1. CONDITIONAL EXPECTATIONS

Minimum-Mean-Square-Error Predictsion

Let y be a continuously distributed random variable whose probability density function is $f(y)$. If we wish to predict the value of y without the help of any other information, then we might take its expected value which is defined by

$$E(y) = \int yf(y)dy.$$

The expected value is a so-called minimum-mean-square-error (m.m.s.e.) predictor. If π is the value of a prediction, then the mean-square error is given by

$$\begin{aligned} M &= \int (y-\pi)^2 f(y)dy \\ &= E\big\{(y-\pi)^2\big\} \\ &= E(y^2) - 2\pi E(y) + \pi^2; \end{aligned} \tag{1.1}$$

and, using the methods of calculus, it is easy to show that this quantity is minimised by taking $\pi = E(y)$.

Now let us imagine that y is statistically related to another random variable x whose value we have already observed. For the sake of argument, let us assume that we know the form of the joint distribution of x and y which is $f(x, y)$. Then the minimum-mean-square-error prediction of y is given by the conditional expectation

$$E(y|x) = \int y\frac{f(x,y)}{f(x)}dy \tag{1.2}$$

wherein

$$f(x) = \int f(x, y)dy \tag{1.3}$$

is the so-called marginal distribution of x. We may state this proposition formally in a way which will assist us in proving it:

(1.4) Let $\hat{y} = \hat{y}(x)$ be the conditional expectation of y given x which is also expressed as $\hat{y} = E(y|x)$. Then we have $E\{(y-\hat{y})^2\} \leq E\{(y-\pi)^2\}$, where $\pi = \pi(x)$ is any other function of x.

Proof. Consider

$$\begin{aligned} E\big\{(y-\pi)^2\big\} &= E\Big[\big\{(y-\hat{y}) + (\hat{y}-\pi)\big\}^2\Big] \\ &= E\big\{(y-\hat{y})^2\big\} + 2E\big\{(y-\hat{y})(\hat{y}-\pi)\big\} + E\big\{(\hat{y}-\pi)^2\big\}. \end{aligned} \tag{1.5}$$

In the second term, there is

$$\begin{aligned} E\big\{(y-\hat{y})(\hat{y}-\pi)\big\} &= \int_x\int_y (y-\hat{y})(\hat{y}-\pi)f(x,y)\partial y\partial x \\ &= \int_x\left\{\int_y (y-\hat{y})f(y|x)\partial y\right\}(\hat{y}-\pi)f(x)\partial x \\ &= 0. \end{aligned} \tag{1.6}$$

Here the second equality depends upon the factorisation $f(x,y) = f(y|x)f(x)$ which expresses the joint probability density function of x and y as the product of the conditional density function of y given x and the marginal density function of x. The final equality depends upon the fact that $\int(y-\hat{y})f(y|x)\partial y = E(y|x) - E(y|x) = 0$. Therefore $E\{(y-\pi)^2\} = E\{(y-\hat{y})^2\} + E\{(\hat{y}-\pi)^2\} \geq E\{(y-\hat{y})^2\}$, and the assertion is proved.

We might note that the definition of the conditional expectation implies that

$$\begin{aligned} E(xy) &= \int_x\int_y xyf(x,y)\partial y\partial x \\ &= \int_x x\left\{\int_y yf(y|x)\partial y\right\}f(x)\partial x \\ &= E(x\hat{y}). \end{aligned} \tag{1.7}$$

When the equation $E(xy) = E(x\hat{y})$ is rewritten as

$$E\big\{x(y-\hat{y})\big\} = 0, \tag{1.8}$$

it may be described as an orthogonality condition. This condition indicates that the prediction error $y-\hat{y}$ is uncorrelated with x. The result is intuitively appealing; for, if the error were correlated with x, then we should not be using the information of x efficiently in forming $\hat{y}$.

Condtional Expectations and Linear Regression

If the joint distribution of x and y is a normal distribution, then we can make rapid headway in finding an expression for the function $E(y|x)$. In the case of a normal distribution, we have

$$E(y|x) = \alpha + \beta x, \tag{1.9}$$

which is to say that the conditional expectation of y given x is a linear function of x. Equation (1.9) is described as a linear regression equation; and we shall explain this terminology later.

The object is to find expressions for α and β which are in terms of the first-order and second-order moments of the joint distribution. That is to say, we wish to express α and β in terms of the expectations $E(x)$, $E(y)$, the variances $V(x)$, $V(y)$ and the covariance $C(x, y)$.

Admittedly, if we had already pursued the theory of the Normal distribution to the extent of demonstrating that the regression equation is a linear equation, then we should have already discovered these expressions for α and β. However, our present purposes are best served by taking equation (1.9) as our starting point; and we are prepared to regard the linearity of the regression equation as an assumption in its own right rather than as a deduction from the assumption of a normal distribution.

Let us begin by multiplying equation (1.9) throughout by $f(x)$, and let us proceed to integrate with respect to x. This gives us the equation

$$E(y) = \alpha + \beta E(x), \tag{1.10}$$

whence

$$\alpha = E(y) - \beta E(x). \tag{1.11}$$

Equation (1.10) shows that the regression line passes through the point $E(x, y) = \{E(x), E(y)\}$ which is the expected value of the joint distribution.

By putting (1.11) into (1.9), we find that

$$E(y|x) = E(y) + \beta\big\{x - E(x)\big\}, \tag{1.12}$$

which shows how the conditional expectation of y differs from the unconditional expectation in proportion to the error of predicting x by taking its expected value.

