ is called the likelihood corresponding to n observations, written $L_n(x; \mu)$. $P(\mu \in \Omega)$ is called the prior, or a priori, probability since it represents the probability that μ is an element of Ω before the sample $x_1, x_2, \cdots, x_n$ has been observed. $P(\mu \in \Omega \mid x_1, x_2, \cdots, x_n)$ is the probability that μ is an element of Ω, given that the sample $x_1, x_2, \cdots, x_n$ has been observed; hence it is called the posterior or a posteriori probability.

$P(\mu \in \Omega)$ can only be assigned subjectively and this is not universally acceptable. However, it can be assumed that $P(\mu \in \Omega)$ is uniformly distributed over Ω, that is, that the investigator cannot distinguish, in terms of prior beliefs, any values of μ that are more likely than others. Such uniform distribution of prior probabilities over Ω is known as the Doctrine of Equal Ignorance. Applying this doctrine to ¯nding the

most likely value of μ in Ω, there results

$$\sup_{\mu \in \Omega} P(\mu \mid x_1, x_2, \cdots, x_n) = \sup_{\mu \in \Omega} \kappa L_n(x; \mu), \tag{7.2}$$

κ being the product of the constant of proportionality from (7:1) and the constant probability covering every $\mu \in \Omega$. Thus maximizing the likelihood for variations in $\mu \in \Omega$ yields the same value of μ as maximizing $P(\mu \mid x_1, x_2, \cdots, x_n)$ and the principle of maximum-likelihood is given a justification as a consequence of (i) the calculus of probability and (ii) the Doctrine of Equal Ignorance.

A second justification for maximum-likelihood methods is found in the properties that maximum-likelihood estimators possess under general conditions. Given that the probability density function, from which that n sample observations have been chosen, is known then the maximum-likelihood estimator of μ in Ω, say $\hat{\mu} \in \Omega$, will generally have the following properties.

1. $\hat{\mu}$ is consistent.

2. $\hat{\mu}$ is a BAN (Best Asymptotic Normal) estimator, that is, $\sqrt{n}\left(\hat{\mu} - \mu_0\right) \xrightarrow{d} N\left(0, B_{\mu_0}^{-1}\right)$ where $B_{\mu_0}^{-1}$ is the $(k \times k)$ positive-definite matrix which attains the Cramér-Rao lower bound, to be discussed below.

3. In view of property 2, $\hat{\mu} \overset{approx}{\sim} N\left(\mu_0, \frac{1}{n} B_{\mu_0}^{-1}\right)$ for large n, and hence standardized quadratic forms in $\hat{\mu}$, and in functions of $\hat{\mu}$ like $h\left(\hat{\mu}\right)$ for example, are χ^2-

distributed. Thus if $H_\mu = \frac{\partial h(\mu)}{\partial \mu^>}$ then, given sufficient regularity,

$$nh^> \left(\hat{\mu}\right)' \left[H_{\hat{\mu}} B_{\hat{\mu}}^{-1} H_{\hat{\mu}}^>\right]^{-1} h\left(\hat{\mu}\right)' \xrightarrow{d} \chi^2(r; \pm)$$

and $\pm = 0$ on $H_0 : \mu \in \Omega$.

4. $\hat{\mu}$ is a function of a sufficient statistic for μ, if a sufficient statistic for μ exists.

5. Maximum-likelihood estimators are invariant to reparametrization in the following sense. Let μ and ψ both be $(k \times 1)$ vectors such that for $g : \Omega \to T$ $\mu \in R^k$ is a smooth mapping that transforms $\mu \in \Omega$ uniquely into $g(\mu) = \psi \in T$. Let the new likelihood based on the same n observations be $L_n^*(x; \psi)$. Then $L_n(x; \mu) = L_n^*(x; \psi)$. Now $L_n\left(x; \hat{\mu}\right)' > L_n(x; \mu)$ for all $\mu \neq \hat{\mu}$ in Ω. It follows that $L_n^*(x; \hat{\psi}) = L_n^*\left(x; g\left(\hat{\mu}\right)'\right) = L_n\left(x; \hat{\mu}\right)' > L_n(x; \mu) = L_n^*(x; \psi)$ for all $\mu \neq \hat{\mu}$ and all $\psi \neq \hat{\psi}$. Thus the maximum-likelihood estimator $\hat{\psi} = g\left(\hat{\mu}\right)'$.

7.3 Notation

$L_n(x; \mu) = \prod_{i=1}^n f(x_i; \mu)$ represents the likelihood corresponding to n observations. $\log L_n(x; \mu)$ is the loglikelihood corresponding to n observations and

$$\frac{1}{n} \log L_n(x; \mu) = \frac{1}{n} \sum_i \log f(x_i; \mu) \tag{7.3}$$

defines the loglikelihood corresponding to a single observation. A justification for this terminology is based on the following argument. $f(x; \mu)$ is the probability density

function for the single observation x and

$$B_\mu = \int \left(-\frac{\partial^2 \log f(x;\mu)}{\partial\mu\,\partial\mu^\top}\right) f(x;\mu) = E_\mu\left(-\frac{\partial^2 \log f(x;\mu)}{\partial\mu\,\partial\mu^\top}\right) \tag{7.4}$$

is a $(k \times k)$ symmetric matrix called the information matrix. B_μ clearly applies to a single observation. Now

$$\begin{aligned}\frac{1}{n}E_\mu\left(-\frac{\partial^2 \log L_n(x;\mu)}{\partial\mu\,\partial\mu^\top}\right) &= \frac{1}{n}\sum_i E_\mu\left(-\frac{\partial^2 \log f(x_i;\mu)}{\partial\mu\,\partial\mu^\top}\right)\\ &= \frac{1}{n} n B_\mu\\ &= B_\mu.\end{aligned}$$

Thus $\frac{1}{n}\log L_n(x;\mu)$ is justified as the loglikelihood corresponding to a single observation because it contains precisely the amount of information contained in a single observation, as given in (7.4). That (7.4) may be interpreted as measuring information contained in a single observation needs some justification; this follows in section 7.5 below (Cramér-Rao Inequality).

