Selection of Significant Regressors from a Large Set: Notes for an Inquiry

By R. C. Geary

This problem may be regarded as solved when the regressors are mutually orthogonal [2]. The more general case will be considered here. The approach will be to try to transform the general case of not necessarily orthogonal regressors into orthogonal, or near orthogonal, form.

The formulation is identical with that of the orthogonal case : given a regressand y_t and a series of K regressors x_{it}, $i = 1, 2, \ldots, K$, $t = 1, 2, \ldots, T$, to identify the $k \leqslant K$ regressors from the full series in significant regression relationship with the regressand. We shall assume, without loss of generality, that all regressors are standarized, i.e.

(1) $$\sum_{t=1}^{T} x_{it} = o \; ; \; \sum_{t=1}^{T} x_{it}^2 = T, \; i = 1, 2, \ldots, K.$$

We also assume that y_t is measured from its mean, i.e.

(2) $$\Sigma y_t = o.$$

The full regression in matrix form is

(3) $$y = \beta X + u,$$

where β, the coefficient row vector is 1×K and the residual vector u (like y 1×T) is assumed a random sample from $N)o, \sigma^2)$. The least squares estimate vector b is given by

(4) $$b = yX' \, (XX')^{-1}$$

which, on substitution for y in (3), gives

(5) $$b = \beta + uX'(XX')^{-1}.$$

The var-covar matrix of b is $\sigma^2(XX')^{-1}$. In the ortho-regressor case $(XX') = TI$ which, from (5), means that $b-\beta$ is a random sample from $N(o, \sigma^2/T)$. In turn, this means that the significant regressors, k in number, are those with the largest absolute value.: a special technique has been evolved (in [2]) where, in the descending order of /b/, one stops so as to be able to state that stochastically the top k are significant and the remaining K-k not significant.

The general case is quite different. The regressors which individually are most highly correlated with y are not necessarily in the significant set of k. From (5) it is evident that $b-\beta$ is distributed on a normal surface of error, of so highly complicated a form, however, that any method of assessing significant single variablewise seems doomed to failure. The use of normal order statistics, effective in the orthogonal case, applied to regressors in their original form, seems unlikely to yield a solution in the general case.

The writer has insisted elsewhere [1] that, in general multivariate regression, the individual regressors are significant only as members of the set of all regressors included. Mathematically, multivariate least square regression is really simple regression in which the single regressor is the linear form of several variables. Ideally, therefore, significance should be tested by examining the relation of y to all 2^K linear forms (or sets of variables derivable from the K potential regressors). As K may be large, perhaps of the same order of magnitude as T, though, of course, less, mere mention of this number of regressions, even having regard to the speed and efficiency of the modern computer, rules this approach out as impracticable. Nonetheless, it is an interesting theoretical problem to consider briefly how we would recognise the right k set when we had found it using this method.

We may perhaps envisage a statistical game with two players A and B, to both of whom are available the X (KxT) matrix. A makes a selection of k variables and constructs the T values of y_t from the formula

(6) $$y_t = \sum_{i=1}^{k}\beta_i x_{it} + u_t, \quad t = 1, 2, \ldots, T,$$

where the u_t are random from N(0,1) and the β_i (positive or negative) are so large that B will have a sporting chance of identification. The set of y_t are handed to B and he is challenged to identify the set of k variables. Of course A, in setting up (6), has renumbered the variables and B has no knowledge of what A has done. B does not even know the number k but he is aware that the residuals are normal with mean zero though he does not know the variance.

Even if the procedure of the second last paragraph were contemplated, what standards could one apply to determine choice? One feels instinctively that no great harm would be done to one's regression if the choice were such that it included all of the correct set but also some variables with zero coefficients, in fact. Set

(7) $$y = \sum_{i=1}^{k} b_i x_{it} + v,$$

where the k selection now includes all the significant regressors and some others, i.e. with true coefficients (i.e. β) zero. From (6) and (7),

(8) $$v = -(b-\beta)\ X + u.$$

The row vector v is observable. It can easily be shown that the sum squares population mean is

(9) $$Evv' = Euu' - Eu\ X'(XX')^{-1}Xu'.$$

The last term on the right is the sum of the trace elements of $X'(XX')^{-1}X$ multiplied by σ^2. The trace sum is

