LECTURE 2

Regression Analysis

The Multiple Regression Model in Matrices

Consider the regression equation

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \tag{1}$$

and imagine that T observations on the variables $y, x_1, \ldots, x_k$ are available, which are indexed by $t = 1, \ldots, T$. Then, the T realisations of the relationship can be written in the following form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{T1} & \ldots & x_{Tk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}. \tag{2}$$

This can be represented in summary notation by

$$y = X\beta + \varepsilon. \tag{3}$$

Our object is to derive an expression for the ordinary least-squares estimates of the elements of the parameter vector $\beta = [\beta_0, \beta_1, \ldots, \beta_k]'$. The criterion is to minimise a sum of squares of residuals, which can be written variously as

$$\begin{aligned} S(\beta) &= \varepsilon'\varepsilon \\ &= (y - X\beta)'(y - X\beta) \\ &= y'y - y'X\beta - \beta'X'y + \beta'X'X\beta \\ &= y'y - 2y'X\beta + \beta'X'X\beta. \end{aligned} \tag{4}$$

Here, to reach the final expression, we have used the identity $\beta'X'y = y'X\beta$, which comes from the fact that the transpose of a scalar—which may be construed as a matrix of order 1×1—is the scalar itself.

The first-order conditions for the minimisation are found by differentiatiating the function with respect to the vector β and by setting the result to zero. According to the rules of matrix differentiation, which are easily verified, the derivative is

$$\frac{\partial S}{\partial \beta} = -2y'X + 2\beta'X'X. \tag{5}$$

Setting this to zero gives $0 = \beta'X'X - y'X$, which is transposed to provide the so-called normal equations:

$$X'X\beta = X'y. \tag{6}$$

On the assumption that the inverse matrix exists, the equations have a unique solution, which is the vector of ordinary least-squares estimates:

$$\hat{\beta} = (X'X)^{-1}X'y. \tag{7}$$

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 $Py = X\hat{\beta}$, 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 exactly the condition that is expressed by the normal equations (6).

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 $e = y - X\hat{\beta} = (I - P)y$, where $P = X(X'X)^{-1}X'$ is a symmetric idempotent matrix, which has the properties that $P = P' = P^2$. Then, in consequence of these conditions and of the equivalent condition that $P'(I - P) = 0$, it follows that

$$\begin{aligned} y'y &= \big\{Py + (I - P)y\big\}'\big\{Py + (I - P)y\big\} \\ &= y'Py + y'(I - P)y \\ &= \hat{\beta}'X'X\hat{\beta} + e'e. \end{aligned} \tag{8}$$

This is 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

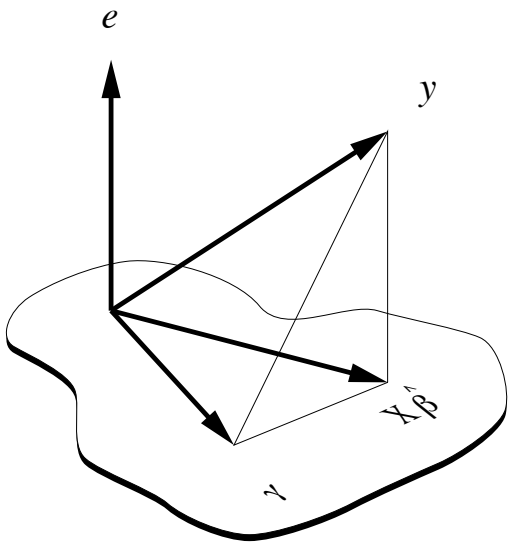

Figure 1. The vector $Py = X\hat{\beta}$ is formed by the orthogonal projection of the vector y onto the subspace spanned by the columns of the matrix X.

$\hat{\beta}'X'X\hat{\beta}$ 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

$$\begin{aligned}(y-\gamma)'(y-\gamma) &= \big\{(y-X\hat{\beta})+(X\hat{\beta}-\gamma)\big\}'\big\{(y-X\hat{\beta})+(X\hat{\beta}-\gamma)\big\} \\ &= \big\{(I-P)y+P(y-g)\big\}'\big\{(I-P)y+P(y-g)\big\}.\end{aligned} \tag{9}$$

The properties of the projector P, which have been used in simplifying equation (9), indicate that

$$\begin{aligned}(y-\gamma)'(y-\gamma) &= y'(I-P)y+(y-g)'P(y-g) \\ &= e'e+(X\hat{\beta}-\gamma)'(X\hat{\beta}-\gamma).\end{aligned} \tag{10}$$

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.

A summary measure of the extent to which the ordinary least-squares regression accounts for the observed vector y is provided by the coefficient of

determination. This is defined by

$$R^2 = \frac{\hat{\beta}' X' X \hat{\beta}}{y'y} = \frac{y'Py}{y'y}. \tag{11}$$

The measure is just the square of the cosine of the angle between the vectors y and $Py = X\hat{\beta}$; and the inequality $0 \leq R^2 \leq 1$ follows from the fact that the cosine of any angle must lie between -1 and $+1$.