Now let us multiply (1.9) by x and $f(x)$ and then integrate with respect to x to provide

$$E(xy) = \alpha E(x) + \beta E(x^2). \tag{1.13}$$

Multiplying (1.10) by $E(x)$ gives

$$E(x)E(y) = \alpha E(x) + \beta\big\{E(x)\big\}^2, \tag{1.14}$$

whence, on taking (1.14) from (1.13), we get

$$E(xy) - E(x)E(y) = \beta\Big[E(x^2) - \big\{E(x)\big\}^2\Big], \tag{1.15}$$

which implies that

$$\begin{aligned}\beta &= \frac{E(xy) - E(x)E(y)}{E(x^2) - \big\{E(x)\big\}^2} \\ &= \frac{E\Big[\big\{x - E(x)\big\}\big\{y - E(y)\big\}\Big]}{E\Big[\big\{x - E(x)\big\}^2\Big]} \\ &= \frac{C(x,y)}{V(x)}.\end{aligned} \tag{1.16}$$

Thus we have expressed α and β in terms of the moments $E(x)$, $E(y)$, $V(x)$ and $C(x,y)$ of the joint distribution of x and y.

Example. Let $x = \xi + \eta$ be an observed random variable which combines a signal component ξ and a noise component η. Imagine that the two components are uncorrelated with $C(\xi,\eta) = 0$, and let $V(\xi) = \sigma_\xi^2$ and $V(\eta) = \sigma_\eta^2$. The object is to extract the signal from the observation.

According to the formulae of (1.12) and (1.16), the expectation of the signal conditional upon the observation is

$$E(\xi|x) = E(\xi) + \frac{C(x,\xi)}{V(x)}\big\{x - E(x)\big\}. \tag{1.17}$$

Given that ξ and η are uncorrelated, it follows that

$$V(x) = V(\xi + \eta) = \sigma_\xi^2 + \sigma_\eta^2 \tag{1.18}$$

and that

$$C(x,\xi) = V(\xi) + C(\xi,\eta) = \sigma_\xi^2. \tag{1.19}$$

Therefore

$$E(\xi|x) = E(\xi) + \frac{\sigma_\xi^2}{\sigma_\xi^2 + \sigma_\eta^2}\big\{x - E(x)\big\}. \tag{1.20}$$

2. SIMULTANEOUS-EQUATIONS BIAS

Bias in the OLS Estiamation of the Consumption Function

In elementary macroeconomic theory, a simple model of the economy is postulated which comprises two equations:

$$y = c + i, \tag{2.1}$$

$$c = \alpha + \beta y + \varepsilon. \tag{2.2}$$

Here y stands for the gross product of the economy, which is also the income of consumers, i stands for investment and c stands for consumption. An additional identity $s = y - c$ or $s = i$, where s is savings, is also entailed. The disturbance term ε, which is omitted from the usual presentation in economics textbooks, is assumed to be independent of the variable i.

On substituting the consumption function of (2.2) into the income identity of (2.1) and rearranging the result, we find that

$$y = \frac{1}{1-\beta}\big(\alpha + i + \varepsilon\big), \tag{2.3}$$

from which

$$y_t - \bar{y} = \frac{1}{1-\beta}\big(i_t - \bar{i} + \varepsilon_t - \bar{\varepsilon}\big). \tag{2.4}$$

The ordinary least-squares estimator of the parameter β, which is called the marginal propensity to consume, gives rise to the following equation:

$$\hat{\beta} = \beta + \frac{\sum(y_t - \bar{y})\varepsilon_t}{\sum(y_t - \bar{y})^2}. \tag{2.5}$$

Equation (2.3), which shows that y is dependent on ε, suggests that $\hat{\beta}$ cannot be a consistent estimator of β.

To determine the probability limit of the estimator, we must assess the separate probability limits of the numerator and the denominator of the term on the RHS of (2.5).

The following results are available:

$$\lim \frac{1}{T}\sum_{t=1}^{T}(i_t - \bar{i})^2 = m_{ii} = V(i),$$

$$\operatorname{plim}\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})^2 = \frac{m_{ii} + \sigma^2}{(1-\beta)^2} = V(y), \tag{2.6}$$

$$\operatorname{plim}\frac{1}{T}\sum_{t=1}^{T}(y_t - \bar{y})\varepsilon_t = \frac{\sigma^2}{1-\beta} = C(y, \varepsilon).$$

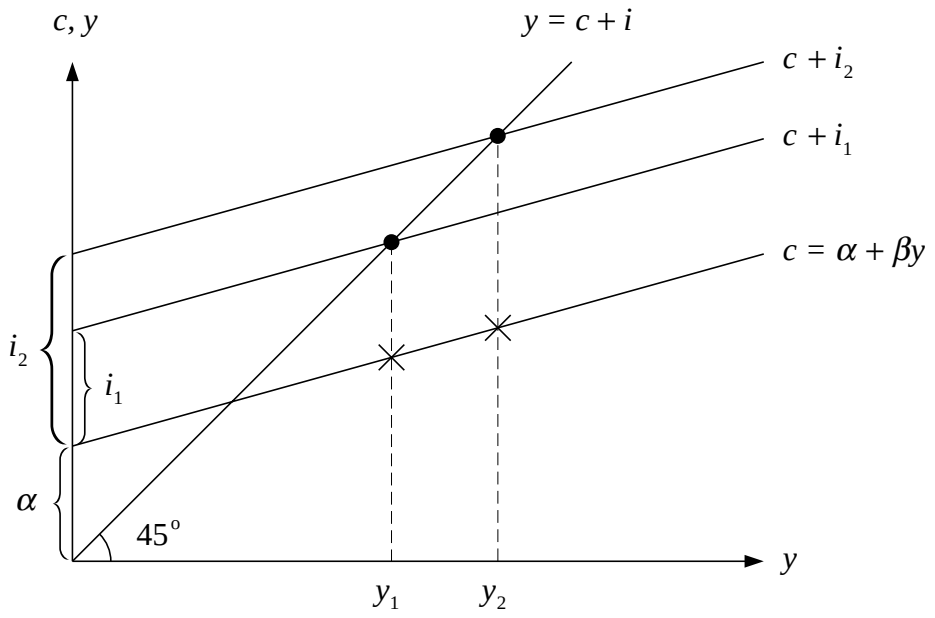

Figure 1. If the only source of variation in y is the variation in i, then the observations on y and c will delineate the consumption function.

