MINIMUM-MEAN-SQUARE-ERROR PREDICTION AND CONDITIONAL EXPECTATIONS

Consider a pair of random vectors x, y whose distribution is characterised by its first-order and second-order moments:

$$E\begin{bmatrix} y \\ x \end{bmatrix} = \begin{bmatrix} E(y) \\ E(x) \end{bmatrix}, \qquad D\begin{bmatrix} y \\ x \end{bmatrix} = \begin{bmatrix} D(y) & C(y,x) \\ C(x,y) & D(x) \end{bmatrix}. \tag{1}$$

A multivariate normal distribution can be characterised in this way.

The object is to predict the departure $y - E(y)$ of y from its expected value on the basis of the observed departure $x - E(x)$ of x from its expected value. If the predicted departure is a linear function of $x - E(x)$, then it can be expressed as

$$\hat{y} - E(y) = B'\{x - E(x)\}, \tag{2}$$

where $\hat{y}$ is the predicted value of y. Let the error of the prediction be denoted by ε. The combination of the prediction and the error gives

$$y - E(y) = B'\{x - E(x)\} + \varepsilon. \tag{3}$$

This is described as the linear regression relationship. An alternative way of denoting the relationship is to write it as

$$y = \alpha + B'x + \varepsilon, \quad \text{with} \quad \alpha = E(y) - B'E(x). \tag{4}$$

The matrix B' may be chosen so as to ensure that the prediction fulfils a criterion of optimality. If the prediction is to fulfill the minimum mean-square-error criterion, then a matrix must be chosen which ensures that the prediction error is uncorrelated with the variables in x. That is to say, we must have

$$C(\varepsilon, x) = E\Big\{\varepsilon\big[x - E(x)\big]'\Big\} = 0; \tag{5}$$

and it follows that

$$D\begin{bmatrix} \varepsilon \\ x \end{bmatrix} = \begin{bmatrix} D(\varepsilon) & 0 \\ 0 & D(x) \end{bmatrix}. \tag{6}$$

The reasoning behind this condition is the observation that, if the prediction error were systematically related to the variables of x, then some part of it could be predicted, with a consequent improvement in the prediction of y.

In order to find the value of B, we may begin by constructing the following equation:

$$\begin{bmatrix} I & -B' \\ 0 & I \end{bmatrix} \begin{bmatrix} y - E(y) \\ x - E(x) \end{bmatrix} = \begin{bmatrix} \varepsilon \\ x - E(x) \end{bmatrix}. \tag{7}$$

This is formed by rearranging equation (3) and thereafter by supplementing it with a trivial identity. Premultiplying equation (7) by the matrix

$$\begin{bmatrix} I & -B' \\ 0 & I \end{bmatrix}^{-1} = \begin{bmatrix} I & B' \\ 0 & I \end{bmatrix} \tag{8}$$

gives an equivalent system in the form of

$$\begin{bmatrix} y - E(y) \\ x - E(x) \end{bmatrix} = \begin{bmatrix} I & B' \\ 0 & I \end{bmatrix} \begin{bmatrix} \varepsilon \\ x - E(x) \end{bmatrix}. \tag{9}$$

Now recall the result that, if z and w are two random vectors related by the equation $z = \Delta w$, then their dispersion matrices are related by the equation $D(z) = \Delta D(w)\Delta'$. Applying this result to the equation (9) gives

$$\begin{aligned} \begin{bmatrix} D(y) & C(y,x) \\ C(x,y) & D(x) \end{bmatrix} &= \begin{bmatrix} I & B' \\ 0 & I \end{bmatrix} \begin{bmatrix} D(\varepsilon) & 0 \\ 0 & D(x) \end{bmatrix} \begin{bmatrix} I & 0 \\ B & I \end{bmatrix} \\ &= \begin{bmatrix} D(\varepsilon) + B'D(x)B & B'D(x) \\ D(x)B & D(x) \end{bmatrix}. \end{aligned} \tag{10}$$

The sole restriction in this construction is that the off-diagonal blocks in the dispersion matrix following the first equality are zero-valued, which corresponds to the condition that $C(x, \varepsilon) = 0$.