The Partitioned Regression Model

Consider taking the regression equation of (3) in the form of

$$y = [\, X_1 \quad X_2 \,] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon. \tag{12}$$

Here, $[X_1, X_2] = X$ and $[\beta_1', \beta_2']' = \beta$ are obtained by partitioning the matrix X and vector β in a conformable manner. The normal equations of (6) can be partitioned likewise. Writing the equations without the surrounding matrix braces gives

$$X_1'X_1\beta_1 + X_1'X_2\beta_2 = X_1'y, \tag{13}$$

$$X_2'X_1\beta_1 + X_2'X_2\beta_2 = X_2'y. \tag{14}$$

From (13), 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 $\hat{\beta}$:

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2). \tag{15}$$

To obtain an expression for $\hat{\beta}_2$, we must eliminate β_1 from equation (14). For this purpose, we multiply equation (13) by $X_2'X_1(X_1'X_1)^{-1}$ to give

$$X_2'X_1\beta_1 + X_2'X_1(X_1'X_1)^{-1}X_1'X_2\beta_2 = X_2'X_1(X_1'X_1)^{-1}X_1'y. \tag{16}$$

When the latter is taken from equation (14), we get

$$\Big\{X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2\Big\}\beta_2 = X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y. \tag{17}$$

On defining

$$P_1 = X_1(X_1'X_1)^{-1}X_1', \tag{18}$$

equation (17) can be written as

$$\Big\{X_2'(I - P_1)X_2\Big\}\beta_2 = X_2'(I - P_1)y, \tag{19}$$

whence

$$\hat{\beta}_2 = \Big\{X_2'(I - P_1)X_2\Big\}^{-1} X_2'(I - P_1)y. \tag{20}$$

The Regression Model with an Intercept

Now consider again the equations

$$y_t = \alpha + x_{t.}\beta + \varepsilon_t, \quad t = 1, \ldots, T, \tag{21}$$

which comprise T observations of a regression model with an intercept term α, denoted by β_0 in equation (1), and with k explanatory variables in $x_{t.} = [x_{t1}, x_{t2}, \ldots, x_{tk}]$. These equations can also be represented in a matrix notation as

$$y = \iota\alpha + Z\beta + \varepsilon. \tag{22}$$

Here, the vector $\iota = [1, 1, \ldots, 1]'$, which consists of T units, is described alternatively as the dummy vector or the summation vector, whilst $Z = [x_{tj}; t = 1, \ldots T; j = 1, \ldots, k]$ is the matrix of the observations on the explanatory variables.

Equation (22) can be construed as a case of the partitioned regression equation of (12). By setting $X_1 = \iota$ and $X_2 = Z$ and by taking $\beta_1 = \alpha$, $\beta_2 = \beta_z$ in equations (15) and (20), we derive the following expressions for the estimates of the parameters α, β_z:

$$\hat{\alpha} = (\iota'\iota)^{-1}\iota'(y - Z\hat{\beta}_z), \tag{23}$$

$$\begin{aligned} \hat{\beta}_z =& \big\{Z'(I - P_\iota)Z\big\}^{-1} Z'(I - P_\iota)y, \quad \text{with} \\ P_\iota =& \iota(\iota'\iota)^{-1}\iota' = \frac{1}{T}\iota\iota'. \end{aligned} \tag{24}$$

To understand the effect of the operator P_ι in this context, consider the following expressions:

$$\begin{aligned} \iota'y &= \sum_{t=1}^{T} y_t, \\ (\iota'\iota)^{-1}\iota'y &= \frac{1}{T}\sum_{t=1}^{T} y_t = \bar{y}, \\ P_\iota y &= \iota\bar{y} = \iota(\iota'\iota)^{-1}\iota'y = [\bar{y}, \bar{y}, \ldots, \bar{y}]'. \end{aligned} \tag{25}$$

Here, $P_\iota y = [\bar{y}, \bar{y}, \ldots, \bar{y}]'$ is a column vector containing T repetitions of the sample mean. From the expressions above, it can be understood that, if $x = [x_1, x_2, \ldots x_T]'$ is vector of T elements, then

$$x'(I - P_\iota)x = \sum_{t=1}^{T} x_t(x_t - \bar{x}) = \sum_{t=1}^{T}(x_t - \bar{x})x_t = \sum_{t=1}^{T}(x_t - \bar{x})^2. \tag{26}$$

The final equality depends on the fact that $\sum(x_t - \bar{x})\bar{x} = \bar{x}\sum(x_t - \bar{x}) = 0$.

The Regression Model in Deviation Form

Consider the matrix of cross-products in equation (24). This is

$$Z'(I - P_\iota)Z = \{(I - P_\iota)Z\}'\{Z(I - P_\iota)\} = (Z - \bar{Z})'(Z - \bar{Z}). \tag{27}$$

Here, $\bar{Z} = [(\bar{x}_j; j = 1, \ldots, k)_t; t = 1, \ldots, T]$ is a matrix in which the generic row $(\bar{x}_1, \ldots, \bar{x}_k)$, which contains the sample means of the k explanatory variables, is repeated T times. The matrix $(I - P_\iota)Z = (Z - \bar{Z})$ is the matrix of the deviations of the data points about the sample means, and it is also the matrix of the residuals of the regressions of the vectors of Z upon the summation vector ι. The vector $(I - P_\iota)y = (y - \iota\bar{y})$ may be described likewise.