$\hat{\mu}$ will represent the maximum-likelihood estimator of μ in Ω; $\tilde{\mu}$ the maximum-likelihood estimator of μ in ω.

The symbol $z(\mu)$ is used for $E_{\mu_0}[\log f(x;\mu)]$, that is

$$z(\mu) = \int \log f(x;\mu)\, f(x;\mu_0)\,dx = E_{\mu_0}[\log f(x;\mu)] \tag{7.5}$$

E_{μ_0} denoting expectation corresponding to the distribution of the random variable x when the true value of μ is μ_0.

H_μ (as de¯ned in section 7.2 above), $z(\mu)$ and B_μ may be evaluated at various values of μ. At $\hat{\mu}$ for example, H_μ, $z(\mu)$ and B_μ become $H_{\hat{\mu}}$, $z\left(\hat{\mu}\right)$ and $B_{\hat{\mu}}$. μ^{α} is the point in Ω at which $z(\mu)$ is greatest.

7.4 Regularity

Certain regularity conditions will be required. Those given below are not the weakest that could be used, but they are fairly simple and su± cient for the purposes of this exposition.

A.1 Potential derivatives with respect to μ of integrals over the sample space of functions of x and μ are the same as the integrals over the sample space of the partially di®erentiated functions with respect to μ.

The main implications of assumption A.1 are that the range of x should be independent of μ and that the tails of the distribution of x should allow convergence of the di®erentiated integral.

A.2 $z(\mu)$ exists for all $\mu \in \Omega$.

The whole problem of maximum-likelihood estimation, restricted or unrestricted, is closely bound up with the behavior of the function z. This is because the Law of Large Numbers ensures that, for each μ, the sequence $\left\{\frac{1}{n}\log L_n(x;\mu)\right\}$ converges, for almost all x, to $z(\mu)$. If this convergence is uniform with respect to μ, then for large n and most x, $\frac{1}{n}\log L_n(x;\mu)$ will be uniformly near z and, under appropriate

conditions, will reach its supremum in Ω near the point where z attains its supremum in Ω. The assumptions given below are designed to achieve this situation.

A.3 Γ is a convex compact subset of R^k.

A.4 $\log f(x;\cdot)$ is continuous on Γ for almost all $x \in R^n$.

A.5 For almost all $x \in R^n$ and for every $\mu \in \Gamma$, $\frac{\partial \log f(x;\mu)}{\partial \mu_i}$ $i = 1, 2, \cdots, k$ exists and its absolute value is less than some function $g(x)$ whose expectation, relative to the distribution at μ_0, is finite.

A.6 The function h is continuous on Γ.

A.7 There exists a point $\mu^* \in \Omega$ such that $z(\mu^*) > z(\mu)$ when $\mu \in \Omega$ and $\mu \neq \mu^*$.

Assumptions A.3–A.5 ensure that, for almost all x, the sequence $\left\{\frac{1}{n}\log L_n(x;\mu)\right\}$ converges to $z(\mu)$ uniformly with respect to $\mu \in \Gamma$. Assumptions A.3 and A.6 ensure that Γ is a compact subset of R^k and consequently that any continuous function on Γ attains its supremum at some point in Γ. In particular, the function $\frac{1}{n}\log L_n(x;\mu)$ attains its supremum in Ω at some point $\tilde{\mu}(x)$ in Ω, for almost all x, the sequence of values of $\tilde{\mu}(x)$ as n increases converges to μ^*. If $\mu_0 \in \Omega$, then usually μ_0 will satisfy the conditions required of μ^*, i.e. $z(\mu_0) > z(\mu)$ if $\mu \neq \mu_0$.

A.8 μ^* is an interior point of Ω.

Assumption A.8 ensures that for large n and most x, $\tilde{\mu}(x)$ will be an interior point of Ω and consequently will emerge as a solution to the restricted maximum-likelihood equations when the function h is differentiable.

The method by which the asymptotic distribution of maximum-likelihood estimates is usually derived, involves expanding the likelihood function by Taylor's Theorem. To adopt this method requires three more assumptions.

A.9 The functions h_i $(i = 1; 2; ¢¢¢; r)$ possess both ¯rst- and second-order partial derivatives which are continuous on - .

A.10 For almost all $x \in R^n$ the function $\log f(x; :)$ possesses continuous second-order partial derivatives in a neighborhood of $\mu^¤$. If μ belongs to this neighborhood, then the absolute value of $\frac{@^2 \log f(x;\mu)}{@\mu_i @\mu_j}$ $(i = 1; 2; ¢¢¢; k)$ is less than a function of x whose expectation relative to the distribution of x at μ_0 is ¯nite.

