THE ERRORS IN VARIABLES MODEL AND THE LINEAR REGRESSION MODEL

Imagine that the variables ξ_1, ξ_2 have an exact linear relationship

$$\xi_1\beta_1 + \xi_2\beta_2 = \alpha. \tag{1}$$

Imagine also that, instead of observing these variables, we observe

$$y_1 = \xi_1 + \eta_1 \quad \text{and} \quad y_2 = \xi_2 + \eta_2, \tag{2}$$

where η_1 and η_2 are errors of observation which are distributed independently of each other and of the true values ξ_1 and ξ_2. We shall assume that

$$E(\eta_i) = 0, \quad V(\eta_i) = \omega_{ii} \quad \text{and} \quad C(\eta_i, \eta_j) = \omega_{ij}, \tag{3}$$

where $i, j = 1, 2$.

The equations of (1) and (2) may be combined to give

$$(y_1 - \eta_1)\beta_1 + (y_2 - \eta_2)\beta_2 = \alpha. \tag{4}$$

The object is to find expressions for the parameters α, β_1 and β_2 which are in terms of the variances and covariances of the observations y_1, y_2 and of the errors which afflict them.

We shall begin the search for these estimators by resorting to the method of moments. The approach is similar to one which we have applied to the simple regression model. Later, we shall develop a least-squares estimator. A maximum-likelihood estimator is also available.

Multiplying (4) by y_1 and taking expectations gives

$$\big\{E(y_1^2) - E(y_1\eta_1)\big\}\beta_1 + \big\{E(y_1y_2) - E(y_1\eta_2)\big\}\beta_2 = E(y_1)\alpha. \tag{5}$$

From the assumption that the error η_j and the true value ξ_i are statistically independent, whether or not the subscripts i and j agree, it follows that

$$E(y_i\eta_j) = E\big\{(\xi_i + \eta_i)\eta_j\big\} = E(\eta_i\eta_j) = \omega_{ij}. \tag{6}$$

Therefore (5) can be written as

$$\big\{E(y_1^2) - \omega_{11}\big\}\beta_1 + \big\{E(y_1y_2) - \omega_{12}\big\}\beta_2 = E(y_1)\alpha. \tag{7}$$

Taking expectations in equation (1) gives

$$E(y_1)\beta_1 + E(y_2)\beta_2 = \alpha, \tag{8}$$

and, on multiplying both sides of this by $E(y_1)$, we get

$$\big\{E(y_1)\big\}^2\beta_1 + E(y_1)E(y_2)\beta_2 = E(y_1)\alpha. \tag{9}$$

On taking (9) from (7) we get

$$\big\{V(y_1) - \omega_{11}\big\}\beta_1 + \big\{C(y_1, y_2) - \omega_{12}\big\}\beta_2 = 0, \tag{10}$$

where we have used

$$\begin{aligned} V(y_1) &= E(y_1^2) - \big\{E(y_1)\big\}^2 \quad \text{and} \\ C(y_1, y_2) &= E(y_1y_2) - E(y_1)E(y_2). \end{aligned} \tag{11}$$

By premultiplying equation (4) by y_2 and taking expectations, and by performing the same set of manipulations as before, we can get

$$\big\{C(y_2, y_1) - \omega_{21}\big\}\beta_1 + \big\{V(y_2) - \omega_{22}\big\}\beta_2 = 0. \tag{12}$$

Putting (10) and (12) together gives a system of homogeneous equations:

$$\left\{\begin{bmatrix} V(y_1) & C(y_1, y_2) \\ C(y_2, y_1) & V(y_2) \end{bmatrix} - \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix}\right\}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{13}$$

This pair of equations cannot be solved uniquely for both β_1 and β_2. In other words, the vector $\beta' = [\beta_1, \beta_2]$ is determined only up to a factor of proportionality. Therefore an arbitrary normalisation must be imposed. One possibility is to set $\beta_1^2 + \beta_2^2 = 1$. Another is to set $\beta_1 = -1$ or $\beta_2 = -1$ which is to give one or other of y_1 and y_2 the role of the dependent variable.

Once values for β_1 and β_2 have been obtained, the value of α is given by equation (8).