It follows that the estimate of β_z is precisely the value which would be obtained by applying the technique of least-squares regression to a meta-equation

$$\begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_T - \bar{y} \end{bmatrix} = \begin{bmatrix} x_{11} - \bar{x}_1 & \ldots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & \ldots & x_{2k} - \bar{x}_k \\ \vdots & & \vdots \\ x_{T1} - \bar{x}_1 & \ldots & x_{Tk} - \bar{x}_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 - \bar{\varepsilon} \\ \varepsilon_2 - \bar{\varepsilon} \\ \vdots \\ \varepsilon_T - \bar{\varepsilon} \end{bmatrix}, \tag{28}$$

which lacks an intercept term. In summary notation, the equation may be denoted by

$$y - \iota\bar{y} = [Z - \bar{Z}]\beta_z + (\varepsilon - \bar{\varepsilon}). \tag{29}$$

Observe that it is unnecessary to take the deviations of y. The result is the same whether we regress y or $y - \iota\bar{y}$ on $[Z - \bar{Z}]$. The result is due to the symmetry and idempotency of the operator $(I - P_\iota)$ whereby $Z'(I - P_\iota)y = \{(I - P_\iota)Z\}'\{(I - P_\iota)y\}$.

Once the value for $\hat{\beta}$ is available, the estimate for the intercept term can be recovered from the equation (23) which can be written as

$$\begin{aligned} \bar{\alpha} &= \bar{y} - \bar{Z}\hat{\beta}_z \\ &= \bar{y} - \sum_{j=1}^{k} \bar{x}_j\hat{\beta}_j. \end{aligned} \tag{30}$$

The Assumptions of the Classical Linear Model

In characterising the properties of the ordinary least-squares estimator of the regression parameters, some conventional assumptions are made regarding the processes which generate the observations.

Let the regression equation be

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \tag{31}$$

which is equation (1) again; and imagine, as before, that there are T observations on the variables. Then, these can be arrayed in the matrix form of (2) for which the summary notation is

$$y = X\beta + \varepsilon, \tag{32}$$

where $y = [y_1, y_2, \ldots, y_T]'$, $\varepsilon = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T]'$, $\beta = [\beta_0, \beta_1, \ldots, \beta_k]'$ and $X = [x_{tj}]$ with $x_{t0} = 1$ for all t.

The first of the assumptions regarding the disturbances is that they have an expected value of zero. Thus

$$E(\varepsilon) = 0 \qquad \text{or, equivalently,} \qquad E(\varepsilon_t) = 0, \quad t = 1, \ldots, T. \tag{33}$$

Next it is assumed that the disturbances are mutually uncorrelated and that they have a common variance. Thus

(34)
$$D(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I \qquad \text{or, equivalently,} \qquad E(\varepsilon_t\varepsilon_s) = \begin{cases} \sigma^2, & \text{if } t = s; \\ 0, & \text{if } t \neq s. \end{cases}$$

If t is a temporal index, then these assumptions imply that there is no inter-temporal correlation in the sequence of disturbances. In an econometric context, this is often implausible; and the assumption will be relaxed at a later stage.

The next set of assumptions concern the matrix X of explanatory variables. A conventional assumption, borrowed from the experimental sciences, is that

$$X \text{ is a nonstochastic matrix with linearly independent columns.} \tag{35}$$

The condition of linear independence is necessary if the separate effects of the k variables are to be distinguishable. If the condition is not fulfilled, then it will not be possible to estimate the parameters in β uniquely, although it may be possible to estimate certain weighted combinations of the parameters.

Often, in the design of experiments, an attempt is made to fix the explanatory or experimental variables in such a way that the columns of the matrix X are mutually orthogonal. The device of manipulating only one variable at a

time will achieve the effect. The danger of miss-attributing the effects of one variable to another is then minimised.

In an econometric context, it is often more appropriate to regard the elements of X as random variables in their own right, albeit that we are usually reluctant to specify in detail the nature of the processes that generate the variables. Thus, it may be declared that

(36) The elements of X are random variables which are distributed independently of the elements of ε.

The consequence of either of these assumptions (35) or (36) is that

$$E(X'\varepsilon|X) = X'E(\varepsilon) = 0. \tag{37}$$

In fact, for present purposes, it makes little difference which of these assumptions regarding X is adopted; and, since the assumption under (35) is more briefly expressed, we shall adopt it in preference.

The first property to be deduced from the assumptions is that

(38) The ordinary least-square regression estimator $\hat{\beta} = (X'X)^{-1}X'y$ is unbiased such that $E(\hat{\beta}) = \beta$.

