21 : CHAPTER

Seemingly-Unrelated Regressions

The Algebra of the Kronecker Product. Consider the matrix equation $Y = AXB'$ where

$$\begin{aligned} Y &= [y_{kl}]; k = 1, \ldots, r, l = 1, \ldots, s, \\ X &= [x_{ij}]; i = 1, \ldots, m, j = 1, \ldots, n, \\ A &= [a_{ki}]; k = 1, \ldots, r, i = 1, \ldots, m, \\ B &= [b_{lj}]; l = 1, \ldots, s, j = 1, \ldots, n. \end{aligned} \tag{1}$$

The object is to reformulate this matrix equation so that it can be treated as an ordinary vector equation. Amongst the advantages which this will bring is the possibility of solving the equation by the methods which are commonly applied in finding the solutions to vector equations.

Therefore consider writing $Y = AXB'$ more explicitly as

$$\begin{aligned} [y_{.1}, y_{.2}, \ldots, y_{.s}] &= A[x_{.1}, x_{.2}, \ldots, x_{.n}][b'_{1.}, b'_{2.}, \ldots, b'_{s.}] \\ &= [Ax_{.1}, Ax_{.2}, \ldots, Ax_{.n}][b'_{1.}, b'_{2.}, \ldots, b'_{s.}]. \end{aligned} \tag{2}$$

In this notation, the expression $x_{.j}$ stands for the jth column of the matrix X whilst the notation $b_{l.}$ stands for the lth *row* of B. Therefore the transposed vector $b'_{l.} = [b_{l1}, b_{l2}, \ldots, b_{ln}]'$ is a *column* vector of n elements—as it must be if the multiplication of the two expressions on the RHS of (2) is to be properly defined. By performing that multiplication, we find that

$$\begin{aligned} [y_{.1}, y_{.2}, \ldots, y_{.s}] = \Big[&\{b_{11}Ax_{.1} + b_{12}Ax_{.2} + \cdots + b_{1n}Ax_{.n}\}, \\ &\{b_{21}Ax_{.1} + b_{22}Ax_{.2} + \cdots + b_{2n}Ax_{.n}\}, \ldots, \\ &\{b_{s1}Ax_{.1} + b_{s2}Ax_{.2} + \cdots + b_{sn}Ax_{.n}\}\Big] \end{aligned} \tag{3}$$

Here, each of the expressions on the the RHS within braces $\{,\}$ stands for one of the vectors $y_{.1}, y_{.2}, \ldots, y_{.s}$ on the LHS. These LHS vectors may be stacked vertically one below the other to form long vectors. When the RHS of the equation is rearranged likewise, a system is derived which takes the form of

$$\begin{bmatrix} y_{.1} \\ y_{.2} \\ \vdots \\ y_{.s} \end{bmatrix} = \begin{bmatrix} b_{11}A & b_{12}A & \ldots & b_{1n}A \\ b_{21}A & b_{22}A & \ldots & b_{2n}A \\ \vdots & \vdots & & \vdots \\ b_{s1}A & b_{s2}A & \ldots & b_{sn}A \end{bmatrix} \begin{bmatrix} x_{.1} \\ x_{.2} \\ \vdots \\ x_{.n} \end{bmatrix}. \tag{4}$$

The system can be written is a summary notation as

$$Y^c = (AXB')^c = (B \otimes A)X^c. \tag{5}$$

Here the long vectors Y^c and X^c are derived simply by slicing the matrices and rearranging the columns in the manner which we have described. The matrix $B \otimes A = [b_{lj}A]$, whose (lj)th partition contains the matrix $b_{lj}A$, is described as the Kronecker product of B and A.

The following rules govern the use of the Kronecker product:

$$\begin{aligned} &\text{(i)} && (A \otimes B)(C \otimes D) = AC \otimes BD, \\ &\text{(ii)} && A \otimes (B + C) = (A \otimes B) + (A \otimes B), \\ &\text{(iii)} && \lambda(A \otimes B) = \lambda A \otimes B = A \otimes \lambda B, \\ &\text{(iv)} && (A \otimes B)^{-1} = (A^{-1} \otimes B^{-1}). \end{aligned} \tag{6}$$

The Kronecker product is non-commutative, which is to say that $A \otimes B \neq B \otimes A$. However, observe that

$$A \otimes B = (A \otimes I)(I \otimes B) = (I \otimes B)(A \otimes I). \tag{7}$$

Systems with Multiple Outputs. The typical regression equation describes a system which transforms k observable inputs and a stochastic disturbance into a single output. We now wish to consider a system which produces M outputs. Consider, therefore, the equations

$$\begin{aligned} [y_{t1}, y_{t2}, \ldots, y_{tM}] &= [x_{t.}\beta_{.1}, x_{t.}\beta_{.2}, \ldots, x_{t.}\beta_{.M}] + [\varepsilon_{t1}, \varepsilon_{t2}, \ldots, \varepsilon_{tM}] \\ &= x_{t.}[\beta_{.1}, \beta_{.2}, \ldots, \beta_{.M}] + [\varepsilon_{t1}, \varepsilon_{t2}, \ldots, \varepsilon_{tM}]. \end{aligned} \tag{8}$$

