Conditional Expectations and Classical Regression Theory

Imagine that $z' = [y', x']$ is a vector of variables which are normally distributed; which means that their distribution is completely characterised by the first-order and second-order moments of the distribution. Then it is always possible to construct a transformation of the form

$$(1) \qquad [\, y' - E(y') \quad x' - E(x')\,] \begin{bmatrix} I & 0 \\ -B & I \end{bmatrix} = [\, \varepsilon' \quad x' - E(x')\,],$$

such that ε, which has $E(\varepsilon) = 0$, and x are statistically independent. The corresponding relationship between the second-order moments is

$$(2) \qquad \begin{bmatrix} I & -B' \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} I & 0 \\ -B & I \end{bmatrix} = \begin{bmatrix} \Sigma_{\varepsilon\varepsilon} & 0 \\ 0 & \Sigma_{xx} \end{bmatrix}.$$

The relationship may be recast in the form of

$$(3) \qquad \begin{bmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{bmatrix} \begin{bmatrix} I & 0 \\ -B & I \end{bmatrix} = \begin{bmatrix} I & B' \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{\varepsilon\varepsilon} & 0 \\ 0 & \Sigma_{xx} \end{bmatrix},$$

from which

$$(4) \qquad \begin{aligned} \Sigma_{yy} - \Sigma_{yx}B &= \Sigma_{\varepsilon\varepsilon}, \\ \Sigma_{xy} - \Sigma_{xx}B &= 0, \end{aligned}$$

which lead to the equations

$$(5) \qquad \begin{aligned} B &= \Sigma_{xx}^{-1}\Sigma_{xy}, \\ \Sigma_{\varepsilon\varepsilon} &= \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}. \end{aligned}$$

The relationships implicit in this structure may be expressed in terms of the expectations operator etc. We recognise $\varepsilon = y - E(y|x)$ as the error from predicting y on the basis of the information of x. We define

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

Then

$$(7) \qquad E(y|x) = E(y) + C(y,x)D^{-1}(x)\big\{x - E(x)\big\},$$

$$(8) \qquad D(y|x) = D(y) - C(y,x)D^{-1}(x)C(x,y),$$

$$(9) \qquad E\big\{E(y|x)\big\} = E(y),$$

$$(10) \qquad D\big\{E(y|x)\big\} = C(y,x)D^{-1}(x)C(x,y),$$

$$(11) \qquad D(y) = D(y|x) + D\big\{E(y|x)\big\},$$

$$(12) \qquad C\big\{y - E(y|x), x\big\} = 0.$$

Recursive Least-Squares Estimation

The tth instance of the regression relationship is represented by

$$y_t = x_t'\beta + \varepsilon_t. \tag{13}$$

It is assumed that the disturbances ε_t are serially independent with

$$E(\varepsilon_t) = 0 \quad \text{and} \quad V(\varepsilon_t) = \sigma^2 \quad \text{for all} \quad t. \tag{14}$$

To initiate the recursion, we attribute a Bayesian prior distribution to β with

$$b_0 = E(\beta) \qquad \text{and} \qquad P_0 = D(\beta).$$

The empirical information at time t is the set of observations $\mathcal{I}_t = \{y_1, \ldots, y_t\}$.

We estimate $b_t = E(\beta|\mathcal{I}_t)$ and $P_t = D(\beta|\mathcal{I}_t)$ making use of the previous estimates $b_{t-1} = E(\beta|\mathcal{I}_{t-1})$ and $P_{t-1} = D(\beta|\mathcal{I}_{t-1})$. From (7) it follows that

$$E(\beta|\mathcal{I}_t) = E(\beta|\mathcal{I}_{t-1}) + C(\beta, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})\big\{y_t - E(y_t|\mathcal{I}_{t-1})\big\}. \tag{15}$$

There are three elements on the RHS. The first is the term is the error from predicting y_t at time $t-1$:

$$\begin{aligned} y_t - E(y_t|\mathcal{I}_{t-1}) &= y_t - x_t'b_{t-1} \\ &= h_t. \end{aligned} \tag{16}$$

Next is the dispersion matrix of associated with this prediction:

$$\begin{aligned} D(y_t|\mathcal{I}_{t-1}) &= D\big\{x_t'(\beta - b_{t|t-1})\big\} + D(\varepsilon_t) \\ &= x_t'P_{t-1}x_t + \sigma^2 = D(h_t). \end{aligned} \tag{17}$$