A.10 ensures that $\frac{@^2 z(\mu)}{@\mu @\mu^>}$ at $\mu^¤$ exists and that the sequence of second-order partial derivatives of $\log f(x; \mu)$ converges for almost all x to $\frac{@^2 z(\mu)}{@\mu @\mu^>}$ at $\mu = \mu^¤$.

A.11 For almost all $x \in R^n$ the function $f(x; :)$ possesses third-order partial derivatives in the neighborhood of $\mu^¤$ and if μ is in this neighborhood then the absolute value of $\frac{@^3 \log f(x;\mu)}{@\mu_i @\mu_j @\mu_k}$ $(i; j; k = 1; 2; ¢¢¢; k)$ is less than some function of x whose expectation relative to the distribution of x at μ_0 is ¯nite.

A.12 The matrix B_{μ_0} is positive de¯nite and the matrix H_{μ_0} has rank r.

7.5 The Cramér-Rao Inequality

Let $\frac{\partial \log f(x;\mu)}{\partial \mu} = s(\mu)$; $s(\mu)$ is a $(k \times 1)$ vector of functions of x, and hence is a random variable. Now

$$\int f(x;\mu)\,dx = 1$$

and hence by assumption A.1

$$\int \frac{\partial f(x;\mu)}{\partial \mu}\frac{1}{f(x;\mu)} f(x;\mu)\,dx = 0;$$

or

$$\int \frac{\partial \log f(x;\mu)}{\partial \mu_i} f(x;\mu)\,dx = 0 \qquad i = 1, 2, \ldots, k. \tag{7.6}$$

Equation (7.6) may be represented as a $(k \times 1)$ vector of integrals. These are written

$$\int \frac{\partial \log f(x;\mu)}{\partial \mu} f(x;\mu)\,dx = 0.$$

Thus $E_\mu[s(\mu)] = 0$, implying that

$$E_\mu\left[s(\mu)\, s'(\mu)\right] = D_\mu[s(\mu)] = A_\mu \tag{7.7}$$

in which D_μ is the operation denoting dispersion relative to the distribution of x when the parameter is μ. Differentiating (7.6) a second time with respect to μ yields

$$\int \frac{\partial \log f(x;\mu)}{\partial \mu}\frac{\partial f(x;\mu)}{\partial \mu'}\frac{1}{f(x;\mu)} f(x;\mu)\,dx + \int \frac{\partial^2 \log f(x;\mu)}{\partial \mu\, \partial \mu'} f(x;\mu)\,dx = 0 \tag{7.8}$$

where the integrals now represent $(k \times k)$ matrices of integrals. Since $\frac{\partial f(x;\mu)}{\partial \mu^{>}} \frac{1}{f(x;\mu)} = \frac{\partial \log f(x;\mu)}{\partial \mu^{>}}$ it follows from (7:7) and (7:8) that the first term on the left-hand side of (7:8) is precisely (7:7) while the second term is $-B_\mu$. Hence

$$A_\mu = B_\mu = E_\mu \left[-\frac{\partial^2 \log f(x;\mu)}{\partial \mu \partial \mu^{>}} \right]. \tag{7.9}$$

Let $t(x)$ be an unbiased estimator of μ. Then

$$\int t(x) f(x;\mu)\, dx = \mu$$

in which $t(x)$ and μ are each $(k \times 1)$ and hence the integral represents the integral of k separate functions. $E_\mu[t_i(x)] = \mu_i$. It follows, again from assumption A.1, that

$$\int t(x) \frac{\partial \log f(x;\mu)}{\partial \mu^{>}} f(x;\mu)\, dx = \frac{\partial \mu}{\partial \mu^{>}} = I_k.$$

Thus

$$E_\mu \left[t(x) s^{>}(\mu) \right] = I_k.$$

But since $E_\mu[s(\mu)] = 0$, $E_\mu\left[[t(x) - \mu] s^{>}(\mu) \right] = E_\mu\left[t(x) s^{>}(\mu) \right] - \mu E_\mu\left[s^{>}(\mu) \right] = I_k - 0$. Thus the covariance matrix between $t(x)$ and s is I_k. Now consider

$$D_\mu \begin{bmatrix} t(x) \\ s(\mu) \end{bmatrix} = \begin{bmatrix} D_\mu[t(x)] & E_\mu\left[[t(x) - \mu] s^{>}(\mu) \right] \\ E_\mu\left[s(\mu)[t(x) - \mu]^{>} \right] & D_\mu[s(\mu)] \end{bmatrix}$$

$$= \begin{bmatrix} D[t(x)] & I_k \\ I_k & B_\mu \end{bmatrix}$$

which, being a dispersion, is required to be at least non-negative definite. But I_k and B_μ are each positive definite whereupon, for every $q \in R^k$,

$$q^\top \begin{bmatrix} I_k \\ -B_\mu^{-1} \end{bmatrix} \begin{bmatrix} D_\mu[t(x)] & I_k \\ I_k & B_\mu \end{bmatrix} \begin{bmatrix} I_k & -B_\mu^{-1} \end{bmatrix} q = q^\top \left\{ D_\mu[t(x)] - B_\mu^{-1} \right\} q \geq 0. \tag{7.10}$$