To demonstrate this, we may write

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'y \\ &= (X'X)^{-1}X'(X\beta + \varepsilon) \\ &= \beta + (X'X)^{-1}X'\varepsilon. \end{aligned} \tag{39}$$

Taking expectations gives

$$\begin{aligned} E(\hat{\beta}) &= \beta + (X'X)^{-1}X'E(\varepsilon) \\ &= \beta. \end{aligned} \tag{40}$$

Notice that, in the light of this result, equation (39) now indicates that

$$\hat{\beta} - E(\hat{\beta}) = (X'X)^{-1}X'\varepsilon. \tag{41}$$

The next deduction is that

(42) The variance–covariance matrix of the ordinary least-squares regression estimator is $D(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

To demonstrate the latter, we may write a sequence of identities:

$$\begin{aligned} D(\hat{\beta}) &= E\Big\{\big[\hat{\beta} - E(\hat{\beta})\big]\big[\hat{\beta} - E(\hat{\beta})\big]'\Big\} \\ &= E\big\{(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\big\} \\ &= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\{\sigma^2 I\}X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned} \tag{43}$$

The second of these equalities follows directly from equation (41).

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:

$$\text{Trace}(A) = \sum_{i=1}^{n} a_{ii}. \tag{44}$$

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

$$\begin{aligned} AB = C = [c_{i\ell}] \quad \text{with} \quad c_{i\ell} = \sum_{j=1}^{m} a_{ij}b_{j\ell} \quad \text{and} \\ BA = D = [d_{kj}] \quad \text{with} \quad d_{kj} = \sum_{\ell=1}^{n} b_{k\ell}a_{\ell j}. \end{aligned} \tag{45}$$

Now,

$$\begin{aligned} \text{Trace}(AB) &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji} \quad \text{and} \\ \text{Trace}(BA) &= \sum_{j=1}^{m}\sum_{\ell=1}^{n} b_{j\ell}a_{\ell j} = \sum_{\ell=1}^{n}\sum_{j=1}^{m} a_{\ell j}b_{j\ell}. \end{aligned} \tag{46}$$

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

$$\text{Trace}(ABC) = \text{Trace}(CAB) = \text{Trace}(BCA). \tag{47}$$

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

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

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{T-k}\sum_{t=1}^{T} e_t^2 \\ &= \frac{1}{T-k}(y - X\hat{\beta})'(y - X\hat{\beta}).\end{aligned} \tag{48}$$

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

$$E\big\{(y - X\hat{\beta})'(y - X\hat{\beta})\big\} = E(\varepsilon'\varepsilon) - E(\varepsilon' P\varepsilon). \tag{49}$$

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

$$E(\varepsilon'\varepsilon) = \sum_{t=1}^{T} E(e_t^2) = T\sigma^2. \tag{50}$$

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

$$\begin{aligned}E(\varepsilon' P\varepsilon) &= \text{Trace}\big\{E(\varepsilon' P\varepsilon)\big\} = E\big\{\text{Trace}(\varepsilon' P\varepsilon)\big\} = E\big\{\text{Trace}(\varepsilon\varepsilon' P)\big\} \\ &= \text{Trace}\big\{E(\varepsilon\varepsilon')P\big\} = \text{Trace}\big\{\sigma^2 P\big\} = \sigma^2\text{Trace}(P) \\ &= \sigma^2 k.\end{aligned} \tag{51}$$

The final equality follows from the fact that $\text{Trace}(P) = \text{Trace}(I_k) = k$. Putting the results of (50) and (51) into (49), gives

$$E\big\{(y - X\hat{\beta})'(y - X\hat{\beta})\big\} = \sigma^2(T - k); \tag{52}$$

and, from this, the unbiasedness of the estimator in (48) follows directly.

Statistical Properties of the OLS Estimator

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

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'(X\beta + \varepsilon)\\ &= \beta + (X'X)^{-1}X'\varepsilon.\end{aligned} \tag{53}$$

The expectation is

$$\begin{aligned}E(\hat{\beta}) &= \beta + (X'X)^{-1}X'E(\varepsilon)\\ &= \beta.\end{aligned} \tag{54}$$

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

$$\begin{aligned}D(\hat{\beta}) &= E\Big[\big\{\hat{\beta} - E(\hat{\beta})\big\}\big\{\hat{\beta} - E(\hat{\beta})\big\}'\Big]\\ &= (X'X)^{-1}X'E(\varepsilon\varepsilon')X(X'X)^{-1}\\ &= \sigma^2(X'X)^{-1}.\end{aligned} \tag{55}$$

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,

(56) 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

$$\begin{aligned}D(\beta^*) &= AD(y)A'\\ &= \sigma^2\big\{(X'X)^{-1}X' + G\big\}\big\{X(X'X)^{-1} + G'\big\}\\ &= \sigma^2(X'X)^{-1} + \sigma^2GG'\\ &= D(\hat{\beta}) + \sigma^2GG'.\end{aligned} \tag{57}$$

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

$$\begin{aligned} V(q'\beta^*) &= q'D(\hat{\beta})q + \sigma^2 q'GG'q \\ &\geq q'D(\hat{\beta})q = V(q'\hat{\beta}); \end{aligned} \tag{58}$$

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

Orthogonality and Omitted-Variables Bias

Let us now investigate the effect that a condition of orthogonality amongst the regressors might have upon the ordinary least-squares estimates of the regression parameters. Let us take the partitioned regression model of equation (12) which was written as

$$y = [\, X_1, X_2 \,] \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon. \tag{59}$$

We may assume that the variables in this equation are in deviation form. Let us 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 imagining that the empirical correlation between variables in X_1 and variables in X_2 is zero.

