Computing (co)variances recursively

Thijs Knaap, CPB

October 2010

For a set of numbers $\{x_t\}_{t=1,\dots,n}$, the sample mean is computed as

$$M_x(n) = \frac{1}{n} \sum_{t=1}^{n} x_t$$

and the sample variance as

$$V_x(n) = \frac{1}{n} \sum_{t=1}^{n} x_t^2 - [M_x(n)]^2 .$$

The covariance between two (ordered) sets of numbers $\{x_t, y_t\}_{t=1,\dots,n}$ is likewise computed as

$$C_{x,y}(n) = \frac{1}{n} \sum_{t=1}^{n} x_t y_t - M_x(n) M_y(n).$$

Computing these statistics requires the entire set of numbers. In some situations, we would like to recompute a statistic when a new observation (x_{n+1}, y_{n+1}) becomes available. Rather than starting from scratch, it would nice if we could use just the old statistic, the new observation and the total number of observations.

It is obvious that such is possible for the mean using the expression

$$M_x(n+1) = \frac{n}{n+1} M_x(n) + \frac{1}{n+1} x_{n+1}.$$

But can it also be done for the variance and the covariance? The answer is yes, and the formulas are as follows:

$$V_x(n+1) = \frac{n}{n+1} V_x(n) + \frac{1}{n} \left[x_{n+1} - M_x(n+1) \right]^2$$

and, more generally,

$$C_{x,y}(n+1) = \frac{n}{n+1} C_{x,y}(n) + \frac{1}{n} \left[x_{n+1} - M_x(n+1) \right] \left[y_{n+1} - M_y(n+1) \right].$$

Note the rather counterintuitive $1/n$ in front of the second term of these expressions. Using this rather than $1/(n+1)$ corrects the effect of the changing mean on the previous statistic.

Derivation

I derive the expression for the covariance, of which the variance is a special case. Compare the two statistics,

$$C_{x,y}(n) = \frac{1}{n} \sum_{t=1}^{n} x_t y_t - M_x(n) M_y(n),$$

$$C_{x,y}(n+1) = \frac{1}{n+1} \sum_{t=1}^{n+1} x_t y_t - M_x(n+1) M_y(n+1).$$

We can write the second as a function of the first by multiplying and adding so that equality is preserved.

$$C_{x,y}(n+1) = \frac{n}{n+1} C_{x,y}(n) + \frac{x_{n+1} y_{n+1}}{n+1}$$
$$- \left(M_x(n+1) M_y(n+1) - \frac{n}{n+1} M_x(n) M_y(n) \right) \quad (1)$$

If we knew the true sample mean, rather than having to estimate it with M, the expression would be a lot simpler. What the part in parentheses on the second line does, is allow for the changing estimate of the mean. Since we already have a recursive expression for the mean, we could stop here as this expression is fully operational. But it turns out that some manipulations give us a much friendlier version.

Develop the last term in brackets in the previous equation as

$$\frac{n}{n+1} M_x(n) M_y(n) = \frac{n}{n+1} \left(\frac{n+1}{n} \left(M_x(n+1) - \frac{1}{n+1} x_{n+1} \right) \right)$$
$$\cdot \left(\frac{n+1}{n} \left(M_y(n+1) - \frac{1}{n+1} y_{n+1} \right) \right)$$
$$= \frac{n+1}{n} \left(M_x(n+1) M_y(n+1) \right.$$
$$- \frac{1}{n+1} \left[y_{n+1} M_x(n+1) + x_{n+1} M_y(n+1) \right]$$
$$\left. + \frac{1}{(n+1)^2} x_{n+1} y_{n+1} \right)$$

Put this back in equation (1) to find

$$C_{x,y}(n+1) = \frac{n}{n+1} C_{x,y}(n) + \left(\frac{1}{n+1} + \frac{1}{n(n+1)} \right) x_{n+1} y_{n+1} +$$
$$\frac{1}{n} M_x(n+1) M_y(n+1) - \frac{1}{n} \left[y_{n+1} M_x(n+1) + x_{n+1} M_y(n+1) \right].$$

Rewriting the first term in large parentheses as $1/n$, this expression can be simplified as

$$C_{x,y}(n+1) = \frac{n}{n+1} C_{x,y}(n) + \frac{1}{n} \left[x_{n+1} - M_x(n+1) \right] \left[y_{n+1} - M_y(n+1) \right]$$

which gives the result. Writing $y_t = x_t$ then gives the expression for the variance.