Thus $\left\{ D_\mu[t(x)] - B_\mu^{-1} \right\}$ in (7.10) is non-negative definite, or, for any fixed $c \in R^k$, the variance of $c^\top t(x)$ is

$$V_\mu \left\{ c^\top t(x) \right\} = c^\top D_\mu[t(x)]\, c \geq c^\top B_\mu^{-1} c. \tag{7.11}$$

It follows that $c^\top B_\mu^{-1} c$ represents a lower bound below which the variance of any unbiased estimator of $c^\top \mu$ cannot go. Another way of expressing (7.10) or (7.11) is

$$\det D_\mu[t(x)] \geq \det B_\mu^{-1}.$$

Moreover, the smaller (greater) is $\det B_\mu^{-1}$, the greater (smaller) is $\det B_\mu$ implying that the greater (smaller) is the information contained in a single observation.

7.6 Maximum-likelihood Estimation in -

Many of the assumptions A.1–A.10 are concerned with establishing the existence of a supremum of $z(\mu)$ at the point $\mu^\circ \in \omega$. Ultimately interest will concentrate on the case of estimation in ω (called restricted maximum-likelihood), but to begin with, unrestricted (or free) estimation in - is considered.

It is, of course, presumed on the maintained hypothesis H_m that $\mu_0 \in \Omega$. Assumption A.3 ensures that the supremum of any continuous function on Ω is attained at some interior point of Ω while assumptions A.3 and A.4 together ensure that $\frac{1}{n}\log L_n(x;\mu)$, for almost all x, attains its supremum at a point $\hat{\mu} \in \Omega$. Finally, as has already been noted, assumptions A.3–A.5 ensure that, for almost all x, the sequence $\left\{\frac{1}{n}\log L_n(x;\mu)\right\}$ converges to $z(\mu)$ uniformly with respect to μ. Now $z(\mu)$ assumes a maximum at the point μ_0. If the distributions on Ω corresponding to different values of μ are essentially different, then for no other μ is $z(\mu)$ equal to $z(\mu_0)$. The fact that $z(\mu)$ attains a maximum at μ_0 in Ω together with the assumptions that ensure that $\frac{1}{n}\log L_n(x;\mu)$ is uniformly near $z(\mu)$ guarantee that $\frac{1}{n}\log L_n(x;\mu)$ assumes its maximum at $\hat{\mu}$ which is near μ_0. Thus as $n \to \infty$ the Strong Law of Large Numbers (SLLN) will ensure that $\hat{\mu} \xrightarrow{a.s.} \mu_0 \Rightarrow \operatorname{plim}\hat{\mu} = \mu_0$.

Since $\hat{\mu}$ is interior to Ω, it can be obtained as a solution to

$$\frac{1}{n}\frac{\partial \log L_n(x;\hat{\mu})}{\partial \mu} = 0. \tag{7.12}$$

Using assumption A.10, a Taylor's expansion around μ_0 yields, from (7.12),

$$0 = \frac{1}{n}\frac{\partial \log L_n(x;\hat{\mu})}{\partial \mu} = \frac{1}{n}\frac{\partial \log L(x;\mu_0)}{\partial \mu} + \frac{1}{n}\frac{\partial^2 \log L_n(x;\mu_0)}{\partial \mu \partial \mu^{\top}}(\hat{\mu} - \mu_0) + o_p(1). \tag{7.13}$$

By the Weak Law of Large Numbers (WLLN)

$$\operatorname{plim}\frac{1}{n}\frac{\partial^2 \log L_n(x;\mu_0)}{\partial \mu \partial \mu^{\top}} = -B_{\mu_0} \tag{7.14}$$

which by assumption A.12 is positive definite. Hence

$$\left(\hat{\mu} - \mu_0\right) = B_{\mu_0}^{-1}\frac{1}{n}\frac{\partial \log L_n(x;\mu_0)}{\partial \mu} + o_p(1). \tag{7.15}$$

Now

$$\frac{1}{n}\frac{\partial \log L_n(x;\mu_0)}{\partial \mu} = s_n(\mu) = \frac{1}{n}\sum_{i=1}\frac{\partial \log f(x_i;\mu_0)}{\partial \mu}$$

and by (7:6) and (7:9) each $\frac{\partial \log f(x_i;\mu_0)}{\partial \mu}$ is an independent random vector having mean zero and dispersion B_{μ_0}. Hence $s_n(\mu_0)$ will have mean zero and dispersion $n\left(\frac{1}{n^2}B_{\mu_0}\right) = \frac{1}{n}B_{\mu_0}$ and, by the Multivariate Central Limit Theorem,

$$\sqrt{n}\left(\frac{1}{n}\frac{\partial \log L_n(x;\mu_0)}{\partial \mu}\right), \sqrt{n}s_n(\mu_0) \rightarrow N(0; B_{\mu_0}).$$

Since B_{μ_0} does not depend on n, $\sqrt{n}\left(\hat{\mu} - \mu_0\right)$ in (7:15) is $O_p(1)$ and hence

$$\sqrt{n}\left(\hat{\mu} - \mu_0\right) \xrightarrow{d} N\left(0; B_{\mu_0}^{-1}\right). \tag{7.16}$$

Notice that the procedure that yields (7:16) as its final conclusion involves four distinct steps.