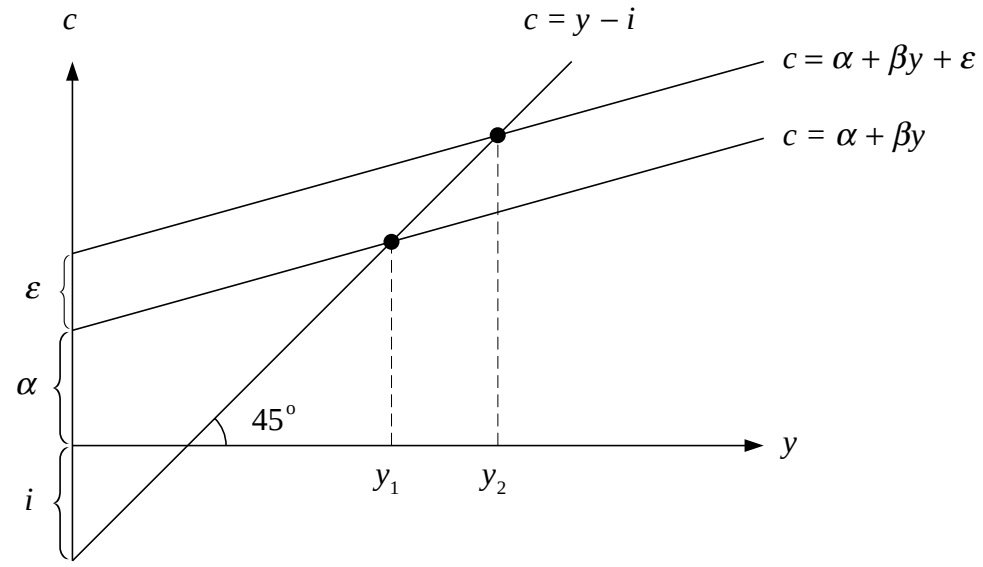

Figure 2. If the only source of variation in y are the disturbances to c, then the observations on y and c will line along a 45° line.

The results indicate that

$$\begin{aligned}\text{plim } \hat{\beta} &= \beta + \frac{\sigma^2(1-\beta)}{m_{ii}+\sigma^2} \\ &= \frac{\beta m_{ii} + \sigma^2}{m_{ii}+\sigma^2};\end{aligned} \tag{2.7}$$

and it can be seen that the limiting value of $\hat{\beta}$ has an upward bias which increases as the ratio σ^2/m_{ii} increases.

On the assumption that the model is valid, it is easy to understand why the parameter of the regression of c on y exceeds the value of the marginal propensity to consume. We can do so by considering the extreme cases.

Imagine, first, that $\sigma^2 = V(\varepsilon) = 0$. Then the only source of variation in y and c is the variation in i. In that case, the parameter of the regression of c on y will coincide with β. This is illustrated in Figure 1. Now imagine, instead, that i is constant and that the only variations in c and y are due ε which is disturbs consumption. Then the expected value of consumption is provided by the equation $c = y - i$ in which the coefficient associated with y is unity. Figure 2 illustrates this case. Assuming now that both $m_{ii} > 0$ and $\sigma^2 > 0$, it follows that the value of the regression parameter must lie somewhere in the interval $[\beta, 1]$.

Although it may be inappropriate for estimating the structural parameter β, the direct regression of c on y does provide the conditional expectation $E(c|y)$; and this endows it with a validity which it retains even if the Keynesian model of (2.1) and (2.2) is misspecified.

In fact, the simple Keynesian model of (2.1) and (2.2) is more an epigram than a serious scientific theory. Common sense dictates that we should give more credence to the estimate of the conditional expectation $E(c|y)$ than to a putative estimate of the marginal propensity to consume devised within the context of a doubtful model.

3. THE CONCEPT OF EXOGENEITY

Exogeneity

As the etymology suggests, the term *exogenenous* variables are generated outside the system or the equation of interest. The dependent variables, that are generated within the system, are described as *endogenous.*

In so far as their values affect those of the dependent variables, the exogenous variables are apt to be described as explanatory variables or regressors. The concept of exogeneity is appropriate to circumstances where regression equations correspond to components of the economy that can be regarded as

structural entities embodying causal relationships running from the explanatory variables to the dependent variables.

The parameters of a structural regression equation are its intrinsic properties; and they are expected to be invariant in respect of any changes in the circumstances affecting the generation of the exogenous variables. Moreover, the validity of the ordinary methods of regression analysis are dependent upon the truth of the assumption that the disturbance term is uncorrelated with the explanatory variables that are regarded as exogenous.

There are also circumstances where regression equations represent a statistical relationships that corresponds neither to a structural relationship nor to a causal connection. The parameters of such a regression equation are a reflection of the joint distribution of the variables comprised by the equation. They are expected to remain constant only in so far as the joint distribution is unchanged.

In the case of a purely statistical regression equation, it is generally inappropriate to categorise the variables as exogenous or endogenous. Moreover, the role of the disturbance term of the structural regression equation, which is deemed to represent the aggregate effect of the omitted exogenous variables, is taken by the prediction error, which, by construction, is uncorrelated with the regressors.

The concept of exogeneity has been analysed and elaborated in an influential article of Engle, Hendry and Richard (1983). Their analysis is concerned primarily with structural regression equations. Nevertheless, it goes some way towards bridging the gap that exists between the structural and the statistical interpretations. It must be said that, in the process, the authors have altered the meaning of the word exogeneity to the extent that they are prepared to find so-called conditions of *weak exogeneity* in equations that are devoid of any structural or behavioural interpretation.

The discussion of exogeneity tends to be heavily burdened by special definitions and by neologisms, but we shall attempt to convey the basic ideas in the simple of terms and within the context of bivariate relationships.

