

1. MINIMUM DISTANCE ESTIMATION AND MINIMUM VARIANCE ESTIMATION

The Decomposition of the Sum of Squares

Ordinary least-squares regression entails the decomposition the vector y into two mutually orthogonal components. These are the vector $Py = X(X'X)^{-1}X'y = X\hat{\beta}$, which estimates the systematic component of the regression equation, and the residual vector $e = y - X\hat{\beta}$, which estimates the disturbance vector ε . The condition that e should be orthogonal to the manifold of X in which the systematic component resides, such that $X'e = X'(y - X\hat{\beta}) = 0$, is the condition which is expressed by the normal equations, which are written more commonly as $X'X\beta = X'y$.

Corresponding to the decomposition of y , there is a decomposition of the sum of squares $y'y$. To express the latter, let us write $X\hat{\beta} = Py$ and $e = y - X\hat{\beta} = (I - P)y$. Then, in consequence of the condition $P = P' = P^2$ and the equivalent condition $P'(I - P) = 0$, it follows that

$$\begin{aligned}
 (1.1) \quad y'y &= \{Py + (I - P)y\}'\{Py + (I - P)y\} \\
 &= y'Py + y'(I - P)y \\
 &= \hat{\beta}'X'X\hat{\beta} + e'e.
 \end{aligned}$$

This is simply an instance of Pythagoras theorem; and the identity is expressed by saying that the total sum of squares $y'y$ is equal to the regression sum of squares $\hat{\beta}'X'X\hat{\beta}$ plus the residual or error sum of squares $e'e$. A geometric interpretation of the orthogonal decomposition of y and of the resulting Pythagorean relationship is given in Figure 1.

It is clear from intuition that, by projecting y perpendicularly onto the manifold of X , the distance between y and $Py = X\hat{\beta}$ is minimised. In order to establish this point formally, imagine that $\gamma = Pg$ is an arbitrary vector in the manifold of X . Then the Euclidean distance from y to γ cannot be less than the distance from y to $X\hat{\beta}$. The square of the former distance is

$$\begin{aligned}
 (1.2) \quad (y - \gamma)'(y - \gamma) &= \{(y - X\hat{\beta}) + (X\hat{\beta} - \gamma)\}'\{(y - X\hat{\beta}) + (X\hat{\beta} - \gamma)\} \\
 &= \{(I - P)y + P(y - g)\}'\{(I - P)y + P(y - g)\}.
 \end{aligned}$$

The properties of the projector P which have been used in simplifying equation (1), indicate that

$$\begin{aligned}
 (1.3) \quad (y - \gamma)'(y - \gamma) &= y'(I - P)y + (y - g)'P(y - g) \\
 &= e'e + (X\hat{\beta} - \gamma)'(X\hat{\beta} - \gamma).
 \end{aligned}$$

Since the squared distance $(X\hat{\beta} - \gamma)'(X\hat{\beta} - \gamma)$ is nonnegative, it follows that $(y - \gamma)'(y - \gamma) \geq e'e$, where $e = y - X\hat{\beta}$; and this proves the assertion.

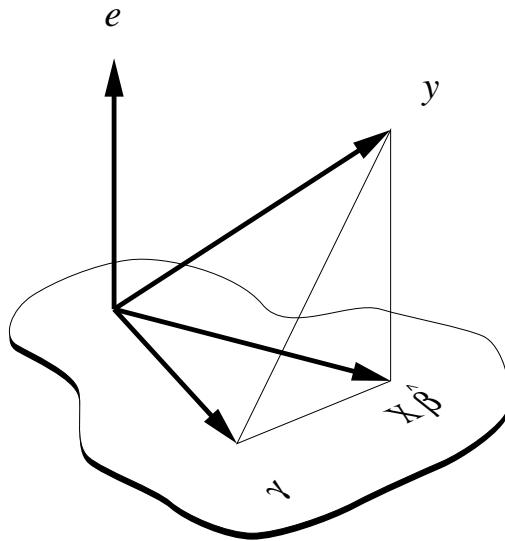


Figure 1. The vector $Py = X\hat{\beta}$ is formed by the orthogonal projection of the vector y onto the subspace spanned by the columns of the matrix X .

Some Statistical Properties of the Estimator

The expectation or mean vector of $\hat{\beta}$, and its dispersion matrix as well, may be found from the expression

$$(1.4) \quad \begin{aligned} \hat{\beta} &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon. \end{aligned}$$

On the assumption that the elements of X are nonstochastic, the expectation is given by

$$(1.5) \quad \begin{aligned} E(\hat{\beta}) &= \beta + (X'X)^{-1}X'E(\varepsilon) \\ &= \beta. \end{aligned}$$

Thus, $\hat{\beta}$ is an unbiased estimator. The deviation of $\hat{\beta}$ from its expected value is $\hat{\beta} - E(\hat{\beta}) = (X'X)^{-1}X'\varepsilon$. Therefore the dispersion matrix, which contains the variances and covariances of the elements of $\hat{\beta}$, is

$$(1.6) \quad \begin{aligned} D(\hat{\beta}) &= E\left[\{\hat{\beta} - E(\hat{\beta})\}\{\hat{\beta} - E(\hat{\beta})\}'\right] \\ &= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

The Gauss–Markov theorem asserts that $\hat{\beta}$ is the unbiased linear estimator of least dispersion. This dispersion is usually characterised in terms of the variance of an arbitrary linear combination of the elements of $\hat{\beta}$, although it may also be characterised in terms of the determinant of the dispersion matrix $D(\hat{\beta})$. Thus,

(1.7) If $\hat{\beta}$ is the ordinary least-squares estimator of β in the classical linear regression model, and if β^* is any other linear unbiased estimator of β , then $V(q'\beta^*) \geq V(q'\hat{\beta})$ where q is any constant vector of the appropriate order.

Proof. Since $\beta^* = Ay$ is an unbiased estimator, it follows that $E(\beta^*) = AE(y) = AX\beta = \beta$, which implies that $AX = I$. Now set $A = (X'X)^{-1}X' + G$. Then $AX = I$ implies that $GX = 0$. Given that $D(y) = D(\varepsilon) = \sigma^2 I$, it follows that

$$(1.8) \quad \begin{aligned} D(\beta^*) &= AD(y)A' \\ &= \sigma^2 \{(X'X)^{-1}X' + G\} \{X(X'X)^{-1} + G'\} \\ &= \sigma^2(X'X)^{-1} + \sigma^2 GG' \\ &= D(\hat{\beta}) + \sigma^2 GG'. \end{aligned}$$

Therefore, for any constant vector q of order k , there is the identity

$$(1.9) \quad \begin{aligned} V(q'\beta^*) &= q'D(\hat{\beta})q + \sigma^2 q'GG'q \\ &\geq q'D(\hat{\beta})q = V(q'\hat{\beta}); \end{aligned}$$

and thus the inequality $V(q'\beta^*) \geq V(q'\hat{\beta})$ is established.

2. THE PARTITIONED REGRESSION MODEL

Consider taking a regression equation in the form of

$$(2.1) \quad y = [X_1 \quad X_2] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

Here $[X_1, X_2] = X$ and $[\beta'_1, \beta'_2]' = \beta$ are obtained by partitioning the matrix X and vector β of the equation $y = X\beta + \varepsilon$ in a conformable manner. The normal equations $X'X\beta = X'y$ can be partitioned likewise. Writing the equations without the surrounding matrix braces gives

$$(2.2) \quad X'_1 X_1 \beta_1 + X'_1 X_2 \beta_2 = X'_1 y,$$

$$(2.3) \quad X'_2 X_1 \beta_1 + X'_2 X_2 \beta_2 = X'_2 y.$$

To obtain an expression for $\hat{\beta}_2$, we must eliminate β_1 from equation (2.3). For this purpose, we multiply equation (2.2) by $X'_2 X_1 (X'_1 X_1)^{-1}$ to give

$$(2.4) \quad X'_2 X_1 \beta_1 + X'_2 X_1 (X'_1 X_1)^{-1} X'_1 X_2 \beta_2 = X'_2 X_1 (X'_1 X_1)^{-1} X'_1 y.$$

