CONDITIONAL EXPECTATIONS AND MINIMUM-MEAN-SQUARE-ERROR PREDICTION

The criterion which is commonly used in judging the performance of an estimator or predictor $\hat{y}$ of a random variable y is its mean-square error defined by $E\{(y - \hat{y})^2\}$. If all of the available information on y is summarised in its marginal distribution, then the minimum mean-square-error prediction is simply the expected value $E(y)$. However, if y is statistically related to another random variable x whose value can be observed, and if the form of the joint distribution of x and y is known, then the minimum mean-square-error prediction of y is the conditional expectation $E(y|x)$. This proposition may be stated this formally:

(1) 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 $E\{(y - \hat{y})^2\} \leq E\{(y - \pi)^2\}$, where $\pi = \pi(x)$ is any other function of x.

Proof. Consider

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

In the second term of the final expression, there is

$$
\begin{aligned}
E\big\{(y-\hat{y})(\hat{y}-\pi)\big\} &= \int_x \int_y (y-\hat{y})(\hat{y}-\pi) f(x,y)\partial y \partial x \\
&= \int_x \bigg\{ \int_y (y-\hat{y}) f(y|x)\partial y \bigg\} (\hat{y}-\pi) f(x)\partial x \\
&= 0.
\end{aligned} \tag{3}
$$

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 $\int (y - \hat{y}) f(y|x)\partial y = E(y|x) - E(y|x) = 0$. Putting (3) into (2) shows that $E\{(y-\pi)^2\} = E\{(y-\hat{y})^2\} + E\{(\hat{y}-\pi)^2\} \geq E\{(y-\hat{y})^2\}$; and the assertion is proved.

The definition of the conditional expectation implies that

$$
\begin{aligned}
E(xy) &= \int_x \int_y xy f(x,y)\partial y \partial x \\
&= \int_x x \bigg\{ \int_y y f(y|x)\partial y \bigg\} f(x)\partial x \\
&= E(x\hat{y}).
\end{aligned} \tag{4}
$$

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

$$E\big\{x(y - \hat{y})\big\} = 0, \tag{5}$$

it takes the form of 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 some part of it would be predictable, which implies that the information of x could be used more efficiently in making the prediction of y.

The proposition of (1) is readily generalised to accommodate the case where, in place of the scalar x, there is a vector $x = [x_1, \ldots, x_p]'$.

Conditional Expectations and Linear Regression

Let x and y be random vectors whose joint distribution is characterised by well-defined first and second-order moments. In particular, let us define the following second-order moments of x and y

$$\begin{aligned} D(x) &= E(xx') - E(x)E(x'), \\ D(y) &= E(yy') - E(y)E(y'), \\ C(y,x) &= E(yx') - E(y)E(x'). \end{aligned} \tag{6}$$

Also, let us postulate that the conditional expectation of y given x is a simple linear function of x:

$$E(y|x) = \alpha + B'x. \tag{7}$$

Then the object is to find expressions for the vector α and the matrix B' which are in terms of the moments listed under (6).

We begin by multiplying $E(y|x)$ by the marginal density function of x and by integrating with respect to x. This converts the conditional expectation into an unconditional expectation. The general result may be expressed by writing

$$E\big\{E(y|x)\big\} = E(y). \tag{8}$$

On applying the latter to equation (7), we find that

$$E(y) = \alpha + B'E(x), \qquad \text{or} \qquad \alpha = E(y) - B'E(x). \tag{9}$$

Next, by multiplying $E(y|x)$ by x' and by the marginal marginal density function of x, and by integrating with respect to x, we obtain the joint moment $E(xy')$. Thus, from equation (7), we get

$$E(yx') = \alpha E(x) + B'E(xx'). \tag{10}$$

But, postmultiplying the first equation under (9) by $E(x')$ gives

$$E(y)E(x') = \alpha E(x') + B'E(x)E(x'), \tag{11}$$

and, when this is subtracted from (10), the result, in view of the definitions under (6), is

$$\begin{aligned} C(y,x) &= E(yx') - E(y)E(x') \\ &= B'\big\{E(xx') - E(x)E(x')\big\} \\ &= B'D(x). \end{aligned} \tag{12}$$

The result from (12) is that

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

This expression for B' and the expression for α under (9) can be substituted into equation (7) to give

$$\begin{aligned} E(y|x) &= \alpha + B'x \\ &= E(y) - B'E(x) + B'x \\ &= E(y) + C(y,x)D^{-1}(x)\big\{x - E(x)\big\}. \end{aligned} \tag{14}$$

In the usual presentation of the theory of the classical regression model, 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 matrices X and Y contain the mean-adjusted observations, 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 is $\hat{B} = (X'X)^{-1}X'Y$.