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

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

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

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

The properties of the projector P which have been used in simplifying equation (1.1), indicate that

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

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.

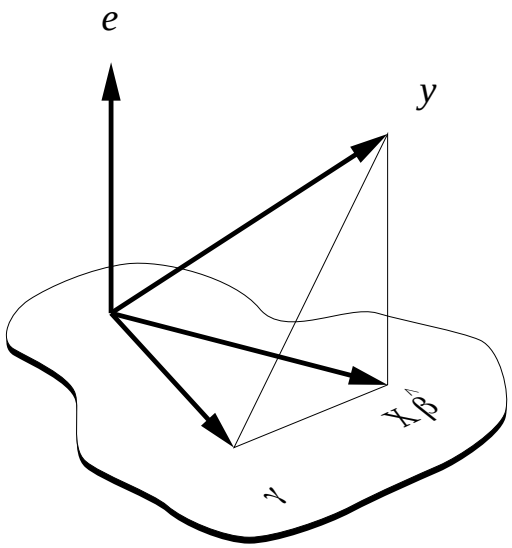

Figure 1. The vector $Py = X\hat{\beta}$ is formed by the orthogonal projection of the vector y onto the subspace spanned by the columns of the matrix X.

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

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

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

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

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

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

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

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

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

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

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

2. THE PARTITIONED REGRESSSION MODEL

Consider taking a regression equation in the form of

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

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

$$X_1'X_1\beta_1 + X_1'X_2\beta_2 = X_1'y, \tag{2.2}$$

$$X_2'X_1\beta_1 + X_2'X_2\beta_2 = X_2'y. \tag{2.3}$$

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

$$X_2'X_1\beta_1 + X_2'X_1(X_1'X_1)^{-1}X_1'X_2\beta_2 = X_2'X_1(X_1'X_1)^{-1}X_1'y. \tag{2.4}$$

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

$$\Big\{X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\Big\}\beta_2 = X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y. \tag{2.5}$$

On defining

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

can we rewrite (2.5) as

$$\Big\{X_2'(I - P_1)X_2\Big\}\beta_2 = X_2'(I - P_1)y, \tag{2.7}$$

whence

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

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

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

Some Algebraic Indentities

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

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

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

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

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

$$P = P_{1/2} + P_{2/1}. \tag{2.13}$$

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 also holds for all values of y, we have

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

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

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

This identity may be written more explicitly as

$$X(X'X)^{-1}X' - X_1(X_1'X_1)^{-1}X_1' = (I - P_1)X_2\Big\{X_2'(I - P_1)X_2\Big\}^{-1}X_2'(I - P_1). \tag{2.16}$$

Regression with an Intercept

We may use the expressions for $\hat{\beta}_1$, $\hat{\beta}_2$ of (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

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

Applying this to the vector y, we get

$$P_i y = i(i'y/T) = i\Big(\sum y_t/T\Big) = i\bar{y}. \tag{2.18}$$

Applying it likewise to the matrix Z, we get

$$P_i Z = i\left[\sum x_{t2}, \ldots, \sum x_{tk}\right]/T = i[\bar{x}_2, \ldots, \bar{x}_k] = \bar{Z}. \tag{2.19}$$

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

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

Thus, the coefficients $\beta_2, \ldots, \beta_k$ 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

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

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

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

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 β 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_2 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_1 by subtracting P_1y to give $(I - P_1)y$. We must also find the components that are peculiar to X_1 by subtracting P_1X_2 to give $(I - P_1)X_2$. We can then obtain a measure of the contribution by finding the ordinary coefficient of determination $R^2\{(I-P_1)y, (I-P_1)X_2\}$ of the regression of $(I-P_1)y$ on $(I-P_1)X_2$. 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 $I - P_1$ and the identity of (2.16), we find that

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

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 (2.18) and (2.19) respectively, we find from (2.23) that

$$R^2(y, Z|i) = \frac{(y-\bar{y})'(Z-\bar{Z})\{(Z-\bar{Z})'(Z-\bar{Z})\}^{-1}(Z-\bar{Z})'(y-\bar{y})}{(y-\bar{y})'(y-\bar{y})} = \frac{y'Py - y'P_i y}{y'(I-P_i)y} = \frac{\hat{\beta}X'X\hat{\beta} - \bar{y}'\bar{y}}{y'y - \bar{y}'\bar{y}}. \tag{2.24}$$

The first equality shows that $R^2(y, Z|i)$ is the ordinary coefficient of determination of the regression (2.20) 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 (2.22).