(10) $$\Sigma_i \Sigma_j \Sigma_k a_{ji}\ A_{jk} a_{ki},$$

where a_{ij} is an element of X and A_{jk} an element of $(XX')^{-1}$. But $\Sigma_i a_{ki} a_{ji}$ is the (jk) element of (XX'). Hence (10) equals the sum of all ements in I_k, the product of a matrix by its inverse, i.e. k. Hence, from (9), the mean of the sum squares

(11) $$Evv' = (T-1-k)\sigma^2.$$

This result is, of course, classical. It is included here for the sake of completeness. In the proof, it will be noted, it was not necessary to assume normality in the residual vector u.

Starting with the full regression, i.e. with K regressors, B, in the game, could probably eliminate with safety all variables with coefficients less than twice SD, i.e. approximately the .05 normal probability level. It is evident that the standard for acceptance of a regressor as significant on a given probability level must be more stringent than in the classical case. B would then set up a new regression with the remaining k variables : he might find some additional variables with coefficients less than their new SD. These latter are eliminated. So far, he has had only two regressions. If the number k surviving (using an unchanged symbol for simplicity) is now reasonably small, say, not exceeding 10, he sets up k regressors, as found by leaving out each of the k in turn. For each he calculates the residual s^2 (i.e. the estimate of σ^2). If the s^2 for a particular elimination seems significantly larger than the lowest in the set and than the s^2 for the second regression above, the likelihood is that it is significant and should be retained. By the procedure outlined B may succeed in identifying with confidence a small set of regressors which contain the true set, known only to A. Unless A has been very generous in according large coefficient values to his selection, it seems unlikely that B will have succeeded in finding the true set, no more nor no less.

In work of this kind, however, and apart from the game, if the object be interpolation or extrapolation, it seems preferable, in cases of doubtful significance, to include the variables concerned. Even if, in fact, theyare, in truth, not significant, accuracy is not impaired by their inclusion. And they might be found significant if a larger number,(i.e. T) of sets were available.

From the foregoing paragraphs evidently some interest attaches to the estimation of s^2 when some significant variables have not been included in the regression. The regression is now

$$y = \beta_1 X_1 + \beta_2 X_2 + u, \tag{12}$$

where the matrix X_2 pertains to the significant variables included in the residue. The estimated regression is

$$y = b_1 X_1 + v, \tag{13}$$

with

$$b_1 = (yX_1')(X_1X_1')^{-1}. \tag{14}$$

Using (12),

$$b_1 = \beta_1 + \beta_2(X_2X_1')(X_1X_1')^{-1} + uX_1'(X_1X_1')^{-1}. \tag{15}$$

From (12) and (13),

$$v = -(b_1-\beta_1)X_1 + \beta_2 X_2 + u. \tag{16}$$

Using (15),

$$\begin{aligned} v = -\beta_2(X_2X_1')(X_1X_1')^{-1}X_1 &+ \beta_2X_2 \\ -uX_1'(X_1X_1')^{-1}X_1 &+ u. \end{aligned} \tag{17}$$

In each pair of terms on the right appears the fascinating symmetrical matrix

$$M = I - X_1'(X_1X_1')^{-1}X_1, \tag{18}$$

which is like the unit matrix in that its every power is equal to itself. From this it follows (cf. (9)) that

$$(19) \qquad \begin{aligned} Evv' &= \beta_2 X_2 M X_2' \beta_2' + EuMu' \\ &= \beta_2 X_2 M X_2' \beta_2' + (T-k_1-1)\sigma^2, \end{aligned}$$

where k_1 is the number of variables in X_1. The first term on the right is, of course, positive. Its dimension in T, when divided by $T-k_1-1$, is zero. It is an ordinary number, therefore of the same dimension as the population variance σ^2. Therefore, so long as the regressor solution does not contain all significant variables the residual mean square will be significantly inflated.

All that can be claimed for the straightforward approach outlined so far is that it may make some contribution to the solution of the problem and might be used in conjunction with more efficient methods.

