Selection of Independent Variables in Multivariate Regression

by

R. C. Geary

A few textbooks consulted are by no means clear on recommended practice in the following, very common , situation. A multivariate regression is set up; R^2 is calculated and the equation is found to be significant by $F(k, T-k-1)$[1], where k is number of independents and T number of sets of observations; invariably some coefficients are found to be significantly different from zero at, say, the 5% probability level, some not. The problem is: in estimating the expected value y_c of the dependent variable for all T sets of observations, should the insignificant variables be (i) included or (ii) excluded? Or, in other words, should the coefficients whose values are not significant be set at their calculated values or at zero in the formula for y_c? As we are uncertain we might compromise with a policy (iii) include coefficients with t's exceeding a particular value (say 1 or 1.5) and assume that the other coefficients are zero. The justification for policies (i) or (iii) is that if T were larger than it is the insignificant values might turn out to be significant, so we try to play it safe, There is, however, an undesirable element of arbitrariness in policy (iii), which nevertheless we instinctively favour. The object of the present note is to clarify the problem: a clear-cut answer still eludes us.

As favouring policies (i) or (iii) we have lately come on a regression with T=18, k=6, in which R^2=.88 but in which none of the coefficient t_i's were significant at the 5% probability level. The F corresponding to R^2=.88 is F=13.44 (=11x.88/6x.12) which with d.f.'s $f_1 = 6$, $f_2 = 11$ is highly significant since the 0.5% probability of F is only 6.10. We conclude that the whole

[1] See, e.g., Handbook of Statistical Tables by Harold B. Owen (Addison Wesley Publishing Company, Inc. 1962), page 514

equation is highly significant and we infer that some of the coefficients must be regarded as significant despite the showing of the t_i's (but not significant at the 5% probability level). It happens that for independent variables numbered 2 and 6 the t_i's are near 2 the remaining four variables having quite small t_i's; we select variables 2 and 6 and reject the rest. The R^2 for the two-variable regression is .87, sufficiently high to justify our choice, which falls into category (iii). The t for variable 2 is now >5 and the t for variable 6 > 3. Because of degrees of freedom considerations an R^2 of .87 for 2 independent variables is overwhelmingly more significant than an R^2 of .88 for 6 independent variables.

In cases like this it therefore seems advisable to purge the regression of variables with the smallest t_i's. In this example the inconsistency between the t_i's and R^2 has vanished. The remaining coefficients are emphatically more meaningful. And, of course, the fewer variables in a regression the better.

The problem could, of course, be regarded as identical with that of selecting <u>ex post</u> significant regressors from a larger set, a problem which, so far, has proved intractable in the general case; it can be solved (by order statistics) only when the large set of regressors is orthogonal.

We have every interest in trying to reduce the number of regressors to a minimum . Let the model be in the original variables,

$$(1) \quad y_t = \sum_{i=0}^{k} \beta_i' x_{it}' + u_t, \quad t = 1,2,\ldots, T,$$

with constant β'_o and $x_{ot} = 1$. It is known that the Σ term can be transformed (using as transformer an orthogonal k + 1 x k + 1 matrix H, the same for all t) into

$$(2) \quad y_t = \sum_{i=0}^{k} \beta_i x_{it} + u_t,$$

so that the k + 1 independents are now orthogonal, i.e.

$$\Sigma_t x_{it} = 0, \quad i = 1,2,\ldots, k$$

$$(3) \quad \Sigma_t x_{it} x_{i't} \; (i' \neq i) = 0, \quad i, i' = 0,1,2,\ldots k.$$

Let the least squares regression of (2) be

(4) $y_{tc} = \Sigma b_i x_{it}$.

Also let

(5) $\eta_t = \Sigma \beta_i x_{it}$

The deviation of the realised y_{tc} from the (unknown) η_t and towards which y_{tc} tends in probability for each t is

(6) $y_{tc} - \eta_t = \sum_{i=0}^{k} (b_i - \beta_i) x_{it}$

with

(7) $b_i - \beta_i = \sum_{t=1}^{T} x_{it} u_t / \Sigma x_{it}^2$, $i = 0, 1, \ldots, k$

Using (3), $Eu_t u_{t'} = 0$ and $Eu_t^2 = \sigma^2$, it can easily be shown that

(8) $E \sum_{t=1}^{T} (y_{tc} - \eta_t)^2 = (k + 1) \sigma^2$.

It is reasonable to regard the l.s. of (8) as the natural measure of the efficiency of the regression: the lower its value the higher the efficiency. The r.s. of (8) shows that we must try to make k, the number of independents, as low as possible. Incidentally, the LS regression b_i is the best linear estimator, by the test of minimizing r.s. of (8), of the class $\Sigma_t c_{it} y_t$.