3. DIAGONALISATION OF A SYMMETRIC MATRIX

The Geometry of Quadratic Forms

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

$$z_1^2 + z_2^2 = r^2. \tag{3.1}$$

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

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

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

$$\lambda_1 y_1^2 + \lambda_2 y_2^2 = r^2, \tag{3.3}$$

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

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

We can write equation (3.6) in matrix notation as

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

This implies

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $x_1, \ldots, x_n$ be a set of n linearly independent characteristic vectors corresponding to its roots $\lambda_1, \ldots, \lambda_n$. Then, we can form a set of normalised vectors

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

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 $C = [c_1, \ldots, c_n]$ is an orthonormal matrix such that $C'C = CC' = I$.

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

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 ith 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

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

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

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

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

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

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

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

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

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

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

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

The first of these results allows us to set $I-P=I-C_1C_1'=C_2C_2'$. Now, if $\varepsilon\sim N(0,\sigma^2I_T)$ and if C is an orthonormal matrix such that $C'C=I_T$, then it follows that $C'\varepsilon\sim N(0,\sigma^2I_T)$. 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

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

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

$$\begin{aligned} \frac{\varepsilon'C_1C_1'\varepsilon}{\sigma^2} &= \frac{\varepsilon'P\varepsilon}{\sigma^2}\sim\chi^2(k) \quad \text{and} \\ \frac{\varepsilon'C_2C_2'\varepsilon}{\sigma^2} &= \frac{\varepsilon'(I-P)\varepsilon}{\sigma^2}\sim\chi^2(T-k) \end{aligned} \tag{4.9}$$

are independent chi-square variates. Since $C_1C_1'+C_2C_2'=I_T$, the sum of these two variates is

$$\frac{\varepsilon'C_1C_1'\varepsilon}{\sigma^2}+\frac{\varepsilon'C_2C_2'\varepsilon}{\sigma^2}=\frac{\varepsilon'\varepsilon}{\sigma^2}\sim\chi^2(T); \tag{4.10}$$

and thus the theorem is proved.

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

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

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) If T is a transformation such that $T\Sigma T' = I$ and $T'T = \Sigma^{-1}$, 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) If $x \sim N(\mu, \Sigma)$ is a vector of order n, 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) If $u \sim \chi^2(m)$ and $v \sim \chi^2(n)$ are independent chi-square variates of m and n degrees of freedom respectively, then $(u + v) \sim \chi^2(m + n)$ is a chi-square variate of $m + n$ 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) If $u \sim \chi^2(m)$ and $v \sim \chi^2(n)$ are independent chi-square variates, then the variate $F = (u/m)/(v/n)$ has an F distribution of m and n 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) If $z \sim N(0, 1)$ and $v \sim \chi^2(n)$ are independent variates, then $t = z/\sqrt{(}v/n)$ has a t distribution of n 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

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

On applying the result under (5.2) to (5.6), we find that

$$\sigma^{-2}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \sim \chi^2(k). \tag{5.7}$$

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 $I - P = C_2C_2'$, where C_2 is matrix of order $T \times (T - k)$ comprising $T - k$ orthonormal columns which are orthogonal to the columns of X such that $C_2'X = 0$. Now, $C_2'C_2 = I_{T-k}$; so it follows that, on premultiplying $y \sim N_T(X\beta, \sigma^2 I)$ by C_2', we get $C_2'y \sim N_{T-k}(0, \sigma^2 I)$. Hence

$$\sigma^{-2}y'C_2C_2'y = \sigma^{-2}(y - X\hat{\beta})'(y - X\hat{\beta}) \sim \chi^2(T - k). \tag{5.8}$$