To see the effect upon the ordinary least-squares estimator, we may examine the partitioned form of the formula $\hat{\beta} = (X'X)^{-1}X'y$. Here, there is

$$X'X = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} [\, X_1 \quad X_2 \,] = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} = \begin{bmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{bmatrix}, \tag{60}$$

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

$$(X'X)^{-1} = \begin{bmatrix} X_1'X_1 & 0 \\ 0 & X_2'X_2 \end{bmatrix}^{-1} = \begin{bmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{bmatrix}. \tag{61}$$

Ther is also

$$X'y = \begin{bmatrix} X_1' \\ X_2' \end{bmatrix} y = \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}. \tag{62}$$

On combining these elements, we find that

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} (X_1'X_1)^{-1} & 0 \\ 0 & (X_2'X_2)^{-1} \end{bmatrix} \begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix} = \begin{bmatrix} (X_1'X_1)^{-1}X_1'y \\ (X_2'X_2)^{-1}X_2'y \end{bmatrix}. \tag{63}$$

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 recognised that this result does not hold true in general. The general formulae for $\hat{\beta}_1$ and $\hat{\beta}_2$ are those that have been given already under (15) and (20):

$$\begin{aligned} \hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'(y - X_2\hat{\beta}_2), \\ \hat{\beta}_2 &= \left\{X_2'(I - P_1)X_2\right\}^{-1}X_2'(I - P_1)y, \quad P_1 = X_1(X_1'X_1)^{-1}X_1'. \end{aligned} \tag{64}$$

It is readily confirmed that these formulae do specialise to those under (63) in the case of $X_1'X_2 = 0$.

The purpose of including X_2 in the regression equation when, in fact, our 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, our estimate will be

$$\begin{aligned} \tilde{\beta}_1 &= (X_1'X_1)^{-1}X_1'y \\ &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon. \end{aligned} \tag{65}$$

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

$$E(\tilde{\beta}_1) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2, \tag{66}$$

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 of the regression parameters does not arise.

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

$$\begin{aligned} L &= (y - X\beta)'(y - X\beta) + 2\lambda'(R\beta - r) \\ &= y'y - 2y'X\beta + \beta'X'X\beta + 2\lambda'R\beta - 2\lambda'r. \end{aligned} \tag{67}$$

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

$$2\beta'X'X - 2y'X + 2\lambda'R = 0, \tag{68}$$

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

$$X'X\beta + R'\lambda = X'y. \tag{69}$$

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:

$$\begin{bmatrix} X'X & R' \\ R & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} = \begin{bmatrix} X'y \\ r \end{bmatrix}. \tag{70}$$

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

$$\text{Rank} \begin{bmatrix} X \\ R \end{bmatrix} = k \tag{71}$$

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

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix}, \tag{72}$$

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 (68) gives an expression for β in the form of

$$\begin{aligned} \beta^* &= (X'X)^{-1}X'y - (X'X)^{-1}R'\lambda \\ &= \hat{\beta} - (X'X)^{-1}R'\lambda, \end{aligned} \tag{73}$$

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

$$r = R\hat{\beta} - R(X'X)^{-1}R'\lambda, \tag{74}$$

from which

$$\lambda = \{R(X'X)^{-1}R'\}^{-1}(R\hat{\beta} - r). \tag{75}$$

On substituting this expression back into equation (73), we get

$$\beta^* = \hat{\beta} - (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}(R\hat{\beta} - r). \tag{76}$$

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, then, on he supposition that the restrictions are valid, it follows that $E(\beta^* - \beta) = 0$, so that β^* is also unbiased.

Next, consider the expression

$$\begin{aligned} \beta^* - \beta &= [I - (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}R](\hat{\beta} - \beta) \\ &= (I - P_R)(\hat{\beta} - \beta), \end{aligned} \tag{77}$$

where

$$P_R = (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}R. \tag{78}$$

The expression comes from taking β from both sides of (76) and from recognising that $R\hat{\beta} - r = R(\hat{\beta} - \beta)$. It can be seen that P_R is an idempotent matrix that is subject to the conditions that

$$P_R = P_R^2, \quad P_R(I - P_R) = 0 \quad \text{and} \quad P_R'X'X(I - P_R) = 0. \tag{79}$$

From equation (77), it can be deduced that

$$\begin{aligned} D(\beta^*) &= (I - P_R)E\{(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\}(I - P_R) \\ &= \sigma^2(I - P_R)(X'X)^{-1}(I - P_R) \\ &= \sigma^2[(X'X)^{-1} - (X'X)^{-1}R'\{R(X'X)^{-1}R'\}^{-1}R(X'X)^{-1}]. \end{aligned} \tag{80}$$

Regressions on Orthogonal Variables

The probability that two vectors of empirical observations should be precisely orthogonal each other must be zero, unless such a circumstance has been carefully contrived by designing an experiment. However, there are some important cases where the explanatory variables of a regression are artificial variables that are either designed to be orthogonal, or are naturally orthogonal.