By relating the submatrices on both sides of this equation, it is found that

$$C(x, y) = D(x)B \tag{11}$$

and that

$$D(y) = B'D(x)B + D(\varepsilon). \tag{12}$$

These are two of the essential relationships associated with the linear regression model of equations (3) and (4). Equation (11) provides an expression for the matrix B of regression coefficients in the form of

$$B = D(x)^{-1}C(x, y). \tag{13}$$

This expression is closely related to a familiar expression from the theory of ordinary least-squares regression. In the usual presentation of the theory, the observations on x and y for $t = 1, \ldots, T$ are accumulated in the matrices X and Y as successions of row vectors, each arrayed below its predecessor. If the observations are in the mean-adjusted form, then the products $T^{-1}X'X$ and $T^{-1}Y'X$ become the empirical counterparts of the moment matrices $D(x)$ and $C(y,x)$ respectively. The estimator of B derived from the principle of the method of moments, which entails ssubstituting the empirical moments for their theoretical counterparts, is $\hat{B} = (X'X)^{-1}X'Y$; and this is also the form of the ordinary least-squares regression estimator.

The Calculus of Conditional Expectations

If we assume that the vectors x and y have a joint normal distribution, then the linear minimum-mean-square error predictor of y, which has been developed in the previous section, can be identified with the ordinary conditional expectation of y given x which is denoted by $E(y|x)$. In that case, we may talk of a calculus of conditional expectations. The essential results of the calculus are as follows:

$$E(y|x) = E(y) + C(y,x)D^{-1}(x)\big\{x - E(x)\big\}, \tag{14}$$

$$D(y|x) = D(y) - C(y,x)D^{-1}(x)C(x,y), \tag{15}$$

$$E\big\{E(y|x)\big\} = E(y), \tag{16}$$

$$D\big\{E(y|x)\big\} = C(y,x)D^{-1}(x)C(x,y), \tag{17}$$

$$D(y) = D(y|x) + D\big\{E(y|x)\big\}, \tag{18}$$

$$C\big\{y - E(y|x), x\big\} = 0. \tag{19}$$

Each of these results corresponds to a result derived in the previous section. However, various changes in notation have occurred. In the first place, we must recognise that the the optimal predictor $\hat{y}$ has becomes the conditional expectation $E(y|x)$. Thus equation (14) can be obtained from equation (2) by replacing $\hat{y}$ by $E(y|x)$ and by substituting for $B = D^{-1}(x)C(x,y)$ from (13).

Next it should be recognised that $D(y|x)$ is simply a synonym for $D(\varepsilon)$, since $\varepsilon = y - E(y|x)$. Then it can be seen that equation (15) is just a restatement of equation (12).

Equation (16) describes the relationship between the conditional and the unconditional expectations. Here the expectation operator which stands outside the braces on the LHS relates to an expectation taken with respect to the variable x. The equation shows how the conditional expectation can be "deconditioned" by taking expectations with respect to the conditioning variable, which is x in this case.

Equation (17) is obtained directly from equation (14) when it is written in the form of

$$E(y|x) - E(y) = C(y,x)D^{-1}(x)\big\{x - E(x)\big\}, \tag{20}$$

Recall, once more, the result that, if z and w are two random vectors related by the equation $z = \Delta w$, then their dispersion matrices are related by the equation $D(z) = \Delta D(w)\Delta'$. Applying this result to the equation above leads to (17) once it is recognised that $D\{x - E(x)\} = D(x)$.

Equation (18) comes from combining equations (15) and (17).

The final equation is a restatement of the condition under (5) which is that the error of prediction, now denoted by $y - E(y|x) = \varepsilon$, must be uncorrelated with the conditioning variable x.