The writer hoped that transformation of the K original regressor variables into a new set of K orthogonal regressors would help towards solution. At least the helpful property of orthogonality would be attained. It is immediately evident, however, that the ensuing change of variables means loss of identity. If, as we shall presently do, operate with a regressor matrix Z (instead of X) so that ZZ' = TI and use the orthogonal theory [1] to distinguish the significant z variables, how, if at all possible, do we infer therefrom the, say k_1, significant x variables? As others may wish to pursue this line to greater length and effect than the writer, he will set down some algebra bearing on this approach.

Let the matrix equations in the scalar λ be

$$(20) \qquad (XX')c = \lambda c,$$

where (XX') is K×K and symmetrical and c is K×1. Required to find c. A solution is possible only if

$$(21) \qquad // XX' - \lambda I // = o,$$

where // // indicates the determinant of the matrix. In every actual application the K roots in λ of (21) will be positive and distinct, say λ_i, i = 1, 2, ..., K. On substitution in (20) each root λ_i will yield proportionately a vector c_i as a solution of (20), with the well-known (and most elegant) property of orthogonality, i.e. $c_i'c_j=o, j\neq i$. To determine the absolute values of the K vectors it may be assumed that $c_i'c_i = 1$, i=1, 2, ..., K. The required linear transformation is then

$$Z = CX, \tag{22}$$

where C is the orthogonal square matrix of which the ith row is c_i'. We shall now show that the transformed regressor matrix Z is orthogonal.

From (22),

$$ZZ' = CXX'C'. \tag{23}$$

But, from (20),

$$XX'C' = C'L, \tag{24}$$

where L is the diagonal matrix of the λ_i. From (23) and (24),

$$ZZ' = CC'L = L, \tag{25}$$

proving the property. There is an infinity of transformations of X into orthogonal form but (22) is unique in that it utilises the original regressors symmetrically. As it does not involve the regressand y it has no stochastic implications. In its regression properties Z is mathematically equivalent to X. Unfortunately, as stated above, the writer does not see how it can be utilised for the present purpose. He has tried other orthogonal transformations equally with lack of success using this approach.

A more promising method seems to be the following. A characteristic feature of the foregoing transformation of the x_{it} was that it pertained to the i and was the same for all t : we may therefore term this a row transformation. As noted above, it had the disadvantage of loss of identity of the original regressors. In this section we consider column transformations.

The simplest of these is the finite differencing procedure. Model (3), in non-matrix form, is

$$(26) \qquad y_t = \sum_{i=1}^{K} \beta_i x_{it} + u_t, \quad t = 1, 2, \ldots, T.$$

Hence

$$(27) \qquad \Delta^p y_t = \Sigma \beta_i \Delta^p x_{it} + \Delta^p u_t, \quad t = p+1, \ldots, T,$$

where, by definition, $\Delta z_t = z_t - z_{t-1}$. One contemplates, in the first instance, the application of the method to time series in which the regressors are mutually correlated because they have the form

$$(28) \qquad x_{it} = f_i(t) + u_{it},$$

where $f_i(t)$ is a polynomial in t of degree p_i and the u_{it} are completely random residuals. It is evident that the Δ^p operator applied to the x_{it}, where p is equal to the largest of the p_i will ensure that the new regressors lack significant intercorrelation. It is, however, by no means certain that this method will be effective with economic time series. For instance, the accompanying table of initially highly correlated data exhibits a dramatic decline after one difference but, in several cases, there seems to be a tendency towards a limiting non-zero correlation with increasing p. However, the method seems worth trying out.

If the problem is not one in time series one could adopt the following procedure; indeed, it might effectively be used even with time series. In the first place, one would calculate the correlation coefficient between y and each of the K regressors. Select the regressor with the largest correlation coefficient and reorder the T equations in descending order of the magnitude of the selected variable. Then apply the differencing process.

Even if the process results at some stage p in near-zero intercorrelation between regressors, an obvious objection to the application of orthogonal theory to assessment of significant regressors is that auto-correlation has been imparted to the residuals by the differencing process. Now, to apply ortho-theory, it is necessary to assume the residuals random, i.e. non-autocorrelated, so that the least squares estimation can be applied. We shall, in fact, *pretend* residual non-autocorrelation. In the full K-regression the new estimates of the coefficients will differ from the *true* estimates b_i which tend in probability to the β_i. But the object of the exercise is *not* to estimate the coefficients - this has been done by the original K-regression - but to *identify* the k significant regressors. Adverting to the A-B game the proposition scarcely requires formal proof that the regressors significant in the original series will be identical with those in the differentiated series. This seems to be all that is required of them.