We should begin by noting that there is nothing inherent is the structure of a bivariate distribution to indicate which of the variables is the dependent variable and which is the explanatory variable. The two variables are appointed to play these roles by choosing one or other of the available factorisations that depict the joint distribution as the product of a marginal distribution and a conditional distribution:

$$f(x, y) = f(y|x)f(x) = f(x|y)f(y). \tag{3.1}$$

We shall choose y as the dependent variable and x as the regressor. In that case, it is the conditional distribution $f(y|x)$ that embodies the regression equation. If x qualifies as a (weakly) exogenous variable, then we are in a

position to ignore the details of the marginal distribution $f(x)$ when making inferences about the parameters of the conditional distribution.

In terms of the existing notation, we have

$$\begin{aligned} E(y|x) &= \mu_{y|x} = \alpha + \beta x, \\ E(x) &= \mu_x, \end{aligned} \tag{3.2}$$

where

$$\beta = \frac{\sigma_{xy}}{\sigma_{xy}} = \frac{\rho\sigma_y}{\sigma_x} \qquad \text{and} \qquad \alpha = \mu_y - \beta\mu_x. \tag{3.3}$$

We may also define the disturbance terms of the conditional and the marginal distributions, which are

$$y_t - \mu_{y|x} = \varepsilon_t \sim N(0, \sigma^2) \qquad \text{and} \qquad x_t - \mu_x = \nu_t \sim N(0, \sigma_x^2). \tag{3.4}$$

By construction, these are statistically independent with $C(\varepsilon_t, \nu_t) = 0$. There is also a specification for $V(\varepsilon_t) = \sigma^2$:

$$\sigma^2 = \sigma_y^2(1 - \rho^2) = \sigma_{yy} - \frac{\sigma_{xy}^2}{\sigma_{xx}}. \tag{3.5}$$

The factorisation of the bivariate distribution has entailed the replacement of the parameter set $\Phi = \{\mu_y, \mu_x, \sigma_x^2, \sigma_y^2, \rho\}$ by two parameter sets, which are $\Lambda_1 = \{\alpha, \beta, \sigma^2\}$ and $\Lambda_2 = \{\mu_x, \sigma_x^2\}$. To show the dependence of the distributions upon the parameters, we may write the chosen factorisation as

$$f(x_t, y_t; \Phi) = f(y_t|x_t; \Lambda_1) f(x_t; \Lambda_2). \tag{3.6}$$

We define the parameters of interest to be a function $\Psi = g(\Lambda_1)$ of the parameters of the conditional distribution.

We are now in a position to supply the central definition of Engle *et al.*:

(3.7) The variable x_t is said to be *weakly exogenous* for $\Psi = g(\Lambda_1)$ if and only if the factorisation of (3.6) generates a parameter space $\Lambda = \Lambda_1 \times \Lambda_2 = \{(\lambda_1, \lambda_2)\}$ in which the elements $\lambda_1 \in \Lambda_1$ and $\lambda_2 \in \Lambda_2$ are free to vary independently of each other.

This definition is concerned essentially with the efficiency of estimation. Thus, a variable x_t is defined to be weakly exogenous for the purposes of estimating the parameters of interest if it entails no loss of information to confine ones attention to the conditional distribution of y_t given x_t and to disregard the marginal distribution of x_t.

The definition would normally allow us to describe the argument x_t of the marginal distribution $f(x_t)$, as well as the argument y_t of $f(y_t)$, as an exogenous variable when there is no structural information regarding the parameters of Λ_1 and Λ_2 to constrain their independent variation. However, the full implications of the definition are best understood in the context of a simple example.

Example. Consider the so-called cobweb model that depicts an agricultural market in which the price is determined in consequence of the current supply and in which the supply has been determined by the price of the previous period. The resulting structural equations are

$$p_t = \beta q_t + \varepsilon_t, \qquad \text{where} \qquad \varepsilon_t \sim N(0, \sigma^2), \tag{3.8}$$

$$q_t = \pi_q p_{t-1} + \nu_{qt}, \qquad \text{where} \qquad \nu_{qt} \sim N(0, \sigma_q^2). \tag{3.9}$$

Here, p_t and q_t are the logarithms of price and quantity respectively, which have been adjusted by subtracting the sample means in order to eliminate the intercept terms from the equations. They are in place of y_t and x_t respectively. The value of $1/\beta$ is the price elasticity of demand, and that of π_q is the price elasticity of supply.

It is assumed that $C(\varepsilon_t, \nu_{qt}) = 0$, which reflects the fact that the circumstances in which the agricultural product is created are remote from those in which it is marketed. It follows that $C(q_t, \nu_{qt}) = 0$, which guarantees that q_t is exogenous with respect to equation (3.8) in the conventional sense. However, in view of the feedback that runs from (3.8) to (3.9), we choose to describe q_t as a predetermined variable rather than an exogenous variable.

The variables p_t and q_t also have a purely statistical joint distribution with constant mean values, which gives rise to what are described as the reduced-form equations:

$$p_t = \pi_p p_{t-1} + \nu_{pt}, \qquad \text{where} \qquad \nu_{pt} \sim N(0, \sigma_p^2), \tag{3.10}$$

$$q_t = \pi_q p_{t-1} + \nu_{qt}, \qquad \text{where} \qquad \nu_{qt} \sim N(0, \sigma_q^2). \tag{3.11}$$

It is assumed that $C(\nu_{pt}, \nu_{qt}) = \sigma_{pq} \neq 0$.

From the conditional distribution of p_t given p_t, we obtain

$$\begin{aligned} E(p_t|q_t) &= E(p_t) + \frac{C(p_t, q_t)}{V(q_t)}\{q_t - E(q_t)\} \\ &= \pi_p p_{t-1} + \frac{\sigma_{pq}}{\sigma_{qq}}\{q_t - \pi_q p_{t-1}\}, \end{aligned} \tag{3.12}$$

from which

$$p_t = (\pi_p - \beta\pi_q)p_{t-1} + \beta q_t + \varepsilon_t. \tag{3.13}$$

Here, ε_t is, by construction, uncorrelated with ν_{qt}.

