RECURSIVE REGRESSION ESTIMATION

Conditional Ecxpectations. 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

$$\begin{aligned} D(x) &= E(xx') - E(x)E(x'),\\ D(y) &= E(yy') - E(y)E(y'),\\ C(y,x) &= E(yx') - E(y)E(x'). \end{aligned} \tag{1}$$

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

$$E(y|x) = \alpha + B'x. \tag{2}$$

Then the object is to find expressions for the vector α and the matrix B which are in terms of the moments listed under (1).

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

$$E\big\{E(y|x)\big\} = E(y). \tag{3}$$

On applying the latter to equation (2), we find that

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

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 (2), we get

$$E(yx') = \alpha E(x) + B'E(xx'). \tag{5}$$

But, postmultiplying the first equation under (4) by $E(x')$ gives

$$E(y)E(x') = \alpha E(x') + B'E(x)E(x'), \tag{6}$$

and, when this is subtracted from (5), the result, in view of the definitions under (1), is

$$\begin{aligned} C(y,x) &= E(yx') - E(y)E(x')\\ &= B'\big\{E(xx') - E(x)E(x')\big\}\\ &= B'D(x). \end{aligned} \tag{7}$$

The result from (7) is that

$$B = D^{-1}(x)C(x,y). \tag{8}$$

This expression for B and the expression for α under (4) can be substituted into equation (2) to give

$$
\begin{aligned}
E(y|x) &= \alpha + B'x \\
&= E(y) - B'E(x) + Bx \\
&= E(y) + C(y,x)D^{-1}(x)\{x - E(x)\}.
\end{aligned} \tag{9}
$$

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 $T^{-1}X'Y$ become the empirical counterparts of the moment matrices $D(x)$ and $C(x,y)$ respectively. The estimator of B derived from the principle of the method of moments is $\hat{B} = (X'X)^{-1}X'Y$.

Several additional results in the algebra of conditional expectations which we shall invoke in the next section can also be derived with ease. To avoid burdening this account with unnecessary developments, let us simply declare in summary that, if x, y are jointly distributed variables which bear the linear relationship $E(y|x) = \alpha + B'x$, then

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

$$D(y|x) = D(y) - C(y,x)D^{-1}(x)C(x,y), \tag{11}$$

$$E\{E(y|x)\} = E(y), \tag{12}$$

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

$$D(y) = D(y|x) + D\{E(y|x)\}, \tag{14}$$

$$C\{y - E(y|x), x\} = 0. \tag{15}$$

Recursive Least-Square Regression. Consider the equation of the linear regression model. The tth instance of the regression relationship is represented by

$$y_t = x_t'\beta + \varepsilon_t. \tag{16}$$

Here y_t is a scalar element and x_t' is a row vector. It is assumed that the disturbances ε_t are serially independent with

$$E(\varepsilon_t) = 0 \quad \text{and} \quad V(\varepsilon_t) = \sigma^2 \quad \text{for all} \quad t. \tag{17}$$

We may regard β as a random variable and we attribute to it a prior distribution with

$$E(\beta) = b_0 \qquad \text{and} \qquad D(\beta) = P_0. \tag{18}$$

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

The Bayesian Derivation. The object is to derive the estimates $b_t = E(\beta|\mathcal{I}_t)$ and $P_t = D(\beta|\mathcal{I}_t)$ from the information available at time t in a manner which makes best use of $b_{t-1} = E(\beta|\mathcal{I}_{t-1})$ and $P_{t-1} = D(\beta|\mathcal{I}_{t-1})$ which are the previous estimates. The essential task is to evaluate the expression

$$E(\beta|\mathcal{I}_t) = E(\beta|\mathcal{I}_{t-1}) + C(\beta, y_t|\mathcal{I}_{t-1})D^{-1}(y_t|\mathcal{I}_{t-1})\{y_t - E(y_t|\mathcal{I}_{t-1})\}, \tag{19}$$

which is derived directly from (39). There are three elements on the RHS which need to be evaluated. The first is the term

$$\begin{aligned} y_t - E(y_t|\mathcal{I}_{t-1}) &= y_t - x_t' b_{t-1} \\ &= h_t. \end{aligned} \tag{20}$$

This is the error from predicting y_t from the information available at time $t-1$. Next is the dispersion matrix of associated with this prediction. This is

$$\begin{aligned} D(y_t|\mathcal{I}_{t-1}) &= D\{x_t'(\beta - b_{t|t-1})\} + D(\varepsilon_t) \\ &= x_t' P_{t-1} x_t + \sigma^2 = D(h_t). \end{aligned} \tag{21}$$

Finally there is the covariance

$$\begin{aligned} C(\beta, y_t|\mathcal{I}_{t-1}) &= E\{(\beta - b_{t-1})y_t'\} \\ &= E\{(\beta - b_{t-1})(x_t'\beta + \varepsilon_t)'\} \\ &= P_{t-1}x_t. \end{aligned} \tag{22}$$