Here the generic equation is

$$y_{tm} = x_{t.}\beta_{.m} + \varepsilon_{tm}; \tag{9}$$

and this has the form of a single regression equation. In a notation which mixes matrices and vectors, the system under (8) may be written as

$$y_{t.} = x_{t.}B + \varepsilon_{t.}, \tag{10}$$

where $B = [\beta_{.1}, \beta_{.2}, \ldots, \beta_{.M}]$, and T realisations of the latter may be compiled to give the equation

$$Y = XB + \mathcal{E}, \tag{11}$$

or

$$[y_{.1}, y_{.2}, \ldots, y_{.M}] = [x_{.1}, x_{.2}, \ldots, x_{.k}]B + [\varepsilon_{.1}, \varepsilon_{.2}, \ldots, \varepsilon_{.M}]. \tag{12}$$

When the latter equation is vectorised, we have

$$Y^c = (XBI)^c + \mathcal{E}^c = (I \otimes X)B^c + \mathcal{E}^c, \tag{13}$$

which can be written more explicitly as

$$\begin{bmatrix} y_{.1} \\ y_{.2} \\ \vdots \\ y_{.s} \end{bmatrix} = \begin{bmatrix} X & 0 & \ldots & 0 \\ 0 & X & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & X \end{bmatrix} \begin{bmatrix} \beta_{.1} \\ \beta_{.2} \\ \vdots \\ \beta_{.M} \end{bmatrix} + \begin{bmatrix} \varepsilon_{.1} \\ \varepsilon_{.2} \\ \vdots \\ \varepsilon_{.M} \end{bmatrix} = \begin{bmatrix} X\beta_{.1} \\ X\beta_{.2} \\ \vdots \\ X\beta_{.M} \end{bmatrix} + \begin{bmatrix} \varepsilon_{.1} \\ \varepsilon_{.2} \\ \vdots \\ \varepsilon_{.M} \end{bmatrix}. \tag{14}$$

Some assumptions must now be made regarding the disturbance terms of the model. We shall assume that, taken separately, the M equations $y_{.m} = X\beta_{.m} + \varepsilon_{.m}; m = 1, \ldots, M$ have the stochastic structure of the classical linear model; which is to say that the disturbances are independently and identically distributed with an expected value of zero and a common variance. However, we shall assume that the M contemporaneous disturbances in the vector $\varepsilon_{t.} = [\varepsilon_{t1}, \ldots, \varepsilon_{tM}]$ have nonzero covariances such that

$$D(\varepsilon_{t.}) = E(\varepsilon'_{t.}\varepsilon_{t.}) = \Sigma = [\sigma_{ml}] \quad \text{for all} \quad t. \tag{15}$$

Thus, if $\varepsilon_{.m}$ and $\varepsilon_{.l}$ are vectors of T disturbances from the equations $y_{.m} = X\beta_{.m} + \varepsilon_{.m}$ and $y_{.l} = X\beta_{.l} + \varepsilon_{.l}$ respectively, then we should have

$$\begin{aligned} &E(\varepsilon_{.m}) = E(\varepsilon_{.l}) = 0 \qquad \text{and} \\ &D(\varepsilon_{.m}) = \sigma_{mm}I_T, \quad D(\varepsilon_{.l}) = \sigma_{ll}I_T, \\ &C(\varepsilon_{.m}, \varepsilon_{.l}) = \sigma_{ml}I_T, \end{aligned} \tag{16}$$

where $C(\varepsilon_{.m}, \varepsilon_{.l}) = E(\varepsilon_{.m}\varepsilon'_{.l})$ is the covariance matrix of the two vectors. Putting these assumptions together, we get

$$E(\mathcal{E}^c) = 0 \qquad \text{and} \qquad D(\mathcal{E}^c) = E(\mathcal{E}^c\mathcal{E}^{c\prime}) = \Sigma \otimes I_T. \tag{17}$$

It may be appropriate to write these in a manner which makes them more explicit. First there is the assumption concerning the expected value of the long vector of disturbances. Writing this vector in transposed form gives

$$E(\mathcal{E}^{c\prime}) = E[\varepsilon'_{.1}, \varepsilon'_{.2}, \ldots, \varepsilon'_{.M}] = [\ 0,\ 0, \ldots, 0\ ]. \tag{18}$$

