THE ECONOMIC RESEARCH ... MEMORANDUM SERIES
NO. 39

The Effect on R^2 of Adding an Independent Variate in Multivariate Regression

Let the (k+1)th independent variate be x_t, the k other variates being x_{ti}, $t=1,2,\ldots, T$, $i=1,2,\ldots,k$. Without loss of generality the x_{ti} are assumed to be orthogonal. Also w.l.g. the dependent variate y_t and all the independents are assumed to be standardized, i.e. their means are zero and their sum squares equal T. Let R^2_k and R^2_{k+1} be the values for the k and (k+1) regressions respectively. The full (k+1) regression is then

$$(1)\qquad y_{tc} = cx_t + \sum_{i=1}^{k} x_{ti}\, b_i$$

Let

$$(2)\qquad \frac{1}{T}\sum_t y_t x_t = p; \quad \frac{1}{T}\sum_t y_t x_{ti} = q_i; \quad \frac{1}{T}\sum_t x_t x_{ti} = r_i, \quad i=1,2,\ldots,k$$

Obviously p and the q_i and t_i are correlation coefficients. They are, in fact, all the simple c.c.'s in the system since, by hypothesis, $\Sigma x_{ti}\, x_{tj}$ $(i,j= 1,2,.,k,\ i\neq j)$ is zero. Then

$$(3)\qquad R^2_k = \sum_{i=1}^{k} q_i^2$$

From (1) and (2),

$$(4)\qquad R^2_{k+1} = c^2 + 2c\,\Sigma\, b_i r_i + \Sigma\, b_i^2,$$

the normal equations being

$$(5)\qquad p = c + \Sigma\, r_i\, b_i$$

$$q_i = r_i c + b_i,\ i = 1,2,..,k.$$

On substitution for i and the b_i from (5) into (4), and on reduction,

$$(6)\qquad R^2_{k+1} = [(p - \Sigma\, q_i r_i)^2 + (1 - \Sigma\, r_i^2)\,\Sigma q_i^2\,]/(1- \Sigma r_i^2),$$

or, using (3),

$$(7)\qquad R^2_{k+1} = R^2_k + (p - \Sigma\, q_i r_i)^2/(1 - \Sigma\, r_i^2)$$

The second term on the right of (7) has all the look of the square of a partial c.c. This turns out to be nearly the case. In fact, let

$$(8)\qquad z_t = \Sigma_i x_{ti} r_i.$$

After some simple reduction we find

(9) $R^2_{k+1} = R^2_k + (1-r^2_{yz})\ r^2(x,y/z)$

with (see (2))

(10) $r_{xy} = p;\quad r_{xz} = \sqrt{\Sigma r_i^2};\quad r_{yz} = \Sigma q_i r_i / \sqrt{\Sigma r_i^2}.$

of course, $R^2_{k+1} \geq R^2_k$.

No novelty is claimed for (7) or (9). The object of this note is to show a particular aspect of the fundamental role of orthogonization of independent variates in multivariate analysis, to say nothing of the simplification in exposition of standardization of all variates. Of course, a result like (7) can be produced using the raw data, but the apparent simplicity in that case of the second term on the right is spurious as consisting of a complicated matrix expression.

5 April 1967 R. C. Geary