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

$$C(e, X\hat{\beta}) = (I - P)D(y)P' = \sigma^2(I - P)P' = 0, \tag{5.9}$$

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

$$\begin{aligned} &\sigma^{-2}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \sim \chi^2(k) \quad \text{and} \\ &\sigma^{-2}(y - X\hat{\beta})'(y - X\hat{\beta}) \sim \chi^2(T - k) \end{aligned} \tag{5.10}$$

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

$$\begin{aligned} F &= \left\{ \frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{k} \Big/ \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T - k} \right\} \\ &= \frac{1}{\hat{\sigma}^2 k}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \sim F(k, T - k). \end{aligned} \tag{5.11}$$

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.

Hypothesis Concerning Subsets of the Coefficients

Consider a set of linear restrictions on the vector β of a classical linear regression model $N(y; X\beta, \sigma^2 I)$ which take the form of

$$R\beta = r, \tag{5.11}$$

where R is a matrix of order $j \times k$ and of rank j, which is to say that the j restrictions are independent of each other and are fewer in number than the parameters within β. We know that the ordinary least-squares estimator of β is a normally distributed vector $\hat{\beta} \sim N\{\beta, \sigma^2(X'X)^{-1}\}$. It follow that

$$R\hat{\beta} \sim N\big\{R\beta = r, \sigma^2 R(X'X)^{-1}R'\big\}; \tag{5.12}$$

and, from this, we can immediately infer that

$$\frac{(R\hat{\beta} - r)'\big\{R(X'X)^{-1}R'\big\}^{-1}(R\hat{\beta} - r)}{\sigma^2} \sim \chi^2(j). \tag{5.13}$$

We have already established the result that

$$\frac{(T-k)\hat{\sigma}^2}{\sigma^2} = \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{\sigma^2} \sim \chi^2(T-k) \tag{5.14}$$

is a chi-square variate which is statistically independent of the chi-square variate

$$\frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{\sigma^2} \sim \chi^2(k) \tag{5.55}$$

derived from the estimator of the regression parameters. The variate of (5.14) must also be independent of the chi-square of (5.13); and it is straightforward to deduce that

$$\begin{aligned} F &= \left\{ \frac{(R\hat{\beta} - r)'\big\{R(X'X)^{-1}R'\big\}^{-1}(R\hat{\beta} - r)}{j} \Bigg/ \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T-k} \right\} \\ &= \frac{(R\hat{\beta} - r)'\big\{R(X'X)^{-1}R'\big\}^{-1}(R\hat{\beta} - r)}{\hat{\sigma}^2 j} \sim F(j, T-k), \end{aligned} \tag{5.16}$$

which is to say that the ratio of the two independent chi-square variates is an F statistic. This statistic, which embodies only know and observable quantities, can be used in testing the validity of the hypothesised restrictions $R\beta = r$.

A specialisation of the statistic under (5.16) can also be used in testing an hypothesis concerning a subset of the elements of the vector β. Let $\beta' = [\beta_1', \beta_2']'$. Then the condition that the subvector β_1 assumes the value of β_1^* can be expressed via the equation

$$[I_{k_1}, 0]\begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix} = \beta_1^*. \tag{5.17}$$

This can be construed as a case of the equation $R\beta = r$ where $R = [I_{k_1}, 0]$ and $r = \beta_1^*$.