The comparison of equation (3.13) with equation (3.8) makes it clear that the cobweb model embodies the restriction that $\pi_p - \beta\pi_q = 0$. (Notice that, in the absence of the condition $C(\varepsilon_t, \nu_{qt}) = 0$ affecting equations (3.8) and (3.9), equation (3.13) could not be identified with (3.8).)

Imagine that $\beta \in \Lambda_1$ alone is the parameter of interest. Then the restriction that excludes p_{t-1} from equation (3.8) will make q_t weakly exogenous for β.

There is, however, the matter of the dynamic stability of the system to be considered. It is natural to assume that the disturbances to the cobweb system will give rise to damped cycles in the prices and quantities. This necessitates that the coefficient of equation (3.10) obeys the condition that $|\pi_p| < 1$. Substitution of equation (3.9) into equation (3.8) gives

$$p_t = \beta\pi_q p_{t-1} + (\varepsilon_t + \beta\nu_{qt}). \tag{3.14}$$

This is an alternative rendering of equation (3.10) which shows again that $\pi_p = \beta\pi_q$.

When the stability of the system is taken into account, there is a connection between the parameters $\beta \in \Lambda_1$ and $\pi_q \in \Lambda_2$ of the conditional and the marginal distributions, such that their values are inversely related. Knowing the value of π_q will enable one to delimit the permissible values of β. Thus, in circumstances where the dynamic stability of the system is a necessary assumption, the variable x_t is no longer weakly exogenous for β according to the definition of (3.7).

The foregoing example can be used to illustrate two further definitions that have been cited by Engle *et al.*:

(3.15) The variable q_t is said to be *predetermined* in equation (3.8) if and only if $C(q_t, \varepsilon_{t+i}) = 0$ for all $i \geq 0$, whereas it is said to be *strictly exogenous* in (3.8) if and only if $C(q_t, \varepsilon_{t+i}) = 0$ for all i.

These are, in fact, the conventional definitions of exogenous and predetermined variables. It is notable that they make reference only to the equation in question and not to the parameters of interest therein.

For the condition of strict exogeneity to be satisfied in equation (3.8), it would be necessary to eliminate the feedback to equation (3.9). This would suggest that the producers have no regard to previous or current prices.

4. DIAGONALISATION OF A SYMMETRIC MATRIX

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.

(4.1) 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.

(4.2) 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

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

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 elements and with zeros in the non-diagonal positions. Post-multiplying 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

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

The Geometry of Quadratice 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

$$z_1^2 + z_2^2 = r^2. \tag{4.4}$$

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

$$z_1 = r\cos(\omega), \quad \text{and} \quad z_2 = r\sin(\omega). \tag{4.5}$$

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

$$\lambda_1 y_1^2 + \lambda_2 y_2^2 = r^2, \tag{4.6}$$

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

$$y_1 = \frac{z_1}{\sqrt{\lambda_1}} = \frac{r}{\sqrt{\lambda_1}}\cos(\omega), \quad y_2 = \frac{z_2}{\sqrt{\lambda_2}} = \frac{r}{\sqrt{\lambda_2}}\sin(\omega). \tag{4.7}$$

We can write equation (4.6) in matrix notation as

$$r^2 = [\, y_1 \quad y_2 \,] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = z_1^2 + z_2^2. \tag{4.8}$$

This implies

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \tag{4.9}$$

and

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}. \tag{4.10}$$

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

$$a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 = r^2; \tag{4.11}$$

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

$$\begin{aligned} r^2 &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \begin{bmatrix} y_1 & y_2 \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = z_1^2 + z_2^2, \end{aligned} \tag{4.12}$$

where θ is the angle which the pincipcal axis of the ellipse makes with the horizontal. The coefficients of the equation (4.11) are the elements of the matrix

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} \lambda_1\cos^2\theta + \lambda_2\sin^2\theta & (\lambda_2 - \lambda_1)\cos\theta\sin\theta \\ (\lambda_2 - \lambda_1)\cos\theta\sin\theta & \lambda_1\sin^2\theta + \lambda_2\cos^2\theta \end{bmatrix}. \tag{4.13}$$

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

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \sqrt{\lambda_1}(x_1\cos\theta - x_2\sin\theta) \\ \sqrt{\lambda_2}(x_1\sin\theta + x_2\cos\theta) \end{bmatrix}, \tag{4.14}$$

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

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}. \tag{4.15}$$

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

$$\begin{aligned}
&\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} \cos\theta \\ -\sin\theta \end{bmatrix} \\
&= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta \\ -\sin\theta \end{bmatrix} \\
&= \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \lambda_1 \begin{bmatrix} \cos\theta \\ -\sin\theta \end{bmatrix}.
\end{aligned} \tag{4.16}$$

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.

5. THE STATISTICAL PROPERTIES OF THE OLS ESIMATOR: UNBIASEDNESS AND EFFICIENCY

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

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'(X\beta + \varepsilon) \\
&= \beta + (X'X)^{-1}X'\varepsilon.
\end{aligned} \tag{5.1}$$

The expectation is

$$\begin{aligned}
E(\hat{\beta}) &= \beta + (X'X)^{-1}X'E(\varepsilon) \\
&= \beta.
\end{aligned} \tag{5.2}$$

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

$$\begin{aligned}
D(\hat{\beta}) &= E\Big[\big\{\hat{\beta} - E(\hat{\beta})\big\}\big\{\hat{\beta} - E(\hat{\beta})\big\}'\Big] \\
&= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}.
\end{aligned} \tag{5.3}$$