1. The imposition of sufficient regularity.

2. A Taylor's expansion applied to the maximum-likelihood equation (7:12).

3. The WLLN applied to $\frac{1}{n}\frac{\partial^2 \log L_n(x;\mu)}{\partial \mu\, \partial \mu^{>}}$.

4. The Multivariate Central Limit Theorem (MCLT) applied to $\sqrt{n}s_n(\mu_0)$.

This procedure is characteristic of all maximum-likelihood theory. Finally note that $\sqrt{n}\,(\hat{\mu} - \mu_0) \to N(0, B_{\mu_0}^{-1})$ implies that approximately, for large n,

$$\hat{\mu} \to N\left(\mu_0, \frac{1}{n} B_{\mu_0}^{-1}\right). \tag{7.17}$$

In practice $B_{\mu_0}^{-1}$ is not known, but it can be estimated consistently since, as $n \to \infty$, the distribution of $\hat{\mu}$ collapses on μ_0. Thus $B_{\mu_0}^{-1}$ is consistently estimated as $B_{\hat{\mu}}^{-1}$, as is suggested by (7:14).

7.7 Maximum-likelihood Estimation in !

To find the maximum-likelihood estimator of μ subject to the r restrictions $h(\mu) = 0$, the method of Lagrange multipliers is employed. The Lagrange function is L and

$$L = \frac{1}{n} \log L_n(x; \mu) + h^{>}(\mu)\,\lambda,$$

where λ is the $(r \times 1)$ vector of Lagrange multipliers. If $\tilde{\lambda}$ and $\tilde{\mu}$ are the restricted maximum-likelihood estimators of λ and μ, then these vectors are obtained at a stationary point on L satisfying

$$\begin{aligned} \frac{1}{n} \frac{\partial \log L_n(x; \tilde{\mu})}{\partial \mu} + H_{\tilde{\mu}}^{>} \tilde{\lambda} &= 0 \\ h(\tilde{\mu}) &= 0, \end{aligned} \tag{7.18}$$

the first of these implying $s_n(\tilde{\mu}) = -H_{\tilde{\mu}}^{>} \tilde{\lambda}$.

Since $\tilde{\mu} \xrightarrow{a.s.} \mu$ for almost all x, applying Taylor's theorem to the first equation in

(7:18) yields

$$\frac{1}{n}\frac{\partial \log L_n\left(x;\tilde{\mu}\right)}{\partial \mu} = \frac{1}{n}\frac{\partial \log L_n(x;\mu^{\alpha})}{\partial \mu} + \frac{1}{n}\frac{\partial^2 \log L_n(x;\mu^{\alpha})}{\partial \mu \partial \mu^{>}}\left(\mu - \mu^{\alpha}\right) + o_p(1) \qquad (7.19)$$

for almost all x with probability near 1. Moreover

$$H_{\tilde{\mu}}^{>}\tilde{\gamma} = H_{\mu^{\alpha}}^{>}\tilde{\gamma} + o_p(1): \qquad (7.20)$$

Equation (7:20) holds because when $\tilde{\mu}$ is near μ^{α}, it is also near μ_0 (and hence $\hat{\mu}$) so that $\tilde{\gamma}$ is relatively small. Thus expanding $H_{\tilde{\mu}}$ around μ^{α}, the first-order terms in the expansion involve $\left(\tilde{\mu} - \mu^{\alpha}\right)$ and $\tilde{\gamma}$ and so are of smaller order than those that have been included in (7:20). Equations (7:19) and (7:20) together enable the re-writing of the first equation of (7:18), to $o_p(1)$, as

$$-\frac{1}{n}\frac{\partial^2 \log L_n(x;\mu)}{\partial \mu \partial \mu^{>}}\sqrt{n}\left(\tilde{\mu} - \mu^{\alpha}\right) - H_{\mu^{\alpha}}^{>}\sqrt{n}\tilde{\gamma} = \sqrt{n}s_n(\mu^{\alpha}): \qquad (7.21)$$

The argument regarding the expansion (7:20) may also be applied to the second equation of (7:18):

$$h\left(\tilde{\mu}\right) = h(\mu^{\alpha}) + H_{\mu^{\alpha}}\left(\tilde{\mu} - \mu^{\alpha}\right) + o_p(1)$$

in which, since $\mu^{\alpha} \in \Gamma$, $h(\mu^{\alpha}) = 0$. This last expression along with (7:21) leads to the matrix equation, to $o_p(1)$,

$$\begin{bmatrix} -\frac{1}{n}\frac{\partial^2 \log L_n(x;\mu^{\alpha})}{\partial \mu \partial \mu^{>}} & -H_{\mu^{\alpha}}^{>} \\ -H_{\mu^{\alpha}} & 0 \end{bmatrix}\begin{bmatrix} \sqrt{n}\left(\tilde{\mu} - \mu^{\alpha}\right) \\ \sqrt{n}\tilde{\gamma} \end{bmatrix} = \begin{bmatrix} \sqrt{n}s_n(\mu^{\alpha}) \\ 0 \end{bmatrix} \qquad (7.22)$$

Before proceeding to apply the WLLN it is useful to consider the role of the function $z(\mu)$. As with unrestricted estimation, $z(\mu)$ plays an important role in restricted maximum-likelihood estimation.