The assumptions concerning the dispersion matrix of this vector can be written as

$$\begin{aligned} D\begin{bmatrix} \varepsilon_{.1} \\ \varepsilon_{.2} \\ \vdots \\ \varepsilon_{.M} \end{bmatrix} &= E\begin{bmatrix} \varepsilon_{.1}\varepsilon'_{.1} & \varepsilon_{.1}\varepsilon'_{.2} & \cdots & \varepsilon_{.1}\varepsilon'_{.M} \\ \varepsilon_{.2}\varepsilon'_{.1} & \varepsilon_{.2}\varepsilon'_{.2} & \cdots & \varepsilon_{.2}\varepsilon'_{.M} \\ \vdots & \vdots & & \vdots \\ \varepsilon_{.M}\varepsilon'_{.1} & \varepsilon_{.M}\varepsilon'_{.2} & \cdots & \varepsilon_{.M}\varepsilon'_{.M} \end{bmatrix} \\ &= \begin{bmatrix} \sigma_{11}I_T & \sigma_{12}I_T & \cdots & \sigma_{1M}I_T \\ \sigma_{21}I_T & \sigma_{22}I_T & \cdots & \sigma_{2M}I_T \\ \vdots & \vdots & & \vdots \\ \sigma_{M1}I_T & \sigma_{M2}I_T & \cdots & \sigma_{MM}I_T \end{bmatrix}. \end{aligned} \tag{19}$$

It is common to denote the regression model $y = X\beta + \varepsilon$ in which $E(\varepsilon) = 0$ and $E(\varepsilon) = \sigma^2 Q$ by the triplet $(y; X\beta, \sigma^2 Q)$. Using the same notation, we may now denote the vectorised version of the model with M outputs as $(Y^c; (I \otimes X)B^c, \Sigma \otimes I)$. It is apparent that the two models are isomorphic, which is to say that they share the same structure. Therefore it is possible to estimate the parameters of the M-output model, once it has been cast in the appropriate form, by using methods which have been developed in the context of a single-equation model.

The approrpriate method is generalised least-squares regression. When it is applied to the model $(y; X\beta, \sigma^2 Q)$, this method delivers the estimate $\hat{\beta} = (X'Q^{-1}X)^{-1}X'Q^{-1}y$. When it is applied to the M-equation model the method delivers the estimate

$$\hat{B} = \Big\{(I \otimes X)'(\Sigma \otimes I)^{-1}(I \otimes X)\Big\}^{-1}(I \otimes X)'(\Sigma \otimes I)^{-1}Y^c. \tag{20}$$

The algebraic rules under (6) can now be invoked to simplify this result. It can be see that

$$\begin{aligned} \hat{B} &= (\Sigma^{-1} \otimes X'X)^{-1}(\Sigma \otimes X')^{-1}Y^c \\ &= \big\{I \otimes (X'X)^{-1}X'\big\}Y^c \\ &= \big\{(X'X)^{-1}X'Y\big\}^c. \end{aligned} \tag{21}$$

Thus it transpires that the efficient system-wide estimator amounts to nothing more than the repeated application of the ordinary least-squares procedure to generate the regression estimates $\hat{\beta}_{.m} = (X'X)^{-1}X'y_{.m}; m = 1, \ldots, M$.

We can use the residual vectors $e_{.m} = y_{.m} - X\hat{\beta}_{.m}$ from these M estimations to derive estimates of the elements of $\Sigma = [\sigma_{ml}]$. Thus an unbiased estimator of σ_{ml} is

$$\begin{aligned}\hat{\sigma}_{ml} &= \frac{e'_{.m}e_{.l}}{T-k} = \frac{(y_{.m} - X\hat{\beta}_{.m})'(y_{.l} - X\hat{\beta}_{.l})}{T-k} \\ &= \frac{y'_{.m}\{I - X(X'X)^{-1}X'\}y_{.l}}{T-k}.\end{aligned} \tag{22}$$

The reduction of the system-wide estimator to an M-fold application of ordinary least-square regression occurs only when all the variables in X are present in each of the M equations and when no other variables are present in any of them. If some of the variables are missing, or if we have *a priori* information relating to the parameter vectors $\beta_{.m}; m = 1, \ldots, M$, then, to obtain efficient estimates, we must use the available information on Σ. For example, let X_m be the submatrix containing only those variables which are present in the mth equation. Then the system of equations assumes the following form:

$$\begin{bmatrix} y_{.1} \\ y_{.2} \\ \vdots \\ y_{.s} \end{bmatrix} = \begin{bmatrix} X_1 & 0 & \ldots & 0 \\ 0 & X_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & X_M \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{bmatrix} + \begin{bmatrix} \varepsilon_{.1} \\ \varepsilon_{.2} \\ \vdots \\ \varepsilon_{.M} \end{bmatrix}. \tag{23}$$

This can be written in summary notation as

$$Y^c = W\delta + \mathcal{E}^c. \tag{24}$$

Now the block-diagonal or "staircase" matrix W no longer has the structure of a Kronecker product. Nor can the subvectors of $\delta' = [\beta'_1, \beta'_2, \ldots, \beta'_M]$ be stacked together in a matrix B, for the reason that they are liable to be of different lengths. The efficient generalised least-squares estimator of the parameters now takes the form of

$$\hat{\delta} = \left\{W'(\Sigma^{-1} \otimes I)W\right\}^{-1}W'(\Sigma^{-1} \otimes I)Y^c; \tag{25}$$

and there in no longer any possibility of simplifying or reducing the expression.