When the latter is taken from equation (2.3), we get

$$(2.5) \quad \left\{ X'_2 X_2 - X'_2 X_1 (X'_1 X_1)^{-1} X'_1 X_2 \right\} \beta_2 = X'_2 y - X'_2 X_1 (X'_1 X_1)^{-1} X'_1 y.$$

On defining

$$(2.6) \quad P_1 = X_1(X_1'X_1)^{-1}X_1' \quad \text{and} \quad P_2 = X_2(X_2'X_2)^{-1}X_2',$$

can we rewrite (2.5) as

$$(2.7) \quad \left\{ X_2'(I - P_1)X_2 \right\} \beta_2 = X_2'(I - P_1)y,$$

whence

$$(2.8) \quad \hat{\beta}_2 = \left\{ X_2'(I - P_1)X_2 \right\}^{-1} X_2'(I - P_1)y.$$

The analogous estimator for β_1 may be obtained from (2.8) simply by interchanging the subscripts 1 and 2. However, an alternative form of the estimator may be obtained directly from (2.2). This is

$$(2.9) \quad \hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2).$$

Some Algebraic Identities

we shall now create some further notation that will enable us to record some useful algebraic identities. We may begin with the following

$$(2.10) \quad X\hat{\beta} = X(X'X)^{-1}X'y = Py$$

$$(2.11) \quad X_1\hat{\beta}_1 = X_1 \left\{ X_1'(I - P_2)X_1 \right\}^{-1} X_1'(I - P_2)y = P_{1/2}y$$

$$(2.12) \quad X_2\hat{\beta}_2 = X_2 \left\{ X_2'(I - P_1)X_2 \right\}^{-1} X_2'(I - P_1)y = P_{2/1}y.$$

In these terms, the identity $X\hat{\beta} = X_1\hat{\beta}_1 + X_2\hat{\beta}_2$ becomes $Py = P_{1/2}y + P_{2/1}y$ and, since this holds for all values of y , we may record that

$$(2.13) \quad P = P_{1/2} + P_{2/1}.$$

Now consider using (2.9) to write $X_1\hat{\beta}_1 = P_1(y - X_2\hat{\beta}_2)$. Using (2.11) and (2.12), this can be written as $P_{1/2}y = P_1(I - P_{2/1})y$ and, since this holds for all values of y , we have

$$(2.14) \quad P_{1/2} = P_1(I - P_{2/1}) = P_1 - P_1P_{2/1}.$$

Adding $P_{2/1}$ to both sides of this gives $P_{1/2} + P_{2/1} = P = P_1 + (I - P_1)P_{2/1}$, from which we get

$$(2.15) \quad P - P_1 = (I - P_1)P_{2/1}.$$

This identity may be written more explicitly as

$$(2.16) \quad X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1' = (I - P_1)X_2 \left\{ X_2'(I - P_1)X_2 \right\}^{-1} X_2'(I - P_1).$$

Regression with an Intercept

We may use the expressions for $\hat{\beta}_1$, $\hat{\beta}_2$ in (2.8) and (2.9) in order to find estimators of the parameters β_1 , β_Z of the regression model $(y, i\beta_i + Z\beta_Z, \sigma^2 I)$. For this purpose, we assimilate the equations $y = i\beta_i + Z\beta_Z + \varepsilon$ to the equations $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ by setting $X_1 = i$, $X_2 = Z$, $\beta_1 = \beta_i$ and $\beta_2' = \beta_Z'$.

To assist us in finding the formulae for the estimators, let us consider the projector

$$(2.17) \quad P_1 = P_i = i(i'i)^{-1}i'y = ii'/T.$$

Applying this to the vector y , we get

$$(2.18) \quad P_i y = i(i'y/T) = i \left(\sum y_t/T \right) = i\bar{y}.$$

Applying it likewise to the matrix Z , we get

$$(2.19) \quad P_i X = i \left[\sum x_{t2}, \dots, \sum x_{tk} \right] /T = i[\bar{x}_2, \dots, \bar{x}_k] = \bar{Z}.$$

On substituting $Z = X_2$ and $i = X_1$ in the formula for $\hat{\beta}_2 = \beta_Z$ in (2.8), and using the identity $(I - P_i) = (I - P_i)'(I - P_i)$ and the notation $\bar{y} = P_i y$, $\bar{Z} = P_i Z$, we obtain

$$(2.20) \quad \begin{aligned} \hat{\beta}_Z &= \{Z'(I - P_i)Z\}^{-1}Z'(I - P_i)y \\ &= \{(Z - \bar{Z})'(Z - \bar{Z})\}^{-1}(Z - \bar{Z})'(y - \bar{y}). \end{aligned}$$

Thus the coefficients β_2, \dots, β_k may be estimated by applying ordinary least-squares regression to data which has been adjusted by subtracting from each observation its respective sample mean.

To find an estimate of the intercept term $\beta_i = \beta_1$, we substitute $i = X_1$ and $Z = X_2$ in the formula for $\hat{\beta}_1$ in (2.9) to get

$$(2.21) \quad \begin{aligned} \hat{\beta}_1 &= (i'i)^{-1}i'y - (i'i)^{-1}i'Z\{Z'(I - P_i)Z\}^{-1}Z'(I - P_i)y \\ &= i'y/T - i'Z\hat{\beta}_Z/T \\ &= \bar{y} - \sum_{j=2}^k \hat{\beta}_j \bar{x}_j. \end{aligned}$$

Coefficients of Determination

To provide a summary measure of the extent to which the ordinary least-squares regression accounts for the observed vector y , we may use the ordinary coefficient of determination. This is defined by

$$(2.22) \quad R^2(y, X) = \frac{y'Py}{y'y} = \frac{\hat{\beta}X'X\hat{\beta}}{y'y}.$$

If $y \in \mathcal{M}(X)$, then $X\hat{\beta} = Py = y$; and it follows that $R^2 = 1$. The value of y is then completely accounted for by the regression. If y is distributed continuously in \mathcal{R}^T , then the event $y \in \mathcal{M}(X)$ has a probability measure of zero unless $\mathcal{M}(X) = \mathcal{R}^T$, in which case the event is a certainty. The condition $\mathcal{M}(X) = \mathcal{R}^T$ is equivalent to the condition $\text{Null}(X) = 0$ which means that X has full row rank which, in turn, implies that the number of rows in X cannot exceed the number of columns. For the parameter vector β to be estimable, the condition $\text{Null}(X) = 0$ must be fulfilled. Thus X must have full column rank, and the number of columns must not exceed the number of rows. It follows that we can expect the regression to yield both a coefficient of determination of unity and a uniquely determined estimate $\hat{\beta}$ if and only if X is a non-singular square matrix comprising equal numbers of variables and observations.

If $y \perp \mathcal{M}(X)$ or, equivalently, $y \in \mathcal{N}(X)$, then $Py = 0$; and it follows that $R^2 = 0$. Then the regression fails to account for any part of y . However, on the assumption that y is distributed continuously in \mathcal{R}^T , the event $y \perp \mathcal{M}(X)$ has a probability measure of zero, and thus we would never expect to find $R^2 = 0$ in practice.

The inequality $0 \leq R^2 \leq 1$ also follows from the properties of cosines once we recognize that R^2 is the cosine of the angle between the vectors y and Py .