On putting these elements together, we get

$$b_t = b_{t-1} + P_{t-1}x_t(x_t'P_{t-1}x_t + \sigma^2)^{-1}(y_t - x_t'b_{t-1}). \tag{23}$$

There must also be a means of deriving the dispersion matrix $D(\beta|\mathcal{I}_t) = P_t$ from its predecessor $D(\beta|\mathcal{I}_{t-1}) = P_{t-1}$. Equation (13) indicates that

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

It follows from (20) and (21) that this is

$$P_t = P_{t-1} - P_{t-1}x_t(x_t'P_{t-1}x_t + \sigma^2)^{-1}x_t'P_{t-1}. \tag{24}$$

Extensions of the Recursive Least-Squares Algorithm

The algorithm which we have presented in the previous section represents little more than an alternative means of computing the ordinary least-squares regression estimates. If the parameters of the underlying process which generates the data are stable, then we can expect the estimate b_t to converge also to a stable value as the number of observations t increases At the same time, the elements of the dispersion matrix P_t will decrease in value.

A further consequence of the growth of the number of observations is that the filter gain κ_t will diminish at t increases. This implies that the impact of successive prediction errors upon the estimate of β will diminish as the amount of information already incorporated in the estimate increases.

If there is doubt about the constancy of the regression parameter, then it may be desirable to give greater weight to the more recent data; and it might even be appropriate to discard data which has reached a certain age and has passed its date of expiry.

One way of accommodating parametric variability is to base the estimate on only the most recent portion of the data. As each new observation is acquired another observation may be removed so that, at any instant, the estimator comprises only n points. Such an estimator has been described as a rolling regression. Implementations are available in the recent versions of the more popular econometric computer packages such as *Microfit 3.0* and *PCGive*.

It is a simple matter to extend the algorithm of the previous section 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 indicates that

$$\begin{aligned} P_t^* &= (P_{t-1}^{-1} - \sigma^{-2}x_{t-n}x_{t-n}')^{-1} \\ &= P_{t-1} - P_{t-1}x_{t-n}(x_{t-n}'P_{t-1}x_{t-n} - \sigma^2)^{-1}x_{t-n}'P_{t-1}, \end{aligned} \tag{25}$$

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

$$\begin{aligned} b_t^* &= b_{t-1} - \sigma^{-2}P_t^{*-1}x_{t-n}(y_{t-n} - x_{t-n}'b_{t-1}) \\ &= b_{t-1} - P_{t-1}x_{t-n}(x_{t-n}'P_{t-1}x_{t-n} - \sigma^2)^{-1}(y_{t-n} - x_{t-n}'b_{t-1}). \end{aligned} \tag{26}$$

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 $t-n$. A rearrangement of the resulting expression for b_{t-1} gives the initial expression for b_t^* under (26). Finally, 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 (23) by replacing b_{t-1} with b_t^* and P_{t-1} with P_t^*.

Discarding observations which have passed a date of expiry is an appropriate procedure when the processes generating the data are liable, from time to time, to undergo sudden structural changes. For it ensures that any misinformation which is conveyed by the data which predate the structural change will not be kept on record permanently. However, if the processes are expected to change gradually in a more or less systematic fashion, then a gradual discounting of old data may be more appropriate. An exponential weighting scheme applied to the data might serve this purpose.

Let the rate at which the data is discounted be given by a parameter $\lambda \in (0, 1]$. Then, in place of the expression for P_t under (24), we should have

$$\begin{aligned} P_t &= (\lambda P_{t-1}^{-1} + \sigma^{-2}x_t x_t')^{-1} \\ &= \frac{1}{\lambda}\Big\{P_{t-1} - P_{t-1}x_t(x_t'P_{t-1}x_t + \lambda\sigma^2)^{-1}x_t P_{t-1}\Big\}. \end{aligned} \tag{27}$$

The formula for the parameter estimate would be

$$b_t = b_{t-1} + P_{t-1}x_t(x_t'P_{t-1}x_t + \lambda\sigma^2)^{-1}(y - x_t'b_{t-1}). \tag{28}$$

It is curious that econometric packages mentioned above have implemented rolling regression but not exponentially-weighted regression.

A wide variety of techniques for shaping the memory of the recursive least-square algorithm may be devised. However, it is clear that such formulations are essentially pragmatic, and one might wish for a theoretical basis from which to develop the algorithms. The basis is provided by the fully-fledged Kalman filter.

The elaboration of the recursive least-square model which is required in order to achieve the generality of the Kalman filter is the addition of a process which describes the variation of the parameter vector β. Such a process might be described by the equation

$$\beta_t = \Phi\beta_{t-1} + \nu_t, \tag{29}$$

which represents a Markov scheme.