In order to discover the specialised form of the requisite test statistic, let us consider the following partitioned form of an inverse matrix:

$$\begin{aligned}(X'X)^{-1} &= \begin{bmatrix}X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2\end{bmatrix}^{-1} \\ &= \begin{bmatrix}\{X_1'(I-P_2)X_1\}^{-1} & -\{X_1'(I-P_2)X_1\}^{-1}X_1'X_2(X_2'X_2)^{-1} \\ -\{X_2'(I-P_1)X_2\}^{-1}X_2'X_1(X_1'X_1)^{-1} & \{X_2'(I-P_1)X_2\}^{-1}\end{bmatrix},\end{aligned} \tag{5.18}$$

Then, with $R = [I, 0]$, we find that

$$R(X'X)^{-1}R' = \big\{X_1'(I-P_2)X_1\big\}^{-1} \tag{5.19}$$

It follows in a straightforward manner that the specialised form of the F statistic of (5.16) is

$$\begin{aligned}F &= \left\{\frac{(\hat\beta_1 - \beta_1^*)'\big\{X_1'(I-P_2)X_1\big\}(\hat\beta_1 - \beta_1^*)}{k_1} \Bigg/ \frac{(y - X\hat\beta)'(y - X\hat\beta)}{T-k}\right\} \\ &= \frac{(\hat\beta_1 - \beta_1^*)'\big\{X_1'(I-P_2)X_1'\big\}(\hat\beta_1 - \beta_1^*)}{\hat\sigma^2 k_1} \sim F(k_1, T-k).\end{aligned} \tag{5.20}$$

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

$$F = \frac{(\hat\beta_i - \beta_{i\diamond})^2}{\hat\sigma^2 w_{ii}}, \tag{5.21}$$

wherein w_{ii} stands for the ith 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

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

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. RESTRICTED LEAST-SQUARES REGRESSION

Sometimes, we find that there is a set of *a priori* restrictions on the elements of the vector β of the regression coefficients which can be taken into account in the process of estimation. A set of j linear restrictions on the vector β can be written as $R\beta = r$, where r is a $j \times k$ matrix of linearly independent rows, such that $\text{Rank}(R) = j$, and r is a vector of j elements.

To combine this *a priori* information with the sample information, we adopt the criterion of minimising the sum of squares $(y - X\beta)'(y - X\beta)$ subject to the condition that $R\beta = r$. This leads to the Lagrangean function

$$\begin{aligned} L &= (y - X\beta)'(y - X\beta) + 2\lambda'(R\beta - r) \\ &= y'y - 2y'X\beta + \beta'X'X\beta + 2\lambda'R\beta - 2\lambda'r. \end{aligned} \tag{6.1}$$

On differentiating L with respect to β and setting the result to zero, we get the following first-order condition $\partial L/\partial\beta = 0$:

$$2\beta'X'X - 2y'X + 2\lambda'R = 0, \tag{6.2}$$

whence, after transposing the expression, eliminating the factor 2 and rearranging, we have

$$X'X\beta + R'\lambda = X'y. \tag{6.3}$$

When these equations are compounded with the equations of the restrictions, which are supplied by the condition $\partial L/\partial\lambda = 0$, we get the following system:

$$\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} = \begin{bmatrix} X'y \\ r \end{bmatrix}. \tag{6.4}$$

For the system to have a unique solution, that is to say, for the existence of an estimate of β, it is not necessary that the matrix $X'X$ should be invertible—it is enough that the condition

$$\text{Rank} \begin{bmatrix} X \\ R \end{bmatrix} = k \tag{6.5}$$

should hold, which means that the matrix should have full column rank. The nature of this condition can be understood by considering the possibility of estimating β by applying ordinary least-squares regression to the equation

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix}, \tag{6.6}$$

which puts the equations of the observations and the equations of the restrictions on an equal footing. It is clear that an estimator exits on the condition that $(X'X + R'R)^{-1}$ exists, for which the satisfaction of the rank condition is necessary and sufficient.