We may also wish to measure the peculiar contribution of the variables in X_1 to the explanation of y when y is regressed on $X = [X_1, X_2]$. To do so, we must remove from y the component that is attributable to X_2 by subtracting P_2y to give $(I - P_2)y$. We must also find the components that are peculiar to X_1 by subtracting P_2X_1 to give $(I - P_2)X_1$. We can then obtain a measure of the contribution by finding the ordinary coefficient of determination $R^2\{(I - P_2)y, (I - P_2)X_1\}$ of the regression of $(I - P_2)y$ on $(I - P_2)X_1$. In the context of the regression of y on X , this is called the partial coefficient of determination of y and X_2 given X_1 and is denoted by $R^2(y, X_2|X_1)$. Using the symmetry and idempotency of $I - P_2$ and the identity of (16), we find that

$$(2.23) \quad \begin{aligned} R^2(y, X_2|X_1) &= \frac{y'(I - P_1)X_2\{X_2(I - P_1)X_2\}^{-1}X_2'(I - P_1)y}{y'(I - P_1)'(I - P_1)y} \\ &= \frac{y'(P - P_1)y}{y'(I - P_1)y}. \end{aligned}$$

In the case of the model $(y, i\beta_i + Z\beta_Z, \sigma^2 I)$, which we also write as $(y, X\beta, \sigma^2 I)$, where $X = [i, Z]$ and $\beta' = [\beta_i, \beta'_Z]$, it is conventional to measure the explanatory power of the regression in terms of the partial coefficient of determination $R^2(y, Z|i)$. This practice is justified by the argument that the explanatory power of the vector i is

given for free. Using the notations $P_i y = \bar{y}$ and $P_i Z = \bar{Z}$ of (7.34) and (7.35) respectively, we find from (7.40) that

$$(2.24) \quad \begin{aligned} R^2(y, Z|i) &= \frac{(y - \bar{y})'(Z - \bar{Z})\{(Z - \bar{Z})'(Z - \bar{Z})\}^{-1}(Z - \bar{Z})'(y - \bar{y})}{(y - \bar{y})'(y - \bar{y})} \\ &= \frac{y'P_i y - y'P_i y}{y'(I - P_i)y} = \frac{\hat{\beta}X'X\hat{\beta} - \bar{y}'\bar{y}}{y'y - \bar{y}'\bar{y}}. \end{aligned}$$

The first equality shows that $R^2(y, Z|i)$ is the ordinary coefficient of determination of the regression (7.36) wherein the variables are the deviations of the observations about their sample means. The final term, which suggests a straightforward way of computing the coefficient, has an interesting comparison with $R^2(y, X) = \hat{\beta}X'X\hat{\beta}/y'y$ defined in (22).

3. DIAGONALISATION OF A SYMMETRIC MATRIX

The Geometry of Quadratic Forms

The Circle. Let the coordinates of the points in the Cartesian plane be denoted by (z_1, z_2) . Then the equation of a circle of radius r centred on the origin is just

$$(3.1) \quad z_1^2 + z_2^2 = r^2.$$

This follows immediately from Pythagoras. The so-called parametric equations for the coordinates of the circle are

$$(3.2) \quad z_1 = r \cos(\omega), \quad \text{and} \quad z_2 = r \sin(\omega).$$

The Ellipse. The equation of an ellipse whose principal axes are aligned with those of the coordinate system in the (y_1, y_2) plane is

$$(3.3) \quad \lambda_1 y_1^2 + \lambda_2 y_2^2 = r^2,$$

On setting $\lambda_1 y_1^2 = z_1^2$ and $\lambda_2 y_2^2 = z_2^2$, we can see that

$$(3.4) \quad y_1 = \frac{z_1}{\sqrt{\lambda_1}} = \frac{r}{\sqrt{\lambda_1}} \cos(\omega), \quad y_2 = \frac{z_2}{\sqrt{\lambda_2}} = \frac{r}{\sqrt{\lambda_2}} \sin(\omega).$$

We can write equation (3.6) in matrix notation as

$$(3.5) \quad r^2 = [y_1 \quad y_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = z_1^2 + z_2^2.$$

This implies

$$(3.6) \quad \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

and

$$(3.7) \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

The Oblique Ellipse. An oblique ellipse is one whose principal axes are not aligned with those of the coordinate system. Its general equation is

$$(3.8) \quad a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = r^2;$$

which is subject to the condition that $a_{11}a_{22} - a_{12}^2 > 0$. We can write this in matrix notation:

$$(3.9) \quad \begin{aligned} r^2 &= [x_1 \ x_2] \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= [y_1 \ y_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = z_1^2 + z_2^2, \end{aligned}$$

where θ is the angle which the principal axis of the ellipse makes with the horizontal. The coefficients of the equation (3.8) are the elements of the matrix

$$(3.10) \quad \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} \lambda_1 \cos^2 \theta + \lambda_2 \sin^2 \theta & (\lambda_2 - \lambda_1) \cos \theta \sin \theta \\ (\lambda_2 - \lambda_1) \cos \theta \sin \theta & \lambda_1 \sin^2 \theta + \lambda_2 \cos^2 \theta \end{bmatrix}.$$

Notice that, if $\lambda_1 = \lambda_2$, which is to say that both axes are rescaled by the same factor, then the equation is that of a circle of radius λ_1 , and the rotation of the circle has no effect.

The mapping from the ellipse to the circle is

$$(3.11) \quad \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1}(x_1 \cos \theta - x_2 \sin \theta) \\ \sqrt{\lambda_2}(x_1 \sin \theta + x_2 \cos \theta) \end{bmatrix},$$

and the inverse mapping, from the circle to the ellipse, is

$$(3.12) \quad \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

We see from the latter that the circle is converted to an oblique ellipse via two operations. The first is an operation of scaling which produces an ellipse whose principal axes are aligned with those of the coordinate system. The second operation is a rotation which tilts the ellipse.

The vectors of the matrix that effects the rotation define the axes of the ellipse. They have the property that, when they are mapped through the matrix A , their orientation is preserved and only their length is altered. Thus

$$(3.13) \quad \begin{aligned} &\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \cos \theta \\ -\sin \theta \end{bmatrix} \\ &= \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \cos \theta \\ -\sin \theta \end{bmatrix} \\ &= \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \lambda_1 \begin{bmatrix} \cos \theta \\ -\sin \theta \end{bmatrix}. \end{aligned}$$

Such vectors are described as the characteristic vectors of the matrix, and the factors λ_1 and λ_2 , by which their lengths are altered under the transformation, are described as the corresponding characteristic roots.

Characteristic Roots and Characteristic Vectors

Let A be an $n \times n$ symmetric matrix such that $A = A'$, and imagine that the scalar λ and the vector x satisfy the equation $Ax = \lambda x$. Then λ is a characteristic root of A and x is a corresponding characteristic vector. We also refer to characteristic roots as latent roots or eigenvalues. The characteristic vectors are also called eigenvectors.

(3.14) The characteristic vectors corresponding to two distinct characteristic roots are orthogonal. Thus, if $Ax_1 = \lambda_1 x_1$ and $Ax_2 = \lambda_2 x_2$ with $\lambda_1 \neq \lambda_2$, then $x_1' x_2 = 0$.

Proof. Premultiplying the defining equations by x_2' and x_1' respectively, gives $x_2' Ax_1 = \lambda_1 x_2' x_1$ and $x_1' Ax_2 = \lambda_2 x_1' x_2$. But $A = A'$ implies that $x_2' Ax_1 = x_1' Ax_2$, whence $\lambda_1 x_2' x_1 = \lambda_2 x_1' x_2$. Since $\lambda_1 \neq \lambda_2$, it must be that $x_1' x_2 = 0$.

The characteristic vector corresponding to a particular root is defined only up to a factor of proportionality. For let x be a characteristic vector of A such that $Ax = \lambda x$. Then multiplying the equation by a scalar μ gives $A(\mu x) = \lambda(\mu x)$ or $Ay = \lambda y$; so $y = \mu x$ is another characteristic vector corresponding to λ .

(3.15) If $P = P' = P^2$ is a symmetric idempotent matrix, then its characteristic roots can take only the values of 0 and 1.

Proof. Since $P = P^2$, it follows that, if $Px = \lambda x$, then $P^2x = \lambda x$ or $P(Px) = P(\lambda x) = \lambda^2 x = \lambda x$, which implies that $\lambda = \lambda^2$. This is possible only when $\lambda = 0, 1$.

Diagonalisation of a Symmetric Matrix

Let A be an $n \times n$ symmetric matrix, and let x_1, \dots, x_n be a set of n linearly independent characteristic vectors corresponding to its roots $\lambda_1, \dots, \lambda_n$. Then we can form a set of normalised vectors

$$c_1 = \frac{x_1}{\sqrt{x_1' x_1}}, \dots, c_n = \frac{x_n}{\sqrt{x_n' x_n}},$$

which have the property that

$$c_i' c_j = \begin{cases} 0, & \text{if } i \neq j; \\ 1, & \text{if } i = j. \end{cases}$$

The first of these reflects the condition that $x_i' x_j = 0$. It follows that $C = [c_1, \dots, c_n]$ is an orthonormal matrix such that $C'C = CC' = I$.

