THE ECONOMIC
AND SOCIAL
RESEARCH INSTITUTE
MEMORANDUM SERIES
NO. 139

A Note on OLS Multivariate Regression with Suggestions for Additions to Routine Computer Programmes

R. C. Geary

A Note on OLS Multivariate Regression with Suggestions for Additions to Routine Computer Programmes

Let the OLS regression be:-

(1) $$Y = a + \sum_{i=1}^{k} b_i X_i + e.$$

Geary has argued, somewhat controvertially, that the individual coefficients b_i are meaningful only in the cases of simple regression or of each pair of the k indvars X_i being not significantly correlated, very rare in practice. Then, and only then, could one state "a rise of 1 in X_i causes a rise of b_i in Y." Otherwise this inference is false. The argument holds that the only valid purpose of the multiple regression is the estimation of Y, usually in the form of extrapolation of time series beyond the estimation period. This means that the coefficient vector $\{a, b_1, b_2, \ldots b_k\}$ is objectively meaningful, but not its individual elements.

There is not much point in extrapolation of time series unless the $\bar{R}^2$ is near unity and the DW or tau are near their white noise values, i.e. near 2 and near half the number of residues respectively. On account of the usual high intercorrelation when the indvars are time series, most researchers prefer to work with the deltas, ΔY and ΔX_i, a procedure which incidentally will usually considerably reduce and in some cases eliminate intercorrelation between the indvars, i.e. the pairs of X_i may be highly correlated but not the pairs of ΔX_i.

DW owes its inception entirely to assessment of residual autocorrelation in the case of time series, and this because of a characteristic property of time series, namely that they are autocorrelated to start with. The thought process is as follows. Y, to be explained, is a time vector. Treating it as an OLS residual (i.e. fitting merely a constant to it) we compute its DW as -

(2) $$DW = \sum_{t=2}^{T} (y_t - y_{t-1})^2 / \sum y_t^2, \; y_t = Y_t - \bar{Y},$$

and customarily find that this DW has a very low value, well below 2, indicating significant autocorrelation. If we found a value near 2 we could go ahead with our

OLS regression but there would be no point in using DW or tau as indicating completeness of relationship, since all the DWs and taus would be near their white noise values, i.e. like the original Ys the successive values of the residuals would be random to one another. Or, given original time series data, $Y, X_1, X_2, \ldots, X_k$, randomizing these would make no difference in computation to the values of the coefficients but it would destroy the useful role of DW or tau, the values of which depend on the ordering of the data.

Reverting then to the typical time series case of Y's having a very low value of DW, we imagine ourselves computing the OLS regressions successively Y on one X (simple regressions), on two Xs etc and computing DW or tau on the residuals in each case. We stop when we have found a DW or tau which indicates that the residuals are probably random to one another. There may be several such sets because of intercorrelation between the X's. There are computer programmes for systematic selection of the best