An important example concerns polynomial regressions. A simple experiment will serve to show that a regression on the powers of the integers $t = 0, 1, \ldots, T$, which might be intended to estimate a function that is trending with time, is fated to collapse if the degree of the polynomial to be estimated is in excess 3 or 4. The problem is that the matrix $X = [1, t, t^2, \ldots, t^n; t = 1, \ldots T]$ of the powers of the integer t is notoriously ill-conditioned. The consequence is that cross-product matrix $X'X$ will be virtually singular. The proper recourse is to employ a basis set of orthogonal polynomials as the regressors.

Another important example of orthogonal regressors concerns a Fourier analysis. Here, the explanatory variables are sampled from a set of trigonometric functions that have angular velocities or frequencies that are evenly distributed in an interval running from zero to π radians per sample period.

If the sample is indexed by $t = 0, 1, \ldots T - 1$, then the frequencies in question will be defined by $\omega_j = 2\pi j/T; j = 0, 1, \ldots, [T/2]$, where $[T/2]$ denotes the integer quotient of the division of T by 2. These are the so-called Fourier frequencies. The object of a Fourier analysis is to express the elements of the sample as a weighted sum of sine and cosine functions as follows:

$$y_t = \alpha_0 + \sum_{j=1}^{[T/2]} \{\alpha_j \cos(\omega_j t) + \beta \sin(\omega_j t)\}; \quad t = 0, 1, \ldots, T - 1. \tag{81}$$

A trigonometric function with a frequency of ω_j completes exactly j cycles in the T periods that are spanned by the sample. Moreover, there will be exactly as many regressors as there are elements within the sample. This is evident is the case where $T = 2n$ is an even number. At first sight, it might appear that the are $T + 2$ trigonometrical functions. However, for integral values of t, is transpires that

$$\begin{aligned} &\cos(\omega_0 t) = \cos 0 = 1, && \sin(\omega_0 t) = \sin 0 = 0, \\ &\cos(\omega_n t) = \cos(\pi t) = (-1)^t, && \sin(\omega_n t) = \sin(\pi t) = 0; \end{aligned} \tag{82}$$

so, in fact, there are only T nonzero functions.

Equally, it can be see that, in the case where T is odd, there are also exactly T nonzero functions defined on the set of Fourier frequencies. These consist of the cosine function at zero frequency, which is the constant function associated with α_0, together with sine and cosine functions at the Fourier frequencies indexed by $j = 1, \ldots, (T-1)/2$.

The vectors of the generic trigonometric regressors may be denoted by

$$c_j = [c_{0j}, c_{1j}, \ldots c_{T-1,j}]' \qquad \text{and} \qquad s_j = [s_{0j}, s_{1j}, \ldots s_{T-1,j}]', \tag{83}$$

where $c_{tj} = \cos(\omega_j t)$ and $s_{tj} = \sin(\omega_j t)$. The vectors of the ordinates of functions of different frequencies are mutually orthogonal. Therefore, amongst these vectors, the following orthogonality conditions hold:

$$\begin{aligned} &c_i' c_j = s_i' s_j = 0 \quad \text{if} \quad i \neq j, \\ &\text{and} \quad c_i' s_j = 0 \quad \text{for all} \quad i, j. \end{aligned} \tag{84}$$

In addition, there are some sums of squares which can be taken into account in computing the coefficients of the Fourier decomposition:

$$\begin{aligned} &c_0' c_0 = \iota' \iota = T, \qquad s_0' s_0 = 0, \\ &c_j' c_j = s_j' s_j = T/2 \quad \text{for} \quad j = 1, \ldots, [(T-1)/2] \end{aligned} \tag{85}$$

Th proofs ar given in a brief appendix at the end of this section. When $T = 2n$, there is $\omega_n = \pi$ and, therefore, in view of (82), there is also

$$s_n' s_n = 0, \qquad \text{and} \qquad c_n' c_n = T. \tag{86}$$

The "regression" formulae for the Fourier coefficients can now be given. First, there is

$$\alpha_0 = (i'i)^{-1} i'y = \frac{1}{T} \sum_t y_t = \bar{y}. \tag{87}$$

Then, for $j = 1, \ldots, [(T-1)/2]$, there are

$$\alpha_j = (c_j'c_j)^{-1}c_j'y = \frac{2}{T}\sum_t y_t \cos\omega_j t, \tag{88}$$

and

$$\beta_j = (s_j's_j)^{-1}s_j'y = \frac{2}{T}\sum_t y_t \sin\omega_j t. \tag{89}$$

If $T = 2n$ is even, then there is no coefficient β_n and there is

$$\alpha_n = (c_n'c_n)^{-1}c_n'y = \frac{1}{T}\sum_t (-1)^t y_t. \tag{90}$$

By pursuing the analogy of multiple regression, it can be seen, in view of the orthogonality relationships, that there is a complete decomposition of the sum of squares of the elements of the vector y, which is given by

$$y'y = \alpha_0^2\iota'\iota + \sum_{j=1}^{[T/2]} \left\{\alpha_j^2 c_j'c_j + \beta_j^2 s_j's_j\right\}. \tag{91}$$