Now consider the equation $A[c_1, \dots, c_n] = [\lambda_1 c_1, \dots, \lambda_n c_n]$ which can also be written as $AC = C\Lambda$ where $\Lambda = \text{Diag}\{\lambda_1, \dots, \lambda_n\}$ is the matrix with λ_i as its i th diagonal elements

and with zeros in the non-diagonal positions. Postmultiplying the equation by C' gives $ACC' = A = C\Lambda C'$; and premultiplying by C' gives $C'AC = C'C\Lambda = \Lambda$. Thus $A = C\Lambda C'$ and $C'AC = \Lambda$; and C is effective in diagonalising A .

Let D be a diagonal matrix whose i th diagonal element is $1/\sqrt{\lambda_i}$ so that $D'D = \Lambda^{-1}$ and $D'\Lambda D = I$. Premultiplying the equation $C'AC = \Lambda$ by D' and postmultiplying it by D gives $D'C'ACD = D'\Lambda D = I$ or $TAT' = I$, where $T = D'C'$. Also, $T'T = CDD'C' = C\Lambda^{-1}C' = A^{-1}$. Thus we have shown that

(3.16) For any symmetric matrix $A = A'$, there exists a matrix T such that $TAT' = I$ and $T'T = A^{-1}$.

4. COCHRANE'S THEOREM: THE DECOMPOSITION OF A CHI-SQUARE

The standard test of an hypothesis regarding the vector β in the model $N(y; X\beta, \sigma^2 I)$ entails a multi-dimensional version of Pythagoras' Theorem. Consider the decomposition of the vector y into the systematic component and the residual vector. This gives

$$(4.1) \quad \begin{aligned} y &= X\hat{\beta} + (y - X\hat{\beta}) \quad \text{and} \\ y - X\beta &= (X\hat{\beta} - X\beta) + (y - X\hat{\beta}), \end{aligned}$$

where the second equation comes from subtracting the unknown mean vector $X\beta$ from both sides of the first. These equations can also be expressed in terms of the projector $P = X(X'X)^{-1}X'$ which gives $Py = X\hat{\beta}$ and $(I - P)y = y - X\hat{\beta} = e$. Using the definition $\varepsilon = y - X\beta$ within the second of the equations, we have

$$(4.2) \quad \begin{aligned} y &= Py + (I - P)y \quad \text{and} \\ \varepsilon &= P\varepsilon + (I - P)\varepsilon. \end{aligned}$$

The reason for rendering the equations in this notation is that it enables us to envisage more clearly the Pythagorean relationship between the vectors. Thus, from the condition that $P = P' = P^2$, which is equivalent to the condition that $P'(I - P) = 0$, it can be established that

$$(4.3) \quad \begin{aligned} \varepsilon'\varepsilon &= \varepsilon'P\varepsilon + \varepsilon'(I - P)\varepsilon \quad \text{or} \\ \varepsilon'\varepsilon &= (X\hat{\beta} - X\beta)'(X\hat{\beta} - X\beta) + (y - X\hat{\beta})'(y - X\hat{\beta}). \end{aligned}$$

The terms in these expressions represent squared lengths; and the vectors themselves form the sides of a right-angled triangle with $P\varepsilon$ at the base, $(I - P)\varepsilon$ as the vertical side and ε as the hypotenuse.

The usual test of an hypothesis regarding the elements of the vector β is based on the foregoing relationships. Imagine that the hypothesis postulates that the true value of

the parameter vector is β_0 . To test this notion, we compare the value of $X\beta_0$ with the estimated mean vector $X\hat{\beta}$. The test is a matter of assessing the proximity of the two vectors which is measured by the square of the distance which separates them. This is given by $\varepsilon'P\varepsilon = (X\hat{\beta} - X\beta_0)'(X\hat{\beta} - X\beta_0)$. If the hypothesis is untrue and if $X\beta_0$ is remote from the true value of $X\beta$, then the distance is liable to be excessive. The distance can only be assessed in comparison with the variance σ^2 of the disturbance term or with an estimate thereof. Usually, one has to make do with the estimate of σ^2 which is provided by

$$(4.4) \quad \begin{aligned} \hat{\sigma}^2 &= \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T - k} \\ &= \frac{\varepsilon'(I - P)\varepsilon}{T - k}. \end{aligned}$$

The numerator of this estimate is simply the squared length of the vector $e = (I - P)y = (I - P)\varepsilon$ which constitutes the vertical side of the right-angled triangle.

The test uses the result that

$$(4.5) \quad \text{If } y \sim N(X\beta, \sigma^2 I) \text{ and if } \hat{\beta} = (X'X)^{-1}X'y, \text{ then}$$

$$F = \left\{ \frac{(X\hat{\beta} - X\beta)'(X\hat{\beta} - X\beta)}{k} \middle/ \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T - k} \right\}$$

is distributed as an $F(k, T - k)$ statistic.

This result depends upon Cochrane's Theorem concerning the decomposition of a chi-square random variate. The following is a statement of the theorem which is attuned to our present requirements:

$$(4.6) \quad \text{Let } \varepsilon \sim N(0, \sigma^2 I_T) \text{ be a random vector of } T \text{ independently and identically distributed elements. Also let } P = X(X'X)^{-1}X' \text{ be a symmetric idempotent matrix, such that } P = P' = P^2, \text{ which is constructed from a matrix } X \text{ of order } T \times k \text{ with } \text{Rank}(X) = k. \text{ Then}$$

$$\frac{\varepsilon'P\varepsilon}{\sigma^2} + \frac{\varepsilon'(I - P)\varepsilon}{\sigma^2} = \frac{\varepsilon'\varepsilon}{\sigma^2} \sim \chi^2(T),$$

which is a chi-square variate of T degrees of freedom, represents the sum of two independent chi-square variates $\varepsilon'P\varepsilon/\sigma^2 \sim \chi^2(k)$ and $\varepsilon'(I - P)\varepsilon/\sigma^2 \sim \chi^2(T - k)$ of k and $T - k$ degrees of freedom respectively.

To prove this result, we begin by finding an alternative expression for the projector $P = X(X'X)^{-1}X'$. First consider the fact that $X'X$ is a symmetric positive-definite matrix. It follows that there exists a matrix transformation T such that $T(X'X)T' = I$ and $T'T = (X'X)^{-1}$. Therefore $P = XT'TX' = C_1C_1'$, where $C_1 = XT'$ is a $T \times k$ matrix comprising k orthonormal vectors such that $C_1'C_1 = I_k$ is the identity matrix of order k .

Now define C_2 to be a complementary matrix of $T - k$ orthonormal vectors. Then $C = [C_1, C_2]$ is an orthonormal matrix of order T such that

$$(4.7) \quad \begin{aligned} CC' &= C_1C_1' + C_2C_2' = I_T \quad \text{and} \\ C'C &= \begin{bmatrix} C_1'C_1 & C_1'C_2 \\ C_2'C_1 & C_2'C_2 \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ 0 & I_{T-k} \end{bmatrix}. \end{aligned}$$

The first of these results allows us to set $I - P = I - C_1C_1' = C_2C_2'$. Now, if $\varepsilon \sim N(0, \sigma^2 I_T)$ and if C is an orthonormal matrix such that $C'C = I_T$, then it follows that $C'\varepsilon \sim N(0, \sigma^2 I_T)$. In effect, if ε is a normally distributed random vector with a density function which is centred on zero and which has spherical contours, and if C is the matrix of a rotation, then nothing is altered by applying the rotation to the random vector. On partitioning $C'\varepsilon$, we find that

$$(4.8) \quad \begin{bmatrix} C_1'\varepsilon \\ C_2'\varepsilon \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 I_k & 0 \\ 0 & \sigma^2 I_{T-k} \end{bmatrix} \right),$$

which is to say that $C_1'\varepsilon \sim N(0, \sigma^2 I_k)$ and $C_2'\varepsilon \sim N(0, \sigma^2 I_{T-k})$ are independently distributed normal vectors. It follows that

$$(4.9) \quad \begin{aligned} \frac{\varepsilon' C_1 C_1' \varepsilon}{\sigma^2} &= \frac{\varepsilon' P \varepsilon}{\sigma^2} \sim \chi^2(k) \quad \text{and} \\ \frac{\varepsilon' C_2 C_2' \varepsilon}{\sigma^2} &= \frac{\varepsilon' (I - P) \varepsilon}{\sigma^2} \sim \chi^2(T - k) \end{aligned}$$

are independent chi-square variates. Since $C_1 C_1' + C_2 C_2' = I_T$, the sum of these two variates is

$$(4.10) \quad \frac{\varepsilon' C_1 C_1' \varepsilon}{\sigma^2} + \frac{\varepsilon' C_2 C_2' \varepsilon}{\sigma^2} = \frac{\varepsilon' \varepsilon}{\sigma^2} \sim \chi^2(T);$$

and thus the theorem is proved.