The foregoing solution depends upon our knowing the precise values of the moments within equation (13). When the moments of y_1 and y_2 are unknown, they may be estimated from a sample of observations $(y_1, y_2)_t; t = 1, \ldots, T$. The estimates are

$$\begin{aligned} s_{11} &= \frac{1}{T}\sum(y_{1t} - \bar{y}_1)^2, \\ s_{22} &= \frac{1}{T}\sum(y_{2t} - \bar{y}_2)^2, \\ s_{21} &= \frac{1}{T}\sum(y_{2t} - \bar{y}_2)(y_{1t} - \bar{y}_1). \end{aligned} \tag{14}$$

The errors are not directly observable; and there is, as yet, no indication of how their moments might be estimated. For the present, we shall assume that these are given in prior knowledge.

When the unknown moments of y_1 and y_2 are replaced by their empirical counterparts, the system will almost certainly become algebraically inconsistent; which means that it can have no solution. To render the system solvable, we must interpolate an additional element λ so as form

$$(15) \qquad \left\{ \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} - \lambda \begin{bmatrix} \omega_{11} & \omega_{12} \\ \omega_{21} & \omega_{22} \end{bmatrix} \right\} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The factor λ should be given the value closest to unity which will reconcile the two equations. This value will converge to unity as the empirical moments converge to the true values.

We shall refer to equation (15) as the errors-in-variables estimator.

To see how λ may be determined, let us assume, for the sake of simplicity, that the two errors η_1, η_2 are uncorrelated, so that $\omega_{12} = \omega_{21} = 0$, and that they have equal variance, so that $\omega_{11} = \omega_{22}$. Then the value of the common variance need not be specified, since it may be absorbed in the value of λ. The resulting equation system is

$$(16) \qquad \left\{ \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right\} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The requirement that the equations should be mutually consistent is equivalent to the condition that

$$(17) \qquad \begin{aligned} 0 &= \mathrm{Det} \begin{bmatrix} s_{11} - \lambda & s_{12} \\ s_{21} & s_{22} - \lambda \end{bmatrix} \\ &= \lambda^2 - \lambda(s_{11} + s_{22}) + (s_{11}s_{22} - s_{12}s_{21}). \end{aligned}$$

Therefore λ is found as the solution to a quadratic equation.

Once the estimates for β_1 and β_2 have been determined, the estimate for α may be obtained from the empirical counterpart of equation (8):

$$(18) \qquad \bar{y}_1\hat{\beta}_1 + \bar{y}_2\hat{\beta}_2 = \hat{\alpha}.$$

Ordinary Least-Squares Regression as a Limiting Case.

Imagine that the variance of the error η_1 is tending to zero. In that case, the covariance of η_1 and η_2 must also be tending to zero. With a change of notation and with a particular normalisation of the parameter vector, the limiting form of equation (15) can be written as

$$\left\{\begin{bmatrix} s_{xx} & s_{xy} \\ s_{yx} & s_{yy} \end{bmatrix} - \lambda \begin{bmatrix} 0 & 0 \\ 0 & \sigma^2 \end{bmatrix}\right\} \begin{bmatrix} \beta \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \tag{23}$$

where

$$\begin{aligned} s_{xx} &= \frac{1}{T}\sum(x_t - \bar{x})^2, \\ s_{yy} &= \frac{1}{T}\sum(y_t - \bar{y})^2, \\ s_{xy} &= \frac{1}{T}\sum(x_t - \bar{x})(y_t - \bar{y}). \end{aligned} \tag{24}$$

On solving the first equation $s_{xx}\beta - s_{xy} = 0$, we find that

$$\hat{\beta} = \frac{\sum(x_t - \bar{x})(y_t - \bar{y})}{(x_t - \bar{x})^2}, \tag{25}$$

which is nothing but the ordinary least-squares estimator of the regression parameter in the equation $E(y|x) - x\beta = \alpha$.

In solving the second equation $s_{yy} - s_{yx}\beta = \lambda\sigma^2$, we are faced with two unknowns, λ and σ^2. If we set $\lambda = 1$, then the solution for σ^2 is

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{T}\sum(y_t - \bar{y})^2 - \frac{1}{T}\sum(y_t - \bar{y})(x_t - \bar{x})\hat{\beta} \\ &= \frac{1}{T}\sum(y_t - \bar{y})^2 - \frac{1}{T}\sum(x_t - \bar{x})^2\hat{\beta}^2. \end{aligned} \tag{26}$$

It is straightforward to demonstrate that this formula is equivalent to the formula

$$\hat{\sigma}^2 = \frac{1}{T}\sum(y_t - \hat{\alpha} - x_t\hat{\beta})^2, \qquad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}, \tag{27}$$