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

(5.4) 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 let us write $A = (X'X)^{-1}X' + G$. Then $AX = I$ implies that $GX = 0$. It follows that

$$\begin{aligned} D(\beta^*) &= AD(y)A' \\ &= \sigma^2\big\{(X'X)^{-1}X' + G\big\}\big\{X(X'X)^{-1} + G'\big\} \\ &= \sigma^2(X'X)^{-1} + \sigma^2 GG' \\ &= D(\hat{\beta}) + \sigma^2 GG'. \end{aligned} \tag{5.5}$$

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

$$\begin{aligned} V(q'\beta^*) &= q'D(\hat{\beta})q + \sigma^2 q'GG'q \\ &\geq q'D(\hat{\beta})q = V(q'\hat{\beta}); \end{aligned} \tag{5.6}$$

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

Estimating the Variance of the Disturbance

The principle of least squares does not, of its own, suggest a means of estimating the disturbance variance $\sigma^2 = V(\varepsilon_t)$. However it is natural to estimate the moments of a probability distribution by their empirical counterparts. Given that $e_t = y_t - x_{t.}\hat{\beta}$ is an estimate of ε_t, it follows that $T^{-1}\sum_t e_t^2$ may be used to estimate σ^2. However, it transpires that this is biased. An unbiased estimate is provided by

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{T-k}\sum_{t=1}^{T} e_t^2 \\ &= \frac{1}{T-k}(y - X\hat{\beta})'(y - X\hat{\beta}). \end{aligned} \tag{5.7}$$

The unbiasedness of this estimate may be demonstrated by finding the expected value of $(y - X\hat{\beta})'(y - X\hat{\beta}) = y'(I - P)y$. Given that $(I - P)y = (I - P)(X\beta + \varepsilon) = (I - P)\varepsilon$ in consequence of the condition $(I - P)X = 0$, it follows that

$$E\big\{(y - X\hat{\beta})'(y - X\hat{\beta})\big\} = E(\varepsilon'\varepsilon) - E(\varepsilon' P\varepsilon). \tag{5.8}$$

The value of the first term on the RHS is given by

$$E(\varepsilon'\varepsilon) = \sum_{t=1}^{T} E(e_t^2) = T\sigma^2. \tag{5.9}$$

The value of the second term on the RHS is given by

$$\begin{aligned} E(\varepsilon' P\varepsilon) &= \text{Trace}\big\{E(\varepsilon' P\varepsilon)\big\} = E\big\{\text{Trace}(\varepsilon' P\varepsilon)\big\} = E\big\{\text{Trace}(\varepsilon\varepsilon' P)\big\} \\ &= \text{Trace}\big\{E(\varepsilon\varepsilon')P\big\} = \text{Trace}\big\{\sigma^2 P\big\} = \sigma^2 \text{Trace}(P) \\ &= \sigma^2 k. \end{aligned} \tag{5.10}$$

The final equality follows from the fact that $\text{Trace}(P) = \text{Trace}(I_k) = k$. Putting the results of (5.9) and (5.10) into (5.8), gives

$$E\big\{(y - X\hat{\beta})'(y - X\hat{\beta})\big\} = \sigma^2(T - k); \tag{5.11}$$

and, from this, the unbiasedness of the estimator in (5.7) follows directly.

A Note on Matrix Traces

The trace of a square matrix $A = [a_{ij}; i, j = 1, \ldots, n]$ is just the sum of its diagonal elements:

$$\text{Trace}(A) = \sum_{i=1}^{n} a_{ii}. \tag{5.12}$$

Let $A = [a_{ij}]$ be a matrix of order $n \times m$ and let $B = [b_{k\ell}]$ a matrix of order $m \times n$. Then

$$\begin{aligned} &AB = C = [c_{i\ell}] \quad \text{with} \quad c_{i\ell} = \sum_{j=1}^{m} a_{ij} b_{j\ell} \quad \text{and} \\ &BA = D = [d_{kj}] \quad \text{with} \quad d_{kj} = \sum_{\ell=1}^{n} b_{k\ell} a_{\ell j}. \end{aligned} \tag{5.13}$$

Now,

$$\begin{aligned}\text{Trace}(AB) &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji} \quad \text{and} \\ \text{Trace}(BA) &= \sum_{j=1}^{m}\sum_{\ell=1}^{n} b_{j\ell}a_{\ell j} = \sum_{\ell=1}^{n}\sum_{j=1}^{m} a_{\ell j}b_{j\ell}.\end{aligned} \tag{5.14}$$

But, apart from a minor change of notation, where ℓ replaces i, the expressions on the RHS are the same. It follows that $\text{Trace}(AB) = \text{Trace}(BA)$. The result can be extended to cover the cyclic permutation of any number of matrix factors. In the case of three factors A, B, C, we have

$$\text{Trace}(ABC) = \text{Trace}(CAB) = \text{Trace}(BCA). \tag{5.15}$$

A further permutation would give $\text{Trace}(BCA) = \text{Trace}(ABC)$, and we should be back where we started.

6. THE PARTITIONED REGRESSSION MODEL

Consider taking a regression equation in the form of

$$y = [\, X_1 \quad X_2 \,] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon. \tag{6.1}$$

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

$$X_1'X_1\beta_1 + X_1'X_2\beta_2 = X_1'y, \tag{6.2}$$

$$X_2'X_1\beta_1 + X_2'X_2\beta_2 = X_2'y. \tag{6.3}$$

From (6.2), we get the equation $X_1'X_1\beta_1 = X_1'(y - X_2\beta_2)$ which gives an expression for the leading subvector of $\hat{\beta}$:

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2). \tag{6.4}$$

To obtain an expression for $\hat{\beta}_2$, we must eliminate β_1 from equation (6.3). For this purpose, we multiply equation (6.2) by $X_2'X_1(X_1'X_1)^{-1}$ to give

$$X_2'X_1\beta_1 + X_2'X_1(X_1'X_1)^{-1}X_1'X_2\beta_2 = X_2'X_1(X_1'X_1)^{-1}X_1'y. \tag{6.5}$$