The statistic under (4.5) can now be expressed in the form of

$$(4.11) \quad F = \left\{ \frac{\varepsilon' P \varepsilon}{k} \middle/ \frac{\varepsilon' (I - P) \varepsilon}{T - k} \right\}.$$

This is manifestly the ratio of two chi-square variates divided by their respective degrees of freedom; and so it has an F distribution with these degrees of freedom. This result provides the means for testing the hypothesis concerning the parameter vector β .

5. TESTING HYPOTHESES CONCERNING THE CLASSICAL LINEAR REGRESSION MODEL

The Normal Distribution and the Sampling Distributions

It is often appropriate to assume that the elements of the disturbance vector ε within the regression equations $y = X\beta + \varepsilon$ are distributed independently and identically according to a normal law. Under this assumption, the sampling distributions of the estimates may be derived and various hypotheses relating to the underlying parameters may be tested.

To denote that x is a normally distributed random variable with a mean of $E(x) = \mu$ and a dispersion matrix of $D(x) = \Sigma$, we shall write $x \sim N(\mu, \Sigma)$. A vector $z \sim N(0, I)$ with a mean of zero and a dispersion matrix of $D(z) = I$ is described as a standard normal vector. Any normal vector $x \sim N(\mu, \Sigma)$ can be standardised:

$$(5.1) \quad \text{If } T \text{ is a transformation such that } T\Sigma T' = I \text{ and } T'T = \Sigma^{-1}, \text{ then } T(x - \mu) \sim N(0, I).$$

Associated with the normal distribution are a variety of so-called sampling distributions which occur frequently in problems of statistical inference. Amongst these are the chi-square distribution, the F distribution and the t distribution.

If $z \sim N(0, I)$ is a standard normal vector of n elements, then the sum of squares of its elements has a chi-square distribution of n degrees of freedom; and this is denoted by $z'z \sim \chi^2(n)$. With the help of the standardising transformation, it can be shown that,

$$(5.2) \quad \text{If } x \sim N(\mu, \Sigma) \text{ is a vector of order } n, \text{ then } (x - \mu)' \Sigma^{-1} (x - \mu) \sim \chi^2(n).$$

The sum of any two independent chi-square variates is itself a chi-square variate whose degrees of freedom equal the sum of the degrees of freedom of its constituents. Thus,

$$(5.3) \quad \text{If } u \sim \chi^2(m) \text{ and } v \sim \chi^2(n) \text{ are independent chi-square variates of } m \text{ and } n \text{ degrees of freedom respectively, then } (u + v) \sim \chi^2(m + n) \text{ is a chi-square variate of } m + n \text{ degrees of freedom.}$$

The ratio of two independent chi-square variates divided by their respective degrees of freedom has a F distribution which is completely characterised by these degrees of freedom. Thus,

$$(5.4) \quad \text{If } u \sim \chi^2(m) \text{ and } v \sim \chi^2(n) \text{ are independent chi-square variates, then the variate } F = (u/m)/(v/n) \text{ has an } F \text{ distribution of } m \text{ and } n \text{ degrees of freedom; and this is denoted by writing } F \sim F(m, n).$$

The sampling distribution which is most frequently used is the t distribution. A t variate is a ratio of a standard normal variate and the root of an independent chi-square variate divided by its degrees of freedom. Thus,

$$(5.5) \quad \text{If } z \sim N(0, 1) \text{ and } v \sim \chi^2(n) \text{ are independent variates, then } t = z/\sqrt{v/n} \text{ has a } t \text{ distribution of } n \text{ degrees of freedom; and this is denoted by writing } t \sim t(n).$$

It is clear that $t^2 \sim F(1, n)$.

Hypothesis Concerning the Coefficients

A linear function of a normally distributed vector is itself normally distributed. Thus, it follows that, if $y \sim N(X\beta, \sigma^2 I)$, then

$$(5.6) \quad \hat{\beta} \sim N_k\{\beta, \sigma^2(X'X)^{-1}\}.$$

Likewise, the marginal distributions of $\hat{\beta}_1, \hat{\beta}_2$ within $\hat{\beta}' = [\hat{\beta}_1, \hat{\beta}_2]$ are given by

$$(5.7) \quad \hat{\beta}_1 \sim N_{k_1}(\beta_1, \sigma^2\{X'_1(I - P_2)X_1\}^{-1}),$$

$$(5.8) \quad \hat{\beta}_2 \sim N_{k_2}(\beta_2, \sigma^2\{X'_2(I - P_1)X_2\}^{-1}),$$

where $P_1 = X_1(X'_1X_1)^{-1}X'_1$ and $P_2 = X_2(X'_2X_2)^{-1}X'_2$. On applying the result under (5.2) to (5.6), we find that

$$(5.9) \quad \sigma^{-2}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \sim \chi^2(k).$$

Similarly, it follows from (5.7) and (5.8) that

$$(5.10) \quad \sigma^{-2}(\hat{\beta}_1 - \beta_1)'X'_1(I - P_2)X_1(\hat{\beta}_1 - \beta_1) \sim \chi^2(k_1),$$

$$(5.11) \quad \sigma^{-2}(\hat{\beta}_2 - \beta_2)'X'_2(I - P_1)X_2(\hat{\beta}_2 - \beta_2) \sim \chi^2(k_2).$$

The distribution of the residual vector $e = y - X\hat{\beta}$ is degenerate in the sense that the mapping $e = \{I - X(X'X)^{-1}X'\}y = \{I - P\}\varepsilon$, where $P = X(X'X)^{-1}X'$, which is from the disturbance vector ε to the residual vector e , entails a singular transformation. Nevertheless, it is possible to obtain a factorisation of the transformation in the form of $I - P = CC'$, where C is matrix of order $T \times (T - k)$ comprising $T - k$ orthonormal columns which are orthogonal to the columns of X such that $C'X = 0$. Now, $C'C = I_{T-k}$; so it follows that, on premultiplying $y \sim N_T(X\beta, \sigma^2 I)$ by C' , we get $C'y \sim N_{T-k}(0, \sigma^2 I)$. Hence

$$(5.12) \quad \sigma^{-2}y'CC'y = \sigma^{-2}(y - X\hat{\beta})'(y - X\hat{\beta}) \sim \chi^2(T - k).$$

The vectors $X\hat{\beta} = Py$ and $y - X\hat{\beta} = (I - P)y$ have a zero-valued covariance matrix. That is

$$(5.13) \quad C(e, X\hat{\beta}) = (I - P)D(y)P' = \sigma^2(I - P)P' = 0,$$

since $D(y) = \sigma^2 I$ and $(I - P)P' = (I - P)P = 0$. If two normally distributed random vectors have a zero covariance matrix, then they are statistically independent. Therefore, it follows that

$$(5.14) \quad \begin{aligned} \sigma^{-2}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) &\sim \chi^2(k) \quad \text{and} \\ \sigma^{-2}(y - X\hat{\beta})'(y - X\hat{\beta}) &\sim \chi^2(T - k) \end{aligned}$$

are mutually independent chi-square variates. From this, it can be deduced that

$$(5.15) \quad \begin{aligned} F &= \left\{ \frac{(\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta)}{k} \middle/ \frac{(y - X\hat{\beta})' (y - X\hat{\beta})}{T - k} \right\} \\ &= \frac{1}{\hat{\sigma}^2 k} (\hat{\beta} - \beta)' X' X (\hat{\beta} - \beta) \sim F(k, T - k). \end{aligned}$$

To test an hypothesis specifying that $\beta = \beta_{\diamond}$, we simply insert this value in the above statistic and compare the resulting value with the critical values of an F distribution of k and $T - k$ degrees of freedom. If a critical value is exceeded, then the hypothesis is liable to be rejected.