Finally there is the covariance

$$\begin{aligned} C(\beta, y_t|\mathcal{I}_{t-1}) &= E\big\{(\beta - b_{t-1})y_t'\big\} \\ &= E\big\{(\beta - b_{t-1})(x_t'\beta + \varepsilon_t)'\big\} \\ &= P_{t-1}x_t. \end{aligned} \tag{18}$$

On putting these elements together, we get

$$b_t = b_{t-1} + P_{t-1}x_t(x_t'P_{t-1}x_t + \sigma^2)^{-1}(y_t - x_t'b_{t-1}). \tag{19}$$

Equation (8) indicates that

$$D(\beta|\mathcal{I}_t) = D(\beta|\mathcal{I}_{t-1}) - C(\beta, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})C(y_t, \beta|\mathcal{I}_{t-1}). \tag{20}$$

It follows from (17) and (18) that this is

$$P_t = P_{t-1} - P_{t-1}x_t(x_t'P_{t-1}x_t + \sigma^2)^{-1}x_t'P_{t-1}. \tag{21}$$

We may anatomise the components of the recursive least-squares algorithm:

$$h_t = y_t - x_t' b_{t-1}, \qquad \textit{Prediction Error} \tag{22}$$

$$f_t = x_t' P_{t-1} x_t + \sigma^2, \qquad \textit{Error Dispersion} \tag{23}$$

$$\kappa_t = P_{t-1} x_t f_t^{-1}, \qquad \textit{Filter Gain} \tag{24}$$

$$b_t = b_{t-1} + \kappa_t h_t, \qquad \textit{Parameter Estimate} \tag{25}$$

$$P_t = (I - \kappa_t x_t) P_{t-1}. \qquad \textit{Estimate Dispersion} \tag{26}$$

Alternative expressions are available for P_t and κ_t:

$$P_t = (P_{t-1}^{-1} + \sigma^{-2} x_t x_t')^{-1}, \tag{27}$$

$$\kappa_t = \sigma^{-2} P_t x_t. \tag{28}$$

The expression of (27) comes via the matrix inversion formula

$$(B + CDC')^{-1} = B^{-1} - B^{-1}C(C'B^{-1}C + D^{-1})^{-1}C'B^{-1}. \tag{29}$$

Equation (27) indicates that

$$\sigma^2 P_t^{-1} = \sigma^2 P_0^{-1} + \sum_{i=1}^{t} x_i x_i'. \tag{30}$$

Apart from the matrix $\sigma^2 P_0^{-1}$, which becomes relatively insignificant for large values of t, this is just the familiar moment matrix of ordinary least-squares regression.

When equations (27) and (28) are used in (25), we get the following expression for recursive least-squares estimate:

$$b_t = b_{t-1} + \sigma^{-2}(P_{t-1}^{-1} + \sigma^{-2} x_t x_t')^{-1} x_t (y_t - x_t' b_{t-1}). \tag{31}$$

The equation serves to show that σ^2, which is a factor of P_t, can be cancelled from the formula for b_t.

The formula of (31) is computationally inefficient compared with the formula of (19). The latter entails finding the inverse of the scalar element $f_t = x_t P_{t-1} x_t' + \sigma^2$. The formula under (30) involves the inversion of the entire matrix P_t.

Extensions of the Recursive Least-Squares Algorithm

It is easy to extend the algorithm to produce a rolling regression. The additional task is to remove the data which was acquired at time $t-n$. The first step is to adjust the moment matrix to give $\sigma^2 P_t^{*-1} = \sigma^2 P_{t-1}^{-1} - x_{t-n}x'_{t-n}$. The matrix inversion formula of (29) indicates that

$$
\begin{aligned}
P_t^* &= (P_{t-1}^{-1} - \sigma^{-2}x_{t-n}x'_{t-n})^{-1} \\
&= P_{t-1} - P_{t-1}x_{t-n}(x'_{t-n}P_{t-1}x_{t-n} - \sigma^2)^{-1}x'_{t-n}P_{t-1},
\end{aligned}
\tag{32}
$$

Next, an intermediate estimate b_t^*, which is based upon the reduced information, is obtained from b_{t-1} via the formula

$$
\begin{aligned}
b_t^* &= b_{t-1} - \sigma^{-2}P_t^{*-1}x_{t-n}(y_{t-n} - x'_{t-n}b_{t-1}) \\
&= b_{t-1} - P_{t-1}x_{t-n}(x'_{t-n}P_{t-1}x_{t-n} - \sigma^2)^{-1}(y_{t-n} - x'_{t-n}b_{t-1}).
\end{aligned}
\tag{33}
$$

This formula can be understood by considering the inverse problem of obtaining b_{t-1} from b_t^* by the *addition* of the information from time $t-n$. The estimate b_t, which is based on the n data points $x_t, \ldots, x_{t-n+1}$, is obtained from the formula under (19) by replacing b_{t-1} with b_t^* and P_{t-1} with P_t^*.