The regularity conditions ensure that $z(\mu)$ reaches a maximum at the point μ_0 in Ω, and reaches a supremum at $z(\mu^*)$ when μ is confined to be chosen from ω. If $\mu_0 \in \omega$, then μ_0 replaces μ^* and $z(\mu_0)$ is the maximum of $z(\mu)$ in the set Ω. In this case $\frac{\partial z(\mu^*)}{\partial \mu} = \frac{\partial z(\mu_0)}{\partial \mu} = 0$ and the Lagrange multiplier that is used to constrain the choice of μ to μ in ω in any sample, must converge to zero as $n \to \infty$. This is because in the limit, as $n \to \infty$, the estimator $\tilde{\mu}$ of μ in ω must converge to its true value μ_0 in ω and restraining it to lie there eventually becomes unnecessary. In this case, then, the solutions, $\tilde{\mu}$ and $\tilde{\lambda}$, to the restrained maximum-likelihood equations (7:18) will tend to μ_0 and zero, since the null hypothesis $H_0 : h(\mu) = 0$ is undoubtedly satisfied.

What happens in a practical case when n is large and μ_0, while not belonging to ω, is nevertheless near to this set? In this case, $z(\mu_0)$ will be $\sup_{\mu \in \Omega} z(\mu)$ and $\mu^* \in \omega$ will be near to μ_0; that is, $H_0 : \mu_0 \in \omega$, while not strictly true, is in fact very close to being true. In particular, $\frac{\partial z(\mu^*)}{\partial \mu}$ will be near to the same expression for $\mu = \mu_0$ and $\frac{\partial z(\mu_0)}{\partial \mu}$ is indeed zero. Then $\tilde{\lambda}$ will be near to zero and moreover $\frac{\partial^2 z(\mu^*)}{\partial \mu \, \partial \mu^{\top}}$ will have elements near to the corresponding element of $\frac{\partial^2 z(\mu_0)}{\partial \mu \, \partial \mu^{\top}} = -B_{\mu_0}$. In short, convergence of expressions determined at μ^* will be toward the corresponding expressions determined at μ_0.

These arguments will now be applied to equation (7:22) in which it may be noted

that, by applying the MCLT,

$$\sqrt{n}s_n(\mu_0) = \frac{1}{\sqrt{n}}\frac{\partial \log L_n(x;\mu_0)}{\partial \mu} \xrightarrow{d} (0; B_{\mu_0}) \tag{7.23}$$

as before. Thus it is in order to write (7:22) as

$$\begin{bmatrix} B_{\mu_0} & -H_{\mu_0} \\ -H_{\mu_0} & 0 \end{bmatrix} \begin{bmatrix} \sqrt{n}(\hat{\mu} - \mu_0) \\ \sqrt{n}\tilde{\lambda}_n \end{bmatrix} = \begin{bmatrix} \sqrt{n}s_n(\mu_0) \\ 0 \end{bmatrix}. \tag{7.24}$$

Now

$$\begin{bmatrix} B_{\mu} & -H_{\mu_0}^{>} \\ -H_{\mu_0} & 0 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} B_{\mu_0}^{-1} - B_{\mu_0}^{-1}H_{\mu_0}^{>}\left(H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1}H_{\mu_0}B_{\mu_0}^{-1} & -B_{\mu_0}^{-1}H_{\mu_0}^{>}\left(H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1} \\ -\left(H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1}H_{\mu_0}B_{\mu_0}^{-1} & -\left(H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1} \end{bmatrix}$$

whereupon, from (7:22), (7:23) and (7:24)

$$\sqrt{n}\tilde{\lambda}_n = -\left(H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1}H_{\mu_0}B_{\mu_0}^{-1}\sqrt{n}s_n(\mu_0)$$

implying that

$$\sqrt{n}\tilde{\lambda}_n \xrightarrow{d} N\left(0; \left(H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1}H_{\mu_0}B_{\mu_0}^{-1}B_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\left(H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1}\right)$$

or

$$\sqrt{n}\tilde{\lambda}_n \xrightarrow{d} N\left(0; \left(H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1}\right). \tag{7.25}$$

In addition, $\sqrt{n}(\tilde{\mu} - \mu_0) = B_{\mu_0}^{-1} - B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1}H_{\mu_0}B_{\mu_0}^{-1}\sqrt{n}s(\mu_0)$ and hence

$$\sqrt{n}(\tilde{\mu} - \mu_0) \xrightarrow{d} N\left(0,\ \begin{matrix} B_{\mu_m}^{-1} - B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1}H_{\mu_0}B_{\mu_0}^{-1}B_{\mu_0} \\ B_{\mu_0}^{-1} - B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1}H_{\mu_0}B_{\mu_0}^{-1} \end{matrix}\right)$$

or

$$\sqrt{n}(\tilde{\mu} - \mu_0) \xrightarrow{d} N\left(0,\ B_{\mu_0} - B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1}H_{\mu_0}B_{\mu_0}^{-1}\right). \tag{7.26}$$

Finally, the covariance term between $\sqrt{n}(\tilde{\mu} - \mu_0)$ and $\sqrt{n}\tilde{\lambda}$ is

$$\begin{aligned} & \left[B_{\mu_0}^{-1} - B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1}H_{\mu_0}B_{\mu_0}^{-1}\right]B_{\mu_0}\left[B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1}\right] \qquad (7.27) \\ = \ & B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1} - B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1}H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1} \\ = \ & 0. \end{aligned}$$