The differencing process of degree p involves the loss of p degrees of freedom in the system - see (27). The writer came across the following Helmert-type transformation with the full T DF, applicable to time series. The T×T transformation matrix D applicable to the y_t, the x_{it} and the u_t is as follows:-

$$(29)\quad D = \left|\begin{array}{ccccccccccc}
a_1 & a_2 & \cdots & a_p & o & o & \cdots & o & o & \cdots & o \\
o & a_1 & \cdots & a_{p-1} & a_p & o & \cdots & o & o & \cdots & o \\
\vdots & & & & & & & & & & \\
o & o & \cdots & o & o & o & \cdots & a_1 & a_2 & \cdots & a_p \\
\hline
b_{11} & b_{12} & \cdots & & \cdots & & \cdots & \cdots & & \cdots & b_{1T} \\
b_{21} & b_{22} & \cdots & & \cdots & & \cdots & \cdots & & \cdots & b_{2T} \\
\vdots & & & & & & & & & & \\
b_{p1} & b_{p2} & \cdots & & \cdots & & \cdots & \cdots & & \cdots & b_{pT}
\end{array}\right| \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \\ \end{array}\right\} T-p \text{ rows} \\ \left.\begin{array}{l} \\ \\ \\ \\ \end{array}\right\} p \text{ rows} \end{array}$$

The a-part are those from Δ^p; e.g, if p = 2, $a_1 = 1$, $a_2 = -2$, $a_3 = 1$. The b-part are the orthopolynomials of successive degrees appropriate to T given in [3] , Table XXIII e.g. the first row is always a succession of T units; the second row, for T = 2w + 1, is -w, -w + 1, ..., -1,0,1,2, ..., w.

Irish Macro-Economic Entities at Current Prices, 1947-1962: Correlation Coefficients

[I: Raw Data; II: ΔI; III: $\Delta^2 I$; IV = $\Delta^3 I$]

	1	2	3	4	5	6	7
1. Personal Expenditure							
I		.995	.887	.934	.981	.997	.996
II		.57	.43	.59	.16	.67	.26
III		.37	.14	.62	-.02	.51	.24
IV		.35	.44	.67	.15	.56	.25
2. Public Authority Expenditure							
I			.913	.939	.980	.990	.991
II			.53	.49	.15	.14	.14
III			.34	.45	-.21	-.39	-.14
IV			.38	.48	-.36	-.39	-.31
3. Fixed Capital Formation							
I				.918	.896	.905	.897
II				.45	.09	.51	.18
III				.29	.30	.01	-.19
IV				.19	-.40	-.22	-.30
4. Imports							
I					.943	.931	.924
II					.22	.35	.02
III					.17	.26	.15
IV					.22	.33	.06
5. Exports							
I						.980	.988
II						.28	.27
III						.22	.34
IV						.35	.40
6. Money Supply							
I							.995
II							.25
III							.18
IV							.27
7. Gross National Product							
I							
II							
III							
IV							

The a-rows are not orthogonal; the b-rows are; and each a-row is orthogonal to all b-rows. The transformation is not applied in the present paper but is placed on record because it may be useful in some other connection.

At the best, the foregoing treatment will result K regressors each pair of which are not significantly correlated, not <u>exactly</u> uncorrelated as in the case of the treatment of [2]. It is interesting to see if a transformation can be found which will yield ideally exact zero correlations. Let such a transformation, applied to the regressor matrix X be

$$(30) \qquad Z = XD,$$

where D is T×T, whence Z, like X, is K×T. There are initially T^2 determinable elements, or DF, in D. Orthogonality required that $ZZ' \equiv XDD'X$ should be a diagonal K×K matrix, to attain which requires K(K-1)/2 D.F. In addition the means of the K new variables, like the rows of the X matrix should be zero - K in all. Clearly, with T^2 elements to dispose of, vastly in excess, with T large, of the number of conditions to be satisfied, it should not be impossible to find such a transformation matrix D. So far, however, a method for finding it has eluded the writer. To complete the transformation of the original T equations, D should be applied to y and to the residual row vector u, to give yD and uD. The elements of the latter row vector will no longer be independent, as postulated for u. Nonetheless, for the reasons given above, one can apply the theory of [2] as if the elements of uD were independent for the purpose of <u>identifying</u> the significant variables in the original formulation.