A gradual discounting of old data may be more appropriate than discarding observations which have passed a date of expiry. Let $\lambda \in (0, 1]$ be the rate at which the data is discounted. Then, in place of P_t of (27), we should have

$$
\begin{aligned}
P_t &= (\lambda P_{t-1}^{-1} + \sigma^{-2}x_t x'_t)^{-1} \\
&= \frac{1}{\lambda}\Big\{P_{t-1} - P_{t-1}x_t(x'_t P_{t-1}x_t + \lambda\sigma^2)^{-1}x_t P_{t-1}\Big\}.
\end{aligned}
\tag{34}
$$

The formula for the parameter estimate would be

$$
b_t = b_{t-1} + P_{t-1}x_t(x'_t P_{t-1}x_t + \lambda\sigma^2)^{-1}(y - x'_t b_{t-1}). \tag{35}
$$

Numerous pragmatic techniques are available for shaping the memory of the recursive least-square algorithm. The fully-fledged Kalman filter provides a firmer theoretical basis.

The Kalman filter incorporates an additional process which describes the variation of the parameter vector β in terms of a Markov scheme:

$$
\beta_t = \Phi\beta_{t-1} + \nu_t, \tag{36}
$$

Equations of the Kalman Filter

The state-space model, which underlies the Kalman filter, consists of two equations

$$y_t = H_t\xi_t + \eta_t, \qquad \textit{Observation Equation} \tag{37}$$

$$\xi_t = \Phi_t\xi_{t-1} + \nu_t, \qquad \textit{Transition Equation} \tag{38}$$

where y_t is the observation on the system and ξ_t is the state vector. The observation error η_t and the state disturbance ν_t are mutually uncorrelated random vectors of zero mean with dispersion matrices

$$D(\eta_t) = \Omega_t \qquad \text{and} \qquad D(\nu_t) = \Psi_t. \tag{39}$$

It is assumed that the matrices H_t, Φ_t, Ω_t and Ψ_t are known for all $t = 1, \ldots, n$ and that an initial estimate x_0 is available for the state vector ξ_0 at time $t = 0$ together with a dispersion matrix $D(\xi_0) = P_0$.

The Kalman-filter equations determine the state-vector estimates $x_{t|t-1} = E(\xi_t|\mathcal{I}_{t-1})$ and $x_t = E(\xi_t|\mathcal{I}_t)$ and their associated dispersion matrices $P_{t|t-1}$ and P_t. From $x_{t|t-1}$, the prediction $\hat{y}_{t|t-1} = H_t x_{t|t-1}$ is formed which has a dispersion matrix F_t. Th equations are

$$x_{t|t-1} = \Phi_t x_{t-1}, \qquad \textit{State Prediction} \tag{40}$$

$$P_{t|t-1} = \Phi_t P_{t-1}\Phi_t' + \Psi_t, \qquad \textit{Prediction Dispersion} \tag{41}$$

$$e_t = y_t - H_t x_{t|t-1}, \qquad \textit{Prediction Error} \tag{42}$$

$$F_t = H_t P_{t|t-1} H_t' + \Omega_t, \qquad \textit{Error Dispersion} \tag{43}$$

$$K_t = P_{t|t-1} H_t' F_t^{-1}, \qquad \textit{Kalman Gain} \tag{44}$$

$$x_t = x_{t|t-1} + K_t e_t, \qquad \textit{State Estimate} \tag{45}$$

$$P_t = (I - K_t H_t) P_{t|t-1}. \qquad \textit{Estimate Dispersion} \tag{46}$$

In comparison with the equations of the recursive regression algorithm listed under (22)–(19), there are two additions: equation (40) for the state prediction and equation (41) for its dispersion. These owe their existence to the presence of the transition equation (38); and they vanish when $\Phi = I$ and $\nu_t = 0$.