When the latter is taken from equation (6.3), we get

$$\Big\{X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\Big\}\beta_2 = X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y. \tag{6.6}$$

On defining

$$P_1 = X_1(X_1'X_1)^{-1}X_1', \tag{6.7}$$

can we rewrite (6.6) as

$$\Big\{X_2'(I - P_1)X_2\Big\}\beta_2 = X_2'(I - P_1)y, \tag{6.8}$$

whence

$$\hat{\beta}_2 = \Big\{X_2'(I - P_1)X_2\Big\}^{-1}X_2'(I - P_1)y. \tag{6.9}$$

Now let us investigate the effect that conditions of orthogonality amongst the regressors have upon the ordinary least-squares estimates of the regression parameters. Consider a partitioned regression model, which can be written as

$$y = [\,X_1, X_2\,]\begin{bmatrix}\beta_1\\ \beta_2\end{bmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon. \tag{6.10}$$

It can be assumed that the variables in this equation are in deviation form. Imagine that the columns of X_1 are orthogonal to the columns of X_2 such that $X_1'X_2 = 0$. This is the same as assuming that the empirical correlation between variables in X_1 and variables in X_2 is zero.

The effect upon the ordinary least-squares estimator can be seen by examining the partitioned form of the formula $\hat{\beta} = (X'X)^{-1}X'y$. Here we have

$$X'X = \begin{bmatrix}X_1'\\ X_2'\end{bmatrix}[\,X_1 \quad X_2\,] = \begin{bmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{bmatrix} = \begin{bmatrix}X_1'X_1 & 0\\ 0 & X_2'X_2\end{bmatrix}, \tag{6.11}$$

where the final equality follows from the condition of orthogonality. The inverse of the partitioned form of $X'X$ in the case of $X_1'X_2 = 0$ is

$$(X'X)^{-1} = \begin{bmatrix}X_1'X_1 & 0\\ 0 & X_2'X_2\end{bmatrix}^{-1} = \begin{bmatrix}(X_1'X_1)^{-1} & 0\\ 0 & (X_2'X_2)^{-1}\end{bmatrix}. \tag{6.12}$$

We also have

$$X'y = \begin{bmatrix}X_1'\\ X_2'\end{bmatrix}y = \begin{bmatrix}X_1'y\\ X_2'y\end{bmatrix}. \tag{6.13}$$

On combining these elements, we find that

$$\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}(X_1'X_1)^{-1} & 0\\ 0 & (X_2'X_2)^{-1}\end{bmatrix}\begin{bmatrix}X_1'y\\ X_2'y\end{bmatrix} = \begin{bmatrix}(X_1'X_1)^{-1}X_1'y\\ (X_2'X_2)^{-1}X_2'y\end{bmatrix}. \tag{6.14}$$

In this special case, the coefficients of the regression of y on $X = [X_1, X_2]$ can be obtained from the separate regressions of y on X_1 and y on X_2.

It should be understood that this result does not hold true in general. The general formulae for $\hat{\beta}_1$ and $\hat{\beta}_2$ are those which we have given already under (6.4) and (6.9):

$$\begin{aligned}\hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2),\\ \hat{\beta}_2 &= \left\{X_2'(I - P_1)X_2\right\}^{-1}X_2'(I - P_1)y, \quad P_1 = X_1(X_1'X_1)^{-1}X_1'.\end{aligned} \tag{6.15}$$

It can be confirmed easily that these formulae do specialise to those under (6.14) in the case of $X_1'X_2 = 0$.

The purpose of including X_2 in the regression equation when, in fact, interest is confined to the parameters of β_1 is to avoid falsely attributing the explanatory power of the variables of X_2 to those of X_1.

Let us investigate the effects of erroneously excluding X_2 from the regression. In that case, the estimate will be

$$\begin{aligned}\tilde{\beta}_1 &= (X_1'X_1)^{-1}X_1'y\\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)\\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon.\end{aligned} \tag{6.16}$$

On applying the expectations operator to these equations, we find that

$$E(\tilde{\beta}_1) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2, \tag{6.17}$$

since $E\{(X_1'X_1)^{-1}X_1'\varepsilon\} = (X_1'X_1)^{-1}X_1'E(\varepsilon) = 0$. Thus, in general, we have $E(\tilde{\beta}_1) \neq \beta_1$, which is to say that $\tilde{\beta}_1$ is a biased estimator. The only circumstances in which the estimator will be unbiased are when either $X_1'X_2 = 0$ or $\beta_2 = 0$. In other circumstances, the estimator will suffer from a problem which is commonly described as *omitted-variables bias*.

We need to ask whether it matters that the estimated regression parameters are biased. The answer depends upon the use to which we wish to put the estimated regression equation. The issue is whether the equation is to be used simply for predicting the values of the dependent variable y or whether it is to be used for some kind of structural analysis.

If the regression equation purports to describe a structural or a behavioral relationship within the economy, and if some of the explanatory variables on

the RHS are destined to become the instruments of an economic policy, then it is important to have unbiased estimators of the associated parameters. For these parameters indicate the leverage of the policy instruments. Examples of such instruments are provided by interest rates, tax rates, exchange rates and the like.

On the other hand, if the estimated regression equation is to be viewed solely as a predictive device—that it to say, if it is simply an estimate of the function $E(y|x_1, \ldots, x_k)$ which specifies the conditional expectation of y given the values of $x_1, \ldots, x_n$—then, provided that the underlying statistical mechanism which has generated these variables is preserved, the question of the unbiasedness the regression estimates does not arise.