To take an absolute position, suppose that in the model (1) k_2 of the β_i were zero and k_1 not zero so that $k = k_1 + k_2$. All the above formal analysis holds, culminating in the relation (8). If we were confident that we could identify the k_1 variables we would clearly have found a more efficient y_c: in fact, the mean square deviation would be $(k_1 + 1) \sigma^2$ instead of $(k + 1)\sigma^2$. The analysis so far favours policies (ii) and (iii) in the omission of variables with estimated coefficients with low values of t.

Perhaps the commoner case is that in which the regression is incomplete: we have identified and included in the regression k_1 of the variables, the model containing $k = k_1 + k_2$ but have failed to identify others. Thus, in matrix form the model is:-

(9) $y = X_1 \beta_1 + X_2 \beta_2 + u,$

$(T1)\ (Tk_1)\ (k_1,)\ (Tk_2)\ (k_2,)\ (T1)$

The dimensions of the various vectors and matrices are shown in brackets ( ) under the symbols. The regression is

(10) $y_c = X_1\ b_1$

with

(11) $b_1 = (X_1'\ \ X_1)^{-1}\ X_1'y$

(12) $= \beta_1 + (X_1'\ X_1)^{-1} X_1'\ X_2 \beta_2 + (X_1' X_1)^{-1}\ X_1' . u$

on substitution for y, given by (9) in (11).

We pause here to remark that very considerable simplification in the following work is effectable by recourse to orthogonization of the raw data. Suppose that in its original form (9) was as follows:-

(13) $y = Z_1\ Y_1 + Z_2\ Y_2 + u$

Let H_1 and H_2 be the square orthogonizing matrices, themselves orthogonal, so that

(14) $H_1\ H_1' = I_{k1} = H_1'\ H_1;\quad H_2\ H_2' = I_{k2} = H_2'\ H_2$

(H_1 and H_2 are most conveniently derived from the latent vectors of Z_1 and Z_2 respectively for which there are computer programmes). Then, from (13),

(15) $y = (Z_1\ H_1)\ (H_1'Y_1) + (Z_2\ H_2)\ (H_2'\ Y_2) + u.$

Then set

(16) $Z_1\ H_1 = X_1,\ Z_2\ H_2 = X_2;\quad H_1'\ \ Y_1 = \beta_1,\ H_2'\ Y_2 = \beta_2$

If the original incomplete regression were

(17) $y_c = Z_1\ c_1,$

this passes over into the form (10) by the same transformations as in (16):-

(18) $Z_1\ H_1 = X_1;\quad H_1'c_1 = b_1.$

The great disadvantage is that in the process that original coefficients Y and c have lost their identify. We shall, however, try to work with their transforms in what follows. X_1 and X_2 can now be regarded as orthogonal, though not generally to one another. Finally we can, with further proportional changes in each column of the respective matrices, arrange that

(19) $X_1'X_1 = TI_{k_1}; \quad X_2' X_2 = TI_{k_2}.$

From (9),

(20) $\eta = X_1 \beta_1 + X_2 \beta_2$

Hence, from (10),

(21) $y_c - \eta = -X_2 \beta_2 + M X_2 \beta_2 + Mu,$

using (12) for $b_1 - \beta_1$, with

(22) $M = X_1 (X_1'X_1)^{-1} X_1'.$

The mean sum square deviation is

(23) $E(y_c' - \eta') (y_c - \eta)$

$$\beta_2' X_2' X_2 \beta_2 - \beta_2' X_2' MX_2 \beta_2 + E u' Mu$$

after some matrix algebra. We have not yet used the standardizing properties at (19); when we do so we find for (23):-

(24) $E(y_c' - \eta') (y_c - \eta) = T\beta_2'\beta_2 - \frac{1}{T} \beta_2' X_2'X_1X_1'X_2 \beta_2 + (k_1 + 1) \sigma^2$

The most interesting and generally manageable case is that in which there has been but a single unidentified significant independent variable, i.e. $k_2 = 1$, Set the single coefficient $\beta_2 = \beta_2' = \beta$ and let the correlation coefficients between the omitted variable and the k_1 identified variables be p_i, $i=1,2,\ldots k_1$. Then, from (24),

(25) $E(y_c' - \eta') (y_c - \eta) = T(1-\Sigma_i \rho_i^2) \beta^2 + (k_1 + 1) \sigma^2$

The first term on the r.s. of (25) must be non-negative. It is curious that $\Sigma\rho_i^2$ cannot exceed unity. This can easily be proved otherwise. Let $X_1 = | x_1 \; x_2 \; \ldots \; x_{k_1} |$ where the x_i are vectors (T x 1) and let $X_2 = \{z_1, z_2, \ldots, z_T\}$, now a vector. Then, if ρ is the coefficient of correlation between the vectors (T x 1) $\Sigma\rho_i x_i$ and X_2, it is easy to show that

$$(26) \quad p^2 = \Sigma_i \rho_i^2,$$

which accordingly cannot exceed unity.

16 June 1967 R. C. Geary