Now consider writing $\alpha_0^2\iota'\iota = \bar{y}^2\iota'\iota = \bar{y}'\bar{y}$, where $\bar{y}' = [\bar{y}, \bar{y}, \ldots, \bar{y}]$ is a vector whose repeated element is the sample mean $\bar{y}$. It follows that $y'y - \alpha_0^2\iota'\iota = y'y - \bar{y}'\bar{y} = (y-\bar{y})'(y-\bar{y})$. Then, in the case where $T = 2n$ is even, the equation can be writen as

$$(y-\bar{y})'(y-\bar{y}) = \frac{T}{2}\sum_{j=1}^{n-1}\left\{\alpha_j^2 + \beta_j^2\right\} + T\alpha_n^2 = \frac{T}{2}\sum_{j=1}^{n}\rho_j^2. \tag{92}$$

where $\rho_j = \alpha_j^2 + \beta_j^2$ for $j = 1, \ldots, n-1$ and $\rho_n = 2\alpha_n$. A similar expression exists when T is odd, with the exceptions that α_n is missing and that the summation runs to $(T-1)/2$. It follows that the variance of the sample can be expressed as

$$\frac{1}{T}\sum_{t=0}^{T-1}(y_t - \bar{y})^2 = \frac{1}{2}\sum_{j=1}^{n}(\alpha_j^2 + \beta_j^2). \tag{93}$$

The proportion of the variance which is attributable to the component at frequency ω_j is $(\alpha_j^2 + \beta_j^2)/2 = \rho_j^2/2$, where ρ_j is the amplitude of the component.

The number of the Fourier frequencies increases at the same rate as the sample size T. Therefore, if the variance of the sample remains finite, and

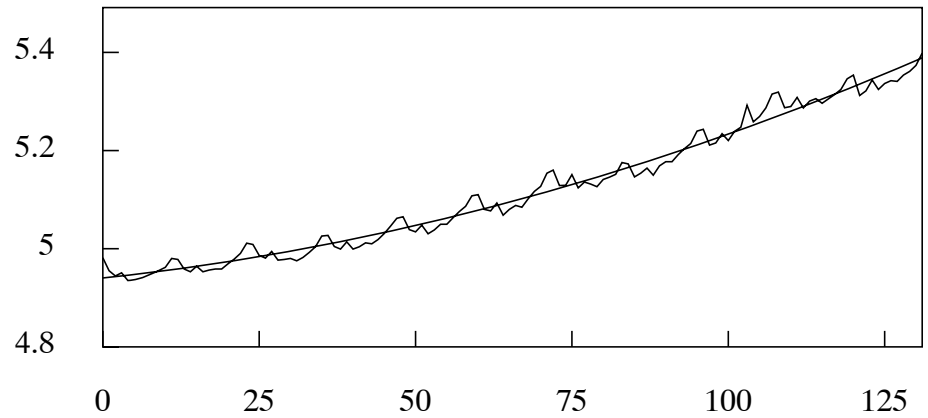

Figure 9. The plot of 132 monthly observations on the U.S. money supply, beginning in January 1960. A quadratic function has been interpolated through the data.

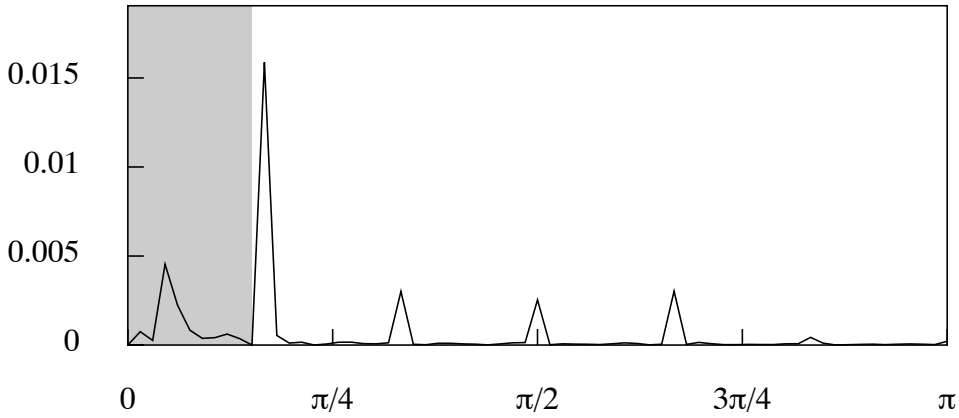

Figure 10. The periodogram of the residuals of the logarithmic money-supply data.

if there are no regular harmonic components in the process generating the data, then we can expect the proportion of the variance attributed to the individual frequencies to decline as the sample size increases. If there is such a regular component within the process, then we can expect the proportion of the variance attributable to it to converge to a finite value as the sample size increases.