The test is readily intelligible, since it is based on a measure of the distance between the hypothesised value $X\beta_{\diamond}$ of the systematic component of the regression and the value $X\hat{\beta}$ which is suggested by the data. If the two values are remote from each other, then we may suspect that the hypothesis is at fault.

It is usual to suppose that a subset of the elements of the parameter vector β are zeros. This represents an instance of a class of hypotheses which specify values for a subvector β_2 within the partitioned model $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$ without asserting anything about the values of the remaining elements in the subvector β_1 . The appropriate test statistic for testing the hypothesis that $\beta_2 = \beta_{2\diamond}$ is

$$(5.16) \quad F = \frac{1}{\hat{\sigma}^2 k_2} (\hat{\beta}_2 - \beta_{2\diamond})' X_2' (I - P_1) X_2 (\hat{\beta}_2 - \beta_{2\diamond}).$$

This will have an $F(k_2, T - k)$ distribution if the hypothesis is true.

We are unlikely to propose that $\beta = 0$ as a whole. Even if we suppose that none of the explanatory variables in a regression model are relevant in explaining the values of the dependent variable, we are likely, nevertheless, to suppose that they have a nonzero mean, which is to say that intercept term is supposed to be nonzero. To test the hypothesis that $\beta_z = 0$ in the model $(y; \iota\alpha + Z\beta_z, \sigma^2 I)$, we could use a statistic in the form of (5.16) with $\beta_2 = \beta_z$ and $X_2 = Z$ and where $P_1 = P_z = T^{-1}u'u$ is the averaging operator.

Of course, the intercept term would be eliminated by taking the variables in deviation form. The hypothesis that $\beta_z = 0$ in the deviations model, which proposes that all of the model's regression coefficients are zero, is the same as the hypothesis that α alone is nonzero in the original model; and the relevant test statistics are identical.

A limiting case of the F statistic concerns the test of an hypothesis affecting a single element β_i within the vector β . By specialising the expression under (5.16), a statistic may be derived in the form of

$$(5.17) \quad F = \frac{(\hat{\beta}_i - \beta_{i\diamond})^2}{\hat{\sigma}^2 w_{ii}},$$

wherein w_{ii} stands for the i th diagonal element of $(X'X)^{-1}$. If the hypothesis is true, then this will be distributed according to the $F(1, T - k)$ law. However, the usual way of assessing such an hypothesis is to relate the value of the statistic

$$(5.18) \quad t = \frac{\hat{\beta}_i - \beta_{i\diamond}}{\sqrt{(\hat{\sigma}^2 w_{ii})}}$$

to the tables of the $t(T - k)$ distribution. The advantage of the t statistic is that it shows the direction in which the estimate of β_i deviates from the hypothesised value as well as the size of the deviation.

6. THE REGRESSION MODEL WITH FIRST ORDER AUTOREGRESSIVE DISTURBANCES

In the classical linear regression model, it is assumed that the disturbances constitute a sequence $\varepsilon(t) = \{\varepsilon_t; t = 0, \pm 1, \pm 2, \dots\}$ of independently and identically distributed random variables such that

$$(6.1) \quad E(\varepsilon_t \varepsilon_s) = \begin{cases} \sigma^2, & \text{if } t = s; \\ 0, & \text{if } t \neq s. \end{cases}$$

The process which generates such disturbances is often called a white-noise process.

Our task is to find models for the disturbance process which are more in accordance with the circumstances of economics where the variables tend to show a high degree of inertia. In econometrics, the traditional means of representing the inertial properties of the disturbance process has been to adopt a simple first-order autoregressive model, or AR(1) model, whose equation takes the form of

$$(6.2) \quad \eta_t = \phi \eta_{t-1} + \varepsilon_t, \quad \text{where } \phi \in (-1, 1).$$

Here it continues to be assumed that ε_t is generated by a white-noise process with $E(\varepsilon_t) = 0$. In many econometric applications, the value of ϕ falls in the more restricted interval $[0, 1)$.

According to this model, the conditional expectation of η_t given η_{t-1} is $E(\eta_t | \eta_{t-1}) = \phi \eta_{t-1}$. That is to say, the expectation of the current disturbance is ϕ times the value of the previous disturbance. This implies that, for a value of ϕ which is closer to unity than to zero, there will be a high degree of correlation amongst successive elements of the sequence $\eta(t) = \{\eta_t; t = 0, \pm 1, \pm 2, \dots\}$.

We can show that the covariance of two elements of the sequence $\eta(t)$ which are separated by τ time periods is given by

$$(6.3) \quad C(\eta_{t-\tau}, \eta_t) = \gamma_\tau = \sigma^2 \frac{\phi^\tau}{1 - \phi^2}.$$

It follows that variance of the process, which is formally the autocovariance of lag $\tau = 0$, is given by

$$(6.4) \quad V(\eta_t) = \gamma_0 = \frac{\sigma^2}{1 - \phi^2}.$$

As ϕ tends to unity, the variance increases without bound.

To find the correlation of two elements from the autoregressive sequence, we note that

$$(6.5) \quad \text{Corr}(\eta_{t-\tau}, \eta_t) = \frac{C(\eta_{t-\tau}, \eta_t)}{\sqrt{V(\eta_{t-\tau})V(\eta_t)}} = \frac{C(\eta_{t-\tau}, \eta_t)}{V(\eta_t)} = \frac{\gamma_\tau}{\gamma_0}.$$

This implies that the correlation of the two elements separated by τ periods is just ϕ^τ ; and thus, as the temporal separation increases, the correlation tends to zero in the manner of a convergent geometric progression.

To demonstrate these results, let us consider substituting for $\eta_{t-1} = \phi\eta_{t-2} + \varepsilon_{t-1}$ in the equation under (6.2) and then substituting for $\eta_{t-2} = \phi\eta_{t-3} + \varepsilon_{t-2}$, and so on indefinitely. By this process, we find that

$$(6.6) \quad \begin{aligned} \eta_t &= \phi\eta_{t-1} + \varepsilon_t \\ &= \phi^2\eta_{t-2} + \varepsilon_t + \phi\varepsilon_{t-1} \\ &\vdots \\ &= \{\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \dots\} \\ &= \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}. \end{aligned}$$

Here the final expression is justified by the fact that $\phi^n \rightarrow 0$ as $n \rightarrow \infty$ in consequence of the restriction that $|\phi| < 1$. Thus we see that η_t is formed as a geometrically declining weighted average of all past values of the sequence $\varepsilon(t)$.

Using this result, we can now write

$$(6.7) \quad \begin{aligned} \gamma_\tau &= C(\eta_{t-\tau}, \eta_t) = E(\eta_{t-\tau}\eta_t) \\ &= E\left(\left\{\sum_{i=0}^{\infty} \phi^i \varepsilon_{t-\tau-i}\right\} \left\{\sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}\right\}\right) \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \phi^i \phi^j E(\varepsilon_{t-\tau-i} \varepsilon_{t-j}). \end{aligned}$$

But the assumption that $\varepsilon(t)$ is a white-noise process with zero-valued autocovariances at all nonzero lags implies that

$$(6.8) \quad E(\varepsilon_{t-\tau-i} \varepsilon_{t-j}) = \begin{cases} \sigma^2, & \text{if } j = \tau + i; \\ 0, & \text{if } j \neq \tau + i. \end{cases}$$

Therefore, on using the above conditions in (6.7) and on setting $j = \tau + i$, we find that

$$(6.9) \quad \begin{aligned} \gamma_\tau &= \sigma^2 \sum_i^{\infty} \phi^i \phi^{i+\tau} = \sigma^2 \phi^\tau \sum_i^{\infty} \phi^{2i} \\ &= \sigma^2 \phi^\tau \{1 + \phi^2 + \phi^4 + \phi^6 + \dots\} \\ &= \sigma^2 \frac{\phi^\tau}{1 - \phi^2}. \end{aligned}$$

This establishes the result under (6.3).