Let us simplify matters by assuming that $(X'X)^{-1}$ *does* exist. Then equation (6.3) gives an expression for β in the form of

$$\begin{aligned}\beta^* &= (X'X)^{-1}X'y - (X'X)^{-1}R'\lambda \\ &= \hat{\beta} - (X'X)^{-1}R'\lambda,\end{aligned} \tag{6.7}$$

where $\hat{\beta}$ is the unrestricted ordinary least-squares estimator. Since $R\beta^* = r$, premultiplying the equation by R gives

$$r = R\hat{\beta} - R(X'X)^{-1}R'\lambda, \tag{6.8}$$

from which

$$\lambda = \{R(X'X)^{-1}R'\}^{-1}(R\hat{\beta} - r). \tag{6.9}$$

On substituting this expression back into equation (6.7), we get

$$\beta^* = \hat{\beta} - (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}(R\hat{\beta} - r). \tag{6.10}$$

This formula is more intelligible than it might appear to be at first, for it is simply an instance of the prediction-error algorithm whereby the estimate of β is updated in the light of the information provided by the restrictions. The error, in this instance, is the divergence between $R\hat{\beta}$ and $E(R\hat{\beta}) = r$. Also included in the formula are the terms $D(R\hat{\beta}) = \sigma^2 R(X'X)^{-1}R'$ and $C(\hat{\beta}, R\hat{\beta}) = \sigma^2(X'X)^{-1}R'$.

The sampling properties of the restricted least-squares estimator are easily established. Given that $E(\hat{\beta} - \beta) = 0$, which is to say that $\hat{\beta}$ is an unbiased estimator, it follows that $E(\beta^* - \beta) = 0$, so that β^* is also unbiased.

Next consider the expression

$$\begin{aligned}\beta^* - \beta &= [I - (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}R](\hat{\beta} - \beta) \\ &= (I - P_R)(\hat{\beta} - \beta),\end{aligned} \tag{6.11}$$

where

$$P_R = (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}R. \tag{6.12}$$

The expression comes from taking β from both sides of (6.10) and from recognising that $R\hat{\beta} - r = R(\hat{\beta} - \beta)$. We may observe that P_R is an idempotent matrix which is subject to the conditions that

$$P_R = P_R^2, \quad P_R(I - P_R) = 0 \quad \text{and} \quad P_R'X'X(I - P_R) = 0. \tag{6.13}$$

From equation (6.11), we deduce that

$$\begin{aligned} D(\beta^*) &= (I - P_R)E\{(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\}(I - P_R) \\ &= \sigma^2(I - P_R)(X'X)^{-1}(I - P_R) \\ &= \sigma^2[(X'X)^{-1} - (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}R(X'X)^{-1}]. \end{aligned} \tag{6.14}$$

7. THE REGRESSION MODEL WITH FIRST ORDER AUTOREGRESSIVE DISTURBANCES

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

$$E(\varepsilon_t\varepsilon_s) = \begin{cases} \sigma^2, & \text{if } t = s; \\ 0, & \text{if } t \neq s. \end{cases} \tag{7.1}$$

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

$$\eta_t = \phi\eta_{t-1} + \varepsilon_t, \qquad \text{where} \qquad \phi \in (-1, 1). \tag{7.2}$$

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 that to zero, there will be a high degree of correlation amongst successive elements of the sequence $\eta(t) = \{\eta_t; t = 0, \pm 1, \pm 2, \ldots\}$.

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

$$C(\eta_{t-\tau}, \eta_t) = \gamma_\tau = \sigma^2 \frac{\phi^\tau}{1 - \phi^2}. \tag{7.3}$$

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

$$V(\eta_t) = \gamma_0 = \frac{\sigma^2}{1 - \phi^2}. \tag{7.4}$$

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

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

$$\text{Corr}(\eta_{t-\tau}, \eta_t) = \frac{C(\eta_{t-\tau}, \eta_t)}{\sqrt{V(\eta_{t-\tau})V(\eta_t)}} = \frac{C(\eta_{t-\tau}, \eta_t)}{V(\eta_t)} = \frac{\gamma_\tau}{\gamma_0}. \tag{7.5}$$

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

$$\begin{aligned} \eta_t &= \phi\eta_{t-1} + \varepsilon_t \\ &= \phi^2\eta_{t-2} + \varepsilon_t + \phi\varepsilon_{t-1} \\ &\vdots \\ &= \{\varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-1} + \cdots\} \\ &= \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}. \end{aligned} \tag{7.6}$$

Here the final expression is justified by the fact that $\phi^n \to 0$ as $n \to \infty$ 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

$$\begin{aligned} \gamma_\tau &= C(\eta_{t-\tau}, \eta_t) = E(\eta_{t-\tau}\eta_t) \\ &= E\left(\Big\{\sum_{i=0}^{\infty} \phi^i \varepsilon_{t-\tau-i}\Big\}\Big\{\sum_{j=0}^{\infty} \phi^j \varepsilon_{t-j}\Big\}\right) \\ &= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \phi^i\phi^j E(\varepsilon_{t-\tau-i}\varepsilon_{t-j}). \end{aligned} \tag{7.7}$$

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

$$E(\varepsilon_{t-\tau-i}\varepsilon_{t-j}) = \begin{cases} \sigma^2, & \text{if } j = \tau + i; \\ 0, & \text{if } j \neq \tau + i. \end{cases} \tag{7.8}$$

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

$$\begin{aligned} \gamma_\tau &= \sigma^2 \sum_{i}^{\infty} \phi^i\phi^{i+\tau} = \sigma^2\phi^\tau \sum_{i}^{\infty} \phi^{2i} \\ &= \sigma^2\phi^\tau\{1 + \phi^2 + \phi^4 + \phi^6 + \cdots\} \\ &= \sigma^2 \frac{\phi^\tau}{1 - \phi^2}. \end{aligned} \tag{7.9}$$

This establishes the result under (6.3).

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

$$y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \cdots + x_{tk}\beta_k + \eta_t, \tag{7.10}$$

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 or order $T \times k$. The variance–covariance or dispersion matrix of the vector $\eta = [\eta_1, \eta_2, \eta_3, \ldots \eta_T]'$ takes the form of $[\gamma_{|i-j|}] = \sigma_\varepsilon^2 Q$, where

$$Q = \frac{1}{1-\phi^2}\begin{bmatrix} 1 & \phi & \phi^2 & \ldots & \phi^{T-1} \\ \phi & 1 & \phi & \ldots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \ldots & \phi^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \ldots & 1 \end{bmatrix}; \tag{7.11}$$

and it can be confirmed directly that

$$Q^{-1} = \begin{bmatrix} 1 & -\phi & 0 & \ldots & 0 & 0 \\ -\phi & 1+\phi^2 & -\phi & \ldots & 0 & 0 \\ 0 & -\phi & 1+\phi^2 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & \ldots & -\phi & 1 \end{bmatrix}. \tag{7.12}$$

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

$$\beta^* = (X'Q^{-1}X)^{-1}X'Q^{-1}y. \tag{7.13}$$

However, by exploiting the factorisation $Q^{-1} = T'T$, we are able to 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:

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

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

$$T = \begin{bmatrix} \sqrt{1-\phi^2} & 0 & 0 & \dots & 0 \\ -\phi & 1 & 0 & \dots & 0 \\ 0 & -\phi & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}. \tag{7.15}$$

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

$$y_t = x_t\beta + \eta_t \qquad \text{with} \quad \eta_t = \phi\eta_{t-1} + \varepsilon_t. \tag{7.16}$$

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

$$y_t - \phi y_{t-1} = (x_t - \phi x_{t-1})\beta + \varepsilon_t, \tag{7.17}$$

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

8. ORDINARY LEAST-SQUARES REGRESSION AND NONSPHERICAL DISTURBANCES

In cases where the structure of the dispersion matrix of the regression disturbances is known to depend on a small set of parameters, it will be possible to estimate the regression parameter β in the model $(y; X\beta, \sigma^2 Q)$ via a method of feasible generalised least squares. This uses an estimate Ω^* of the dispersion matrix of the disturbances within the formula $\beta^* = (X'\Omega^{*-1}X)^{-1}X'\Omega^{*-1}y$. In other cases, where there is no knowledge of the structure of the dispersion matrix, we may have to use the ordinary least-squares (OLS) estimator $\hat{\beta} = (X'X)^{-1}X'y$.

The OLS estimator will be unbiased and, subject to certain restrictions limiting the serial dependence of the disturbances, it will also be consistent. However, the dispersion matrix of the estimator will differ from that which obtains in the case of the OLS estimator of the classical model $(y; X\beta, \sigma^2 I)$, which is $D(\hat{\beta}) = \sigma^2(X'X)^{-1}$. In fact, the dispersion matrix of the OLS estimator of β in the model $(y; X\beta, \sigma^2 Q)$ is given by

$$\begin{aligned} D(\hat{\beta}) &= (X'X)^{-1}X'D(y)X(X'X)^{-1} \\ &= (X'X)^{-1}\{\sigma^2 X'\Omega X\}(X'X)^{-1}, \end{aligned} \tag{8.1}$$

which is commonly referred to as the sandwich formula. Here, $D(y) = E(\varepsilon\varepsilon') = \sigma^2\Omega = \Sigma$ is a symmetric matrix of order T, which cannot be estimated on the basis of a sample

of size T, unless there are sufficient restrictions on its structure. However, in order to implement the sandwich formula, what is required is an estimate of the matrix

$$W = \sigma^2 X'\Omega X = X'\Sigma X, \tag{8.2}$$

which is of the order k, which is the number of explanatory variables in X.

To reveal the structure of this matrix, let us consider the elements of the matrices $W = [w_{ij}]$, $X = [x_{tj}]$, $X' = [x_{si}]'$ and $\Sigma = [\sigma_{st}]$. Then, there is

$$\begin{aligned} w_{ij} &= \sum_t \sum_s x_{si}\sigma_{st}x_{tj} \\ &= \sum_t \sum_s x_{si}E(\varepsilon_s\varepsilon_t)x_{tj}; \end{aligned} \tag{8.3}$$

and the matrix as a whole is given by $W = \sum_t \sum_s x'_{s\bullet}E(\varepsilon_s\varepsilon_t)x_{t\bullet}$, in a more summary notation. For this to be estimable, some further restrictions are necessary. The restriction that removes the serial dependence from the disturbances, but which allows them to be heteroskedastic, sets

$$E(\varepsilon_s\varepsilon_t) = \begin{cases} \sigma_t^2, & \text{if } t = s; \\ 0, & \text{if } t \neq s. \end{cases} \tag{8.4}$$

Then, there is

$$w_{ij} = \sum_t \sigma_t^2 x_{ti}x_{tj} \tag{8.5}$$

and $W = \sum_t \sigma_t^2 x'_{t\bullet}x_{t\bullet}$. There are still as many parameters within the matrix $\Sigma = \operatorname{diag}\{\sigma_1^2, \ldots, \sigma_T^2\}$ as there are observations. Therefore, it cannot be estimated consistently. Nevertheless, under certain assumptions, the product $T^{-1}W = T^{-1}X'\Sigma X$ can be estimated consistently via

$$\frac{1}{T}\hat{w}_{ij} = \frac{1}{T}\sum_t e_t^2 x_{ti}x_{tj} \tag{8.6}$$

which is obtained by replacing $\sigma^2 = E(\varepsilon_t^2)$ by the squared residual e_t^2. This is the heteroskedasticity-consistent estimator of White (1982).

To demonstrate the consistency, we note that, if $\hat{\beta} \to \beta$ as $T \to \infty$, then $e_t^2 \to \varepsilon_t^2$. Therefore, it is sufficient to consider the limiting behaviour of

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^2 x_{ti}x_{tj} &= \frac{1}{T}\sum_{t=1}^{T}(\sigma_t^2 + \nu_t)x_{ti}x_{tj} \\ &= \frac{1}{T}\sum_{t=1}^{T}\sigma_t^2 x_{ti}x_{tj} + \frac{1}{T}\sum_{t=1}^{T}\nu_t x_{ti}x_{tj}. \end{aligned} \tag{8.7}$$

In the second term on the LHS, there is a random variable ν_t, representing the deviation of ε_t^2 from its expected value $E(\varepsilon_t^2) = \sigma_t^2$, which has $E(\nu_t) = 0$ and which is independent of the elements x_{ti} and x_{tj}. We can expect the second term to converge to zero. Since $e_t^2 \to \varepsilon_t^2$, it follows that $T^{-1}\hat{w}_{ij} = T^{-1}\sum_t e_t^2 x_{ti}x_{tj}$ converges to $T^{-1}w_{ij}$.

The restriction that eliminates the serial dependence of the disturbances is much stronger than it need be. Given that the matrix W is of a constant order k, whereas the sample size T may grow indefinitely, there is hope of estimating W consistently in most practical circumstances. Consider the writing the matrix $T^{-1}W = T^{-1}X'E(\varepsilon\varepsilon')X$ as

$$\begin{aligned}\frac{1}{T}W &= \sum_{s=1}^{T}\sum_{t=1}^{T}\frac{1}{T}x'_{s\bullet}E(\varepsilon_s\varepsilon_t)x_{t\bullet} \\ &= \sum_{j=1-T}^{T-1}\Big\{\frac{1}{T}\sum_{t=j+1}^{T}x'_{t\bullet}E(\varepsilon_t\varepsilon_{t-j})x_{[t-j]\bullet}\Big\}.\end{aligned} \tag{8.8}$$

The elements that are subject to these summations may be assigned to a square matrix of order T of which the rows and columns are indexed by $s, t = 1, \ldots, T$. In the first expression, the summation runs across each of the matrix rows in succession. In the second expression, it runs along the NE–SW diagonals of the matrix, beginning in the bottom left corner and rising through the principal diagonal to the top right corner. The dispersion matrix of a stationary stochastic process has constant values along these diagonals. Therefore, the second expression is appropriate to cases where both the data and the disturbances are generated by stationary processes.

The second equality of (8.8) can be written as

$$\frac{1}{T}W = \sum_{j=1-T}^{T-1}\Gamma_j = \Gamma_0 + \sum_{j=1}^{T-1}(\Gamma_j + \Gamma'_j), \tag{8.9}$$

where Γ_j is the expression within the braces. The empirical counterpart of this matrix is

$$G_j = \frac{1}{T}\sum_{t=j+1}^{T}x'_{t\bullet}e_te_{t-j}x_{[t-j]\bullet}. \tag{8.10}$$

If the number j is small in comparison with T, then we can expect G_j to be an adequate estimate of Γ_j. Moreover, for a fixed j, we can expect $G_j \to \Gamma_j$ as $T \to \infty$.

Replacing Γ_j by G_j in (8.9) for all j results in the matrix $T^{-1}X'ee'X$, which does not constitute a viable estimator. The difficulty lies in the estimates G_j when j is close to T. In that case, the estimate will comprise a limited amount of information from $T - j$ sample points. Various recourses for avoiding the problem are available. The simplest of these is to limit the range of the index j such that its absolute value does not exceed some threshold value p. Then, we obtain the estimator of Hansen (1982), which is

$$W_H = \sum_{j=p}^{p}G_j = G_0 + \sum_{j=1}^{p}(G_j + G'_j). \tag{8.11}$$

An alternative estimator, which is due to Newey and West (1987), applies a gradual discount to the matrices G_j as j increases. It takes the form of

$$W_N = G_0 + \sum_{j=1}^{p} \Big(1 - \frac{j}{p+1}\Big)(G_j + G_j'). \tag{8.12}$$

References

Hansen, L.P., (1982), Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica,* 50, 1029–1054.

Newey, W.K. and K.D. West, (1987), A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix, *Econometrica,* 55, 703–708.

White, H., (1980), A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity, *Econometrica,* 48, 817–838.