In order provide a graphical representation of the decomposition of the sample variance, we must scale the elements of equation (36) by a factor of T. The graph of the function $I(\omega_j) = (T/2)(\alpha_j^2 + \beta_j^2)$ is know as the periodogram.

Figure 9 shows the logarithms of a monthly sequence of 132 observations of the US money supply through which a quadratic function has been interpolated.

This provides a simple way of characterising the growth of the money supply over the period in question. The pattern of seasonal fluctuations is remarkably regular, as can be see from the residuals from the quadratic detrending.

The peridogram of the residual sequence is shown in Figure 10. This has a prominent spike at the frequency value of $\pi/6$ radians or 30 degrees per month, which is the fundamental seasonal frequency. Smaller spikes are seen at 60, 90, 120, and 150 degrees, which are the harmonics of the fundamental frequency. Their presence reflects the fact that the pattern of the seasonal fluctuations is more complicated than that of a simple sinusoidal fluctuation at the seasonal frequency.

The peridodogram also shows a significant spectral mass within the frequency range $[0\pi/6]$. This mass properly belongs to the trend; and, if the trend had been adequately estimated, then its effect would not be present in the residual, which would then show even greater regularity. In lecture 9, we will show how a more fitting trend function can be estimated.

Appendix: Harmonic Cycles

If a trigonometrical function completes an integral number of cycles in T periods, then the sum of its ordinates at the points $t = 0, 1, \ldots, T-1$ is zero. We state this more formally as follows:

(94) Let $\omega_j = 2\pi j/T$ where $j \in \{0, 1, \ldots, T/2\}$, if T is even, and $j \in \{0, 1, \ldots, (T-1)/2\}$, if T is odd. Then

$$\sum_{t=0}^{T-1} \cos(\omega_j t) = \sum_{t=0}^{T-1} \sin(\omega_j t) = 0.$$

Proof. We have

$$\begin{aligned}\sum_{t=0}^{T-1} \cos(\omega_j t) &= \frac{1}{2} \sum_{t=0}^{T-1} \{\exp(i\omega_j t) + \exp(-i\omega_j t)\} \\ &= \frac{1}{2} \sum_{t=0}^{T-1} \exp(i2\pi jt/T) + \frac{1}{2} \sum_{t=0}^{T-1} \exp(-i2\pi jt/T).\end{aligned}$$

By using the formula $1 + \lambda + \cdots + \lambda^{T-1} = (1 - \lambda^T)/(1 - \lambda)$, we find that

$$\sum_{t=0}^{T-1} \exp(i2\pi jt/T) = \frac{1 - \exp(i2\pi j)}{1 - \exp(i2\pi j/T)}.$$

But Euler's equation indicates that $\exp(i2\pi j) = \cos(2\pi j) + i \sin(2\pi j) = 1$, so the numerator in the expression above is zero, and hence $\sum_t \exp(i2\pi j/T) = 0$.

By similar means, it can be show that $\sum_t \exp(-i2\pi j/T) = 0$; and, therefore, it follows that $\sum_t \cos(\omega_j t) = 0$.

An analogous proof shows that $\sum_t \sin(\omega_j t) = 0$.

The proposition of (94) is used to establish the orthogonality conditions affecting functions with an integral number of cycles.

(95) Let $\omega_j = 2\pi j/T$ and $\psi_k = 2\pi k/T$ where $j, k \in 0, 1, \ldots, T/2$ if T is even and $j, k \in 0, 1, \ldots, (T-1)/2$ if T is odd. Then

$$\text{(a)} \qquad \sum_{t=0}^{T-1} \cos(\omega_j t)\cos(\psi_k t) = 0 \qquad \text{if} j \neq k,$$

$$\sum_{t=0}^{T-1} \cos^2(\omega_j t) = T/2,$$

$$\text{(b)} \qquad \sum_{t=0}^{T-1} \sin(\omega_j t)\sin(\psi_k t) = 0 \qquad \text{if} j \neq k,$$

$$\sum_{t=0}^{T-1} \sin^2(\omega_j t) = T/2,$$

$$\text{(c)} \qquad \sum_{t=0}^{T-1} \cos(\omega_j t)\sin(\psi_k t) = 0 \qquad \text{if} j \neq k.$$

Proof. From the formula $\cos A \cos B = \frac{1}{2}\{\cos(A+B) + \cos(A-B)\}$, we have

$$\begin{aligned}\sum_{t=0}^{T-1} \cos(\omega_j t)\cos(\psi_k t) &= \frac{1}{2}\sum \{\cos([\omega_j + \psi_k]t) + \cos([\omega_j - \psi_k]t)\} \\ &= \frac{1}{2}\sum_{t=0}^{T-1} \{\cos(2\pi[j+k]t/T) + \cos(2\pi[j-k]t/T)\}.\end{aligned}$$

We find, in consequence of (94), that if $j \neq k$, then both terms on the RHS vanish, which gives the first part of (a). If $j = k$, then $\cos(2\pi[j-k]t/T) = \cos 0 = 1$ and so, whilst the first term vanishes, the second terms yields the value of T under summation. This gives the second part of (a).

The proofs of (b) and (c) follow along similar lines once the relevant trigonometrical identities have been invoked.