Now let us imagine a linear regression model in the form of

$$(6.10) \quad y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \cdots + x_{tk}\beta_k + \eta_t,$$

where η_t follows a first-order autoregressive process. A set of T instances of the relationship would be written as $y = X\beta + \eta$, where y and η are vectors of T elements and X is a matrix of order $T \times k$. The variance-covariance or dispersion matrix of the vector $\eta = [\eta_1, \eta_2, \eta_3, \dots, \eta_T]'$ takes the form of $[\gamma_{|i-j|}] = \sigma_\varepsilon^2 Q$, where

$$(6.11) \quad Q = \frac{1}{1-\phi^2} \begin{bmatrix} 1 & \phi & \phi^2 & \dots & \phi^{T-1} \\ \phi & 1 & \phi & \dots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \dots & \phi^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \dots & 1 \end{bmatrix};$$

and it can be confirmed directly that

$$(6.12) \quad Q^{-1} = \begin{bmatrix} 1 & -\phi & 0 & \dots & 0 & 0 \\ -\phi & 1 + \phi^2 & -\phi & \dots & 0 & 0 \\ 0 & -\phi & 1 + \phi^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1 + \phi^2 & -\phi \\ 0 & 0 & 0 & \dots & -\phi & 1 \end{bmatrix}.$$

This is a matrix of three nonzero diagonal bands. The elements of principal diagonal, apart from the first and the last, have the value of $1 + \phi^2$. The first and last elements are units. The elements of the supradiagonal band and of the subdiagonal band have the value of $-\phi$.

Given its sparsity, the matrix Q^{-1} could be used directly in implementing the generalised least-squares estimator for which the formula is

$$(6.13) \quad \beta^* = (X'Q^{-1}X)^{-1}X'Q^{-1}y.$$

However, by exploiting the factorisation $Q^{-1} = T'T$, we are able to implement the estimator by applying an ordinary least-squares procedure to the transformed data $W = TX$ and $g = Ty$. The following equation demonstrates the equivalence of the procedures:

$$(14) \quad \begin{aligned} \beta^* &= (W'W)^{-1}W'g \\ &= (X'T'TX)^{-1}X'T'Ty \\ &= (X'Q^{-1}X)^{-1}X'Q^{-1}y \end{aligned}$$

The factor T of the matrix $Q^{-1} = T'T$ takes the form of

$$(6.15) \quad T = \begin{bmatrix} \sqrt{1-\phi^2} & 0 & 0 & \dots & 0 \\ -\phi & 1 & 0 & \dots & 0 \\ 0 & -\phi & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

This effects a very simple transformation the data. Thus, for example, the element y_1 within the vector $y = [y_1, y_2, y_3, \dots, y_T]'$ is replaced $y_1\sqrt{1-\phi^2}$ whilst y_t is replaced by $y_t - \phi y_{t-1}$, for all $t > 1$.

Consider, for example, the simple regression model

$$(6.16) \quad y_t = x_t\beta + \eta_t \quad \text{with} \quad \eta_t = \phi\eta_{t-1} + \varepsilon_t.$$

For $t > 1$, the transformation gives the equation

$$(6.17) \quad y_t - \phi y_{t-1} = (x_t - \phi x_{t-1})\beta + \varepsilon_t,$$

which represents a model that fulfils the classical assumptions and for which ordinary least squares regression is the appropriate method of estimation.

7. THE ALMON LAG

A useful way of reducing the number of parameters to be estimated in the distributed-lag model

$$(7.1) \quad y(t) = \sum_{i=0}^k \beta_j x(t-i) + \varepsilon(t)$$

is to assume that the $k+1$ coefficients of $\beta = [\beta_0, \dots, \beta_k]'$, can be represented by the ordinates of a polynomial $P(i)$ of a degree q which is less than k . Thus it may be specified that

$$(7.2) \quad \beta_i = P(i) = \sum_{j=0}^q \gamma_j i^j \quad \text{for} \quad i = 0, \dots, k.$$

Let $\Pi = [i^j]$ and let $\gamma = [\gamma_0, \gamma_1, \dots, \gamma_q]$. Then the equations of equation (7.2) can be be compiled as $\beta = \Pi\gamma$, which can be written more explicitly as

$$(7.3) \quad \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 2 & 4 & \dots & 2^q \\ 1 & 3 & 9 & \dots & 3^q \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & k & k^2 & \dots & k^q \end{bmatrix} \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_q \end{bmatrix}.$$

On substituting (7.2) into (7.1), we get

$$\begin{aligned}
 (7.4) \quad y(t) &= \sum_{i=0}^k \sum_{j=0}^q \gamma_j i^j x(t-i) + \varepsilon(t) \\
 &= \sum_{j=0}^q \gamma_j z_j(t) + \varepsilon(t),
 \end{aligned}$$

where $z_j(t) = \sum_{i=0}^k i^j x(t-i)$. Given a set of T observations $(x_1, y_1), \dots, (x_T, y_T)$, we can form the variables z_{0t}, \dots, z_{qt} for $t = 1, \dots, T$. Then, by using the method of ordinary least squares to regress y_t on these variables, we can find estimates of the polynomial parameters $\gamma_0, \dots, \gamma_q$. By substituting these estimates into the equation (7.2), we can find estimates of the lag coefficients β_0, \dots, β_k .

If the equation of (7.1) were written in matrix form as $y = X\beta + \varepsilon$, then setting $\beta = \Pi\gamma$ would give

$$(7.5) \quad y = X\Pi\gamma + \varepsilon = Z\gamma + \varepsilon.$$

From the estimate $\hat{\gamma} = (Z'Z)^{-1}Z'y$ we should find $\hat{\beta} = \Pi\hat{\gamma}$.

In her original exposition, Almon employed an alternative form of the q th-degree polynomial $P(i)$. This takes account of the fact that, if τ_0, \dots, τ_q are any $q+1$ points in the domain of $P(i)$, then the polynomial is completely specified by knowing the values $p_0 = P(\tau_0), \dots, p_q = P(\tau_q)$. Taking these values as our parameters, we can write the alternative representation as

$$(7.6) \quad P(\tau) = \sum_{j=0}^q p_j \delta_j(\tau),$$

wherein

$$(7.7) \quad \delta_j(\tau) = \frac{\prod_{\ell \neq j}^q (\tau - \tau_\ell)}{\prod_{l \neq j}^q (\tau_j - \tau_\ell)}$$

is a polynomial of degree q with the property that $\delta_j(\tau_j) = 1$ and $\delta_j(\tau_\ell) = 0$ for all $\ell \neq j$. The functions $\delta_j(\tau); j = 0, \dots, q$ are called Lagrangean interpolation polynomials; and together they form a basis for the set of all polynomials of degrees less than or equal to q . A complete array of the $q+1$ basis function can be represented as follows:

$$\begin{aligned}
 (7.8) \quad \delta_0(\tau) &= \frac{(\tau - \tau_1)(\tau - \tau_2) \cdots (\tau - \tau_q)}{(\tau_0 - \tau_1)(\tau_0 - \tau_2) \cdots (\tau_0 - \tau_q)}, \\
 \delta_1(\tau) &= \frac{(\tau - \tau_0)(\tau - \tau_2) \cdots (\tau - \tau_q)}{(\tau_1 - \tau_0)(\tau_1 - \tau_2) \cdots (\tau_1 - \tau_q)}, \\
 &\vdots \\
 \delta_q(\tau) &= \frac{(\tau - \tau_0)(\tau - \tau_1) \cdots (\tau - \tau_{q-1})}{(\tau_q - \tau_0)(\tau_q - \tau_1) \cdots (\tau_q - \tau_{q-1})}.
 \end{aligned}$$

If, instead of (7.2), we substitute into (7.1) the expression

$$(7.9) \quad \beta_i = P(i) = \sum_{j=0}^q p_j \delta_j(i),$$

then we obtain the equation

$$(7.10) \quad \begin{aligned} y(t) &= \sum_{i=0}^k \sum_{j=0}^q p_j \delta_j(i) x(t-i) + \varepsilon(t) \\ &= \sum_{j=0}^q p_j z_j(t) + \varepsilon(t), \end{aligned}$$

where $z_j(t) = \sum_{i=0}^q \delta_j(i) x(t-i)$. Given a set of T observations, we can estimate the lag coefficients in the same way as before.