Putting (7:25), (7:26) and (7:27) into one equation

$$\begin{bmatrix} \sqrt{n}(\tilde{\mu} - \mu_0) \\ \sqrt{n}\tilde{\lambda} \end{bmatrix} \xrightarrow{d} N(0, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} B_{\mu_0}^{-1} - B_{\mu_0}^{-1}H_{\mu_0}^{\top}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1}H_{\mu_0}B_{\mu_0}^{-1} & 0 \\ 0 & -\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{\top}\right]^{-1} \end{bmatrix}.$$

By assumption A.12 B_{μ_0} is positive definite. Hence there will exist a non-singular matrix Q_{μ_0} such that $Q_{\mu_0}^{\top}Q_{\mu_0} = B_{\mu_0}^{-1}$. Writing $G_{\mu_0} = Q_{\mu_0}H_{\mu_0}^{\top}$, then

$$\Sigma = \begin{bmatrix} Q_{\mu_0}^{\top}\left[I_k - G_{\mu_0}\left[G_{\mu_0}^{\top}G_{\mu_0}\right]^{-1}G_{\mu_0}^{\top}\right]Q_{\mu_0} & 0 \\ 0 & -\left[G_{\mu_0}^{\top}G_{\mu_0}\right]^{-1} \end{bmatrix}.$$

and, if $P_{\mu_0} = G_{\mu_0}\left(G_{\mu_0}^{>}G_{\mu_0}\right)^{-1}G_{\mu_0}$ and $M_{\mu_0} = (I_k - P_{\mu_0})$, then § reduces to

$$\S = \begin{bmatrix} Q_{\mu_0}^{>}M_{\mu_0}Q_{\mu_0} & 0 \\ 0 & G_{\mu_0}^{>}G_{\mu_0} \end{bmatrix}:$$

Since $Q_{\mu_0}^{>}M_{\mu_0}Q_{\mu_0}$ must be non-negative definite (since Q_{μ_0} is non-singular and M_{μ_0} is non-negative definite), § must be non-negative definite. This is reasonable since - is a k-dimensional subset of R^k which is restricted by r restraint equations; thus $\tilde{\mu} \in \Omega$ has k elements, $(k - r)$ of which are independently estimated. Thus the dispersion of $\sqrt{n}\left(\tilde{\mu} - \mu_0\right)$ would be expected to have rank $(k - r)$. Since Q_{μ_0} is non-singular, and by assumption A.12 H_{μ_0} has rank r, $Q_{\mu_0}^{>}M_{\mu_0}Q_{\mu_0}$ has rank $(k - r)$.

7.8 Associated Tests of Significance

Notice from (7:18) that

$$H_{\tilde{\mu}}^{>}\sqrt{n}\tilde{\lambda} = \frac{1}{\sqrt{n}}\frac{\partial \log L_n\left(x;\tilde{\mu}\right)}{\partial \mu} = \sqrt{n}s_n\left(\tilde{\mu}\right): \tag{7.28}$$

Since for large n the approximate distribution of $\sqrt{n}\tilde{\lambda}$ is $N\left(0; H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1}$, then

$$\sqrt{n}\tilde{\lambda}^{>}\left(\frac{1}{\sqrt{n}}\tilde{\lambda}\right)\left(\frac{1}{\sqrt{n}}\tilde{\lambda}\right)^{-1}\sqrt{n}\tilde{\lambda} = \tilde{\lambda}^{>}\left[H_{\tilde{\mu}}B_{\tilde{\mu}}^{-1}H_{\tilde{\mu}}^{>}\right]\tilde{\lambda} \tag{7.29}$$

is approximately $\chi^2(r;\pm)$ with $\pm = 0$ on H_0. This is called the Lagrange Multiplier (LM) test. Equation 7:29 is also equivalent, by (7:28), to

$$ns_n^{>}\left(\tilde{\mu}\right)B_{\tilde{\mu}}^{-1}s_n\left(\tilde{\mu}\right) \xrightarrow{d} \chi^2(r;0) \tag{7.30}$$

where $\frac{1}{\sqrt{n}} s_n(\tilde{\mu}) = \frac{1}{n} \frac{\partial \log L_n(x;\tilde{\mu})}{\partial \mu}$ is the score statistic. This implies that the LM and the score methods lead to the same test.

Also, since $h(\hat{\mu}) \simeq h(\mu_0) + H_{\mu_0}(\hat{\mu} - \mu_0)$,

$$\sqrt{n}\left[h(\hat{\mu}) - h(\mu_0)\right] \simeq H_{\mu_0} \sqrt{n}(\hat{\mu} - \mu_0) \xrightarrow{d} N\left(0, H_{\mu_0} B_{\mu_0}^{-1} H_{\mu_0}^{\top}\right) \tag{7.31}$$

and hence

$$W = nh^{\top}(\hat{\mu}) \left[H_{\hat{\mu}} B_{\hat{\mu}}^{-1} H_{\hat{\mu}}^{\top}\right]^{-1} h(\hat{\mu}) \xrightarrow{d} \chi^2(r; \pm) \tag{7.32}$$

and $\pm = 0$ on H_0. W is known as the Wald statistic.