It may be of interest to observe that, as the elements of D satisfying the foregoing conditions are presumed determined, the maximum likelihood solutions b and σ^2 of the transformed equation set is identical with the original ML solution. In fact, the original integral element of the frequency distribution

$$(31) \qquad f(u)du.$$

For the ML solution u, in f(u), is to be regarded as a function of the parameters β from (3) as well as the parameters (σ^2 etc) of $f(u)$, deemed unknown. The ML solution consists of the values b of β and the frequency parameters s^2 etc which maximize $f(u)$. Transform (31) by $u = D^{-1}v$ and the frequency element of v becomes

$$(32) \qquad f(D^{-1}v)/J(u,v)/dv,$$

where J is the Jacobian of the transformation. It equals, in fact, $/\!/ D^{-1} /\!/$ and is therefore parameter-free. Hence the ML solution is the parameter-set which maximizes $f(D^{-1}v) = f(u)$ which proves the proposition since, with the presumption of normality in u, the only frequency parameter is σ^2.

The most promising line of those contemplated is the Δ^p transformation. Its effectiveness remains to be seen by applications preferably to actual data but, if these are not readily available, to A-B statistical games as indicated above.

REFERENCES

1 GEARY, R. C. (1963). Some Remarks about Relations between Stochastic Variables : A Discussion Document. Review of the International Statistical Institute, Vol. 31 : 2.

2 GEARY, R. C. (1965). Ex-Post Determination of Significance in Multivariate Regression when the Independent Variables are Orthogonal. ERI Memorandum Series, No. 27.

3 FISHER, R.A. & YATES, F. (1957). Statistical Tables for Biological, Agricultural and Medical Research. Table XXIII.

Addendum to Memorandum No. 28

A much simpler form may be given to the first term on the right of (19), namely

(30) $$\beta_2 X_2 M X_2' \beta_2'.$$

Suppose that the LS regression of X_2 $(k_2 \times T)$ on $X_1(k_1 \times T)$ is

(31) $$X_2 = CX_1 + D,$$

where the coefficient matrix $C(k_2 \times k_1)$ is found from relation

(32) $$X_1D' = 0 \text{ or } DX_1' = 0.$$

Note that the relationship (31) is non-stochastic. Using (18) and (31), (30) becomes

(33) $$\beta_2(CX_1 + D)\,[\,I - X_1'(X_1X_1')^{-1}X_1\,]\,(X_1'C' + D')\beta_2'$$

which is made up of two terms of which the first is

(34) $$\beta_2(CX_1 + D)(X_1'C' + D')\beta_2' = \beta_2(CX_1X_1'C' + DD')\,\beta_2'.$$

The second term, more complicated, is

$$-\,\beta_2(CX_1 + D)\,[\,X_1'(X_1X_1')^{-1}X_1\,]\,(X_1'C' + D')\beta_2'$$

$$= -\beta_2 CX_1(X_1'C' + D')\,\beta_2'$$

(35) $$= -\beta_2 CX_1X_1'C'\beta_2'$$

Adding (34) and (35), (30) becomes simply

(36) $$\beta_2 DD'\beta_2'.$$

This result is almost intuitive: having regressed y on any set X_1 the addition to the variance Evv' of the remaining set X_2 can depend only on the residual contribution of X_2, having allowed for X_1, already in the regression. Finally, from (19)

$$(37) \qquad Evv' = \beta_2 DD'\beta_2' + (T - k_1 - 1)\sigma^2$$

The statement in the text about dimensions in T under (19) is, from (37), wrong. The first term on the right of (37) is of dimension O in T.

It may be worth noting that the matrix M, given by (18) has the property

$$(38) \qquad MX_1' = O \quad \text{or} \quad X_1M = O$$

M does not appear to have been accorded an explicit role in regression literature, perhaps because, in the pre-computer age, a T x T matrix was practically inconceivable, operationally.

By the way, the second term on the right of (37) is derived like (11).