First, we choose values τ_0, \dots, τ_q . These should be a sequence of rising values in the vicinity of the interval $[0, k]$, which is the range of the index of the coefficients $\{\beta_i\}$. Then, having formed the associated Lagrangean interpolation polynomials, we can find the T values of the transformed variables z_{0t}, \dots, z_{qt} . Next, the estimates of the polynomial parameters p_0, \dots, p_q are found by using the method of ordinary least squares to regress y_t on the transformed variables. Finally, the estimates of the lag coefficients β_0, \dots, β_q can be obtained from equation (7.9) by putting the estimates of the polynomial parameters in place of the unknown values.

The Lagrangean polynomials have the advantage that they enable us to tie down the sequence of coefficients β_0, \dots, β_k at either end by ensuring that $\beta_{-1} = P(-1) = 0$ and $\beta_{k+1} = P(k+1) = 0$. This is achieved by setting $\tau_0 = -1$ and $\tau_q = k+1$ and by specifying that $P(\tau_0) = 0$ and $P(\tau_q) = 0$. With these values in place, only $q-1$ ordinates p_1, \dots, p_{q-1} need to be estimated since equation (7.8) reduces to

$$(7.11) \quad \beta_i = P(i) = \sum_{j=2}^q p_j \delta_j(\tau_j) \quad \text{where } i = 0, \dots, k.$$

The polynomial $P(i)$ can be expressed as a linear combination of other sets of basis polynomials. Although the choice of the Lagrangean interpolation polynomials as a basis renders the problem of incorporating certain types of *a priori* information particularly tractable, it may be less useful for incorporating other types of information.

8. TRANSFER FUNCTIONS

Consider a simple dynamic model of the form

$$(8.1) \quad y(t) = \phi y(t-1) + x(t)\beta + \varepsilon(t).$$

With the use of the lag operator, we can rewrite this as

$$(8.2) \quad (1 - \phi L)y(t) = \beta x(t) + \varepsilon(t)$$

or, equivalently, as

$$(8.3) \quad y(t) = \frac{\beta}{1 - \phi L} x(t) + \frac{1}{1 - \phi L} \varepsilon(t).$$

The latter is the so-called rational transfer-function form of the equation. We can replace the operator L within the transfer functions or filters associated with the signal sequence $x(t)$ and disturbance sequence $\varepsilon(t)$ by a complex number z . Then, for the transfer function associated with the signal, we get

$$(8.4) \quad \frac{\beta}{1 - \phi z} = \beta \{1 + \phi z + \phi^2 z^2 + \dots\},$$

where the RHS comes from a familiar power-series expansion.

The sequence $\{\beta, \beta\phi, \beta\phi^2, \dots\}$ of the coefficients of the expansion constitutes the impulse response of the transfer function. That is to say, if we imagine that, on the input side, the signal is a unit-impulse sequence of the form

$$(8.5) \quad x(t) = \{\dots, 0, 1, 0, 0, \dots\},$$

which has zero values at all but one instant, then its mapping through the transfer function would result in an output sequence of

$$(8.6) \quad r(t) = \{\dots, 0, \beta, \beta\phi, \beta\phi^2, \dots\}.$$

Another important concept is the step response of the filter. We may imagine that the input sequence is zero-valued up to a point in time when it assumes a constant unit value:

$$(8.7) \quad x(t) = \{\dots, 0, 1, 1, 1, \dots\}.$$

The mapping of this sequence through the transfer function would result in an output sequence of

$$(8.8) \quad s(t) = \{\dots, 0, \beta, \beta + \beta\phi, \beta + \beta\phi + \beta\phi^2, \dots\}$$

whose elements, from the point when the step occurs in $x(t)$, are simply the partial sums of the impulse-response sequence.

This sequence of partial sums $\{\beta, \beta + \beta\phi, \beta + \beta\phi + \beta\phi^2, \dots\}$ is described as the step response. Given that $|\phi| < 1$, the step response converges to a value

$$(8.9) \quad \gamma = \frac{\beta}{1 - \phi}$$

which is described as the steady-state gain or the long-term multiplier of the transfer function.

These various concepts apply to models of any order. Consider the equation

$$(8.10) \quad \alpha(L)y(t) = \beta(L)x(t) + \varepsilon(t),$$

where

$$(8.11) \quad \begin{aligned} \alpha(L) &= 1 + \alpha_1 L + \cdots + \alpha_p L^p \\ &= 1 - \phi_1 L - \cdots - \phi_p L^p, \\ \beta(L) &= 1 + \beta_1 L + \cdots + \beta_k L^k \end{aligned}$$

are polynomials of the lag operator. The transfer-function form of the model is simply

$$(8.12) \quad y(t) = \frac{\beta(L)}{\alpha(L)}x(t) + \frac{1}{\alpha(L)}\varepsilon(t),$$

The rational function associated with $x(t)$ has a series expansion

$$(8.13) \quad \begin{aligned} \frac{\beta(z)}{\alpha(z)} &= \omega(z) \\ &= \{\omega_0 + \omega_1 z + \omega_2 z^2 + \cdots\}; \end{aligned}$$

and the sequence of the coefficients of this expansion constitutes the impulse-response function. The partial sums of the coefficients constitute the step-response function. The gain of the transfer function is defined by

$$(8.14) \quad \gamma = \frac{\beta(1)}{\alpha(1)} = \frac{\beta_0 + \beta_1 + \cdots + \beta_k}{1 + \alpha_1 + \cdots + \alpha_p}.$$

The method of finding the coefficients of the series expansion of the transfer function in the general case can be illustrated by the second-order case:

$$(8.15) \quad \frac{\beta_0 + \beta_1 z}{1 - \phi_1 z - \phi_2 z^2} = \{\omega_0 + \omega_1 z + \omega_2 z^2 + \cdots\}.$$

We rewrite this equation as

$$(8.16) \quad \beta_0 + \beta_1 z = \{1 - \phi_1 z - \phi_2 z^2\} \{\omega_0 + \omega_1 z + \omega_2 z^2 + \cdots\}.$$

Then, by performing the multiplication on the RHS, and by equating the coefficients of the same powers of z on the two sides of the equation, we find that

$$(8.17) \quad \begin{aligned} \beta_0 &= \omega_0, & \omega_0 &= \beta_0, \\ \beta_1 &= \omega_1 - \phi_1 \omega_0, & \omega_1 &= \beta_1 + \phi_1 \omega_0, \\ 0 &= \omega_2 - \phi_1 \omega_1 - \phi_2 \omega_0, & \omega_2 &= \phi_1 \omega_1 + \phi_2 \omega_0, \\ \vdots & & \vdots & \\ 0 &= \omega_n - \phi_1 \omega_{n-1} - \phi_2 \omega_{n-2}, & \omega_n &= \phi_1 \omega_{n-1} + \phi_2 \omega_{n-2}. \end{aligned}$$

By examining this scheme, we are able to distinguish between the different roles which are played by the numerator parameters β_0, β_1 and the denominator parameters ϕ_1, ϕ_2 . The parameters of the numerator serve as initial conditions for the process which generates the impulse response. The denominator parameters determine the dynamic nature of the impulse response.

Consider the case where the impulse response takes the form a damped sinusoid. This case arises when the roots of the equation $\alpha(z) = 1 - \phi z - \phi_2 z^2 = 0$ are a pair of conjugate complex numbers falling outside the unit circle—as they are bound to do if the response is to be a damped one. Then the parameters β_0 and β_1 are jointly responsible for the initial amplitude and for the phase of the sinusoid. The phase is the time lag which displaces the peak of the sinusoid so that it occurs after the starting time $t = 0$ of the response, which is where the peak of an undisplaced cosine response would occur.

The parameters ϕ_1 and ϕ_2 , on the other hand, serve to determine the period of the sinusoidal fluctuations and the degree of damping, which is the rate at which the impulse response converges to zero.

It seems that all four parameters ought to be present in a model which aims at capturing any of the dynamic responses of which a second-order system is capable. To omit one of the numerator parameters of the model would be a mistake unless, for example, there is good reason to assume that the impulse response attains its maximum value at the starting time $t = 0$. We are rarely in the position to make such an assumption.