The Wald test is associated with estimation of μ in Ω and the LM test with estimation of μ in ω. The likelihood ratio (LR) test, in contrast, is essentially a test based upon a ratio of the value of the likelihood for estimation in ω relative to the value of the likelihood for estimation in Ω. If λ is the LR, then

$$LR = \frac{L_n(x;\tilde{\mu})}{L_n(x;\hat{\mu})}. \tag{7.33}$$

Since $L_n(x;\tilde{\mu}) \leq L_n(x;\hat{\mu})$ it is clear that $0 < LR \leq 1$. The test is based on the statistic $-2\log\lambda$, thus ensuring that the statistic is never negative. The reason for the multiple 2 will become clear as the test statistic is developed below.

The first step is to expand in a Taylor's series $\log L_n(x;\tilde{\mu})$ around the point $\hat{\mu}$:

$$\log L_n(x;\tilde{\mu}) \simeq \log L_n(x;\hat{\mu}) + \frac{\partial \log L_n(x;\hat{\mu})}{\partial \mu^{\top}}(\tilde{\mu} - \hat{\mu}) + \frac{1}{2}(\tilde{\mu} - \hat{\mu})^{\top} \frac{\partial^2 \log L_n(x;\hat{\mu})}{\partial \mu \partial \mu^{\top}}(\tilde{\mu} - \hat{\mu}).$$

Noting that the second term on the right-hand side is zero,

$$\log L_n\left(x;\tilde{\mu}\right) - \log L_n\left(x;\hat{\mu}\right) = \frac{1}{2}\left(\tilde{\mu}-\hat{\mu}\right)^{>}\frac{\partial^2 \log L_n\left(x;\hat{\mu}\right)}{\partial\mu\,\partial\mu^{>}}\left(\tilde{\mu}-\hat{\mu}\right)$$

and this may be expressed as

$$-2\log\lambda = \sqrt{n}\left(\tilde{\mu}-\hat{\mu}\right)^{>} B_{\hat{\mu}}\left(\tilde{\mu}-\hat{\mu}\right)\sqrt{n}. \tag{7.34}$$

Several previously developed results will now become important: from (7:15) and $Q_{\mu_0}^{>}Q_{\mu_0} = B_{\mu_0}^{-1}$

$$\sqrt{n}\left(\hat{\mu}-\mu_0\right) = B_{\mu_0}^{-1}\sqrt{n}s_n(\mu_0) = Q_{\mu_0}^{>}Q_{\mu_0}\sqrt{n}s_n(\mu_0); \tag{7.35}$$

from the solution to (7:24)

$$\sqrt{n}\left(\tilde{\mu}-\mu_0\right) = \left[B_{\mu_0}^{-1} - B_{\mu_0}^{-1}H_{\mu_0}^{>}\left(H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right)^{-1}H_{\mu_0}B_{\mu_0}^{-1}\right]\sqrt{n}s_n(\mu_0)$$

which reduces to

$$\sqrt{n}\left(\tilde{\mu}-\mu_0\right) = Q_{\mu_0}^{>}M_{\mu_0}Q_{\mu_0}\sqrt{n}s_n(\mu_0); \tag{7.36}$$

in which $M_{\mu_0} = I_k - Q_{\mu_0}H_{\mu_0}^{>}\left(H_{\mu_0}Q_{\mu_0}^{>}Q_{\mu_0}H_{\mu_0}^{>}\right)^{-1}H_{\mu_0}Q_{\mu_0}^{>} = I_k - P_{\mu_0}$. Since $\sqrt{n}s_n(\mu_0) \xrightarrow{d} N(0;B_{\mu_0})$, $Q_{\mu_0}\sqrt{n}s_n(\mu_0) \xrightarrow{d} N(0;I_k)$ and hence taking (7:35) from (7:36)

$$\sqrt{n}\left(\tilde{\mu}-\hat{\mu}\right) = Q_{\mu_0}^{>}\left(M_{\mu_0}-I_k\right)Q_{\mu_0}\sqrt{n}s_n(\mu_0)$$

and the last expression reduces to

$$Q_{\mu_0}^{>}\left(I_k - P_{\mu_0} - I_k\right)Q_{\mu_0}\sqrt{n}s_n(\mu_0) = -Q_{\mu_0}^{>}P_{\mu_0}Q_{\mu_0}\sqrt{n}s_n(\mu_0):$$

Hence for (7:34)

$$-2\log\lambda = \sqrt{n}s_n(\mu_0)\left[Q_{\mu_0}^{>}P_{\mu_0}Q_{\mu_0}B_{\hat{\mu}}Q_{\mu_0}^{>}P_{\mu_0}Q_{\mu_0}\right]s_n(\mu_0)\sqrt{n}$$

which reduces to

$$-2\log\lambda = \sqrt{n}s_n^{>}(\mu_0)Q_{\mu_0}^{>}\left[Q_{\mu_0}H_{\mu_0}^{>}\left[H_{\mu_0}B_{\mu_0}^{-1}H_{\mu_0}^{>}\right]^{-1}H_{\mu_0}Q_{\mu_0}^{>}\right]Q_{\mu_0}\sqrt{n}s_n(\mu_0)$$

which is clearly approximately $\chi^2(r;\pm)$ with $\pm = 0$ on $H_0 : h(\mu) = 0$, since the expression in square brackets is a projection matrix of rank r and $Q_{\mu_0}\sqrt{n}s_n(\mu_0) \sim N(0; I_k)$.