7. 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

$$\begin{aligned} y &= X\hat{\beta} + (y - X\hat{\beta}) \quad \text{and} \\ y - X\beta &= (X\hat{\beta} - X\beta) + (y - X\hat{\beta}), \end{aligned} \tag{7.1}$$

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

$$\begin{aligned} y &= Py + (I - P)y \qquad \text{and} \\ \varepsilon &= P\varepsilon + (I - P)\varepsilon. \end{aligned} \tag{7.2}$$

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

$$\begin{aligned} \varepsilon'\varepsilon &= \varepsilon' P\varepsilon + \varepsilon'(I - P)\varepsilon \qquad \text{or} \\ \varepsilon'\varepsilon &= (X\hat{\beta} - X\beta)'(X\hat{\beta} - X\beta) + (y - X\hat{\beta})'(y - X\hat{\beta}). \end{aligned} \tag{7.3}$$

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

$$\begin{aligned}\hat{\sigma}^2 &= \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T - k} \\ &= \frac{\varepsilon'(I - P)\varepsilon}{T - k}.\end{aligned} \tag{7.4}$$

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

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

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

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:

(7.6) 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

$$\begin{aligned} CC' &= C_1C_1' + C_2C_2' = I_T \qquad \text{and} \\ C'C &= \begin{bmatrix} C_1'C_1 & C_1'C_2 \\ C_2'C_1 & C_2'C_2 \end{bmatrix} = \begin{bmatrix} I_k & 0 \\ 0 & I_{T-k} \end{bmatrix}. \end{aligned} \tag{7.7}$$

The first of these results allows us to set $I - P = I - C_1C_1' = C_2C_2'$. Now, if $\varepsilon \sim N(0, \sigma^2 I_T)$ 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

$$\begin{bmatrix} C_1'\varepsilon \\ C_2'\varepsilon \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 I_k & 0 \\ 0 & \sigma^2 I_{T-k} \end{bmatrix} \right), \tag{7.8}$$

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

$$\begin{aligned} \frac{\varepsilon' C_1C_1'\varepsilon}{\sigma^2} &= \frac{\varepsilon' P\varepsilon}{\sigma^2} \sim \chi^2(k) \quad \text{and} \\ \frac{\varepsilon' C_2C_2'\varepsilon}{\sigma^2} &= \frac{\varepsilon'(I-P)\varepsilon}{\sigma^2} \sim \chi^2(T-k) \end{aligned} \tag{7.9}$$

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

$$\frac{\varepsilon' C_1C_1'\varepsilon}{\sigma^2} + \frac{\varepsilon' C_2C_2'\varepsilon}{\sigma^2} = \frac{\varepsilon'\varepsilon}{\sigma^2} \sim \chi^2(T); \tag{7.10}$$

and thus the theorem is proved.

The statistic under (7.5) can now be expressed in the form of

$$F = \left\{ \frac{\varepsilon' P\varepsilon}{k} \Big/ \frac{\varepsilon'(I-P)\varepsilon}{T-k} \right\}. \tag{7.11}$$

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.

8. 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:

(8.1) 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,

(8.2) 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,

(8.3) 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,

(8.4) 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,

(8.5) 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

$$\hat{\beta} \sim N_k\{\beta, \sigma^2(X'X)^{-1}\}. \tag{8.6}$$

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

$$\hat{\beta}_1 \sim N_{k_1}\big(\beta_1, \sigma^2\{X_1'(I - P_2)X_1\}^{-1}\big), \tag{8.7}$$

$$\hat{\beta}_2 \sim N_{k_2}\big(\beta_2, \sigma^2\{X_2'(I - P_1)X_2\}^{-1}\big), \tag{8.8}$$

where $P_1 = X_1(X_1'X_1)^{-1}X_1'$ and $P_2 = X_2(X_2'X_2)^{-1}X_2'$. On applying the result under (8.2) to (8.6), we find that

$$\sigma^{-2}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \sim \chi^2(k). \tag{8.9}$$

Similarly, it follows from (8.7) and (8.8) that

$$\sigma^{-2}(\hat{\beta}_1 - \beta_1)'X_1'(I - P_2)X_1(\hat{\beta}_1 - \beta_1) \sim \chi^2(k_1), \tag{8.10}$$

$$\sigma^{-2}(\hat{\beta}_2 - \beta_2)'X_2'(I - P_1)X_2(\hat{\beta}_2 - \beta_2) \sim \chi^2(k_2). \tag{8.11}$$

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

$$\sigma^{-2} y'CC'y = \sigma^{-2}(y - X\hat{\beta})'(y - X\hat{\beta}) \sim \chi^2(T-k). \tag{8.12}$$

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

$$C(e, X\hat{\beta}) = (I-P)D(y)P' = \sigma^2(I-P)P' = 0, \tag{8.13}$$

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

$$\begin{aligned} &\sigma^{-2}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \sim \chi^2(k) \quad \text{and} \\ &\sigma^{-2}(y - X\hat{\beta})'(y - X\hat{\beta}) \sim \chi^2(T-k) \end{aligned} \tag{8.14}$$

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

$$\begin{aligned} F &= \left\{ \frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{k} \Big/ \frac{(y - X\hat{\beta})'(y - X\hat{\beta})}{T-k} \right\} \\ &= \frac{1}{\hat{\sigma}^2 k}(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta) \sim F(k, T-k). \end{aligned} \tag{8.15}$$

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 $y = X_1\beta_1 + X\beta_2 + \varepsilon$ 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

$$F = \frac{1}{\hat{\sigma}^2 k_2}(\hat{\beta}_2 - \beta_{2\diamond})'X_2'(I - P_1)X_2(\hat{\beta}_2 - \beta_{2\diamond}). \tag{8.16}$$

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 (8.16) with $\beta_2 = \beta_z$ and $X_2 = Z$ and where $P_1 = P_\iota = T^{-1}\iota\iota'$ 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 in 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 (8.16), a statistic may be derived in the form of

$$F = \frac{(\hat{\beta}_i - \beta_{i\diamond})^2}{\hat{\sigma}^2 w_{ii}}, \tag{8.17}$$

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

$$t = \frac{\hat{\beta}_i - \beta_{i\diamond}}{\sqrt{(\hat{\sigma}^2 w_{ii})}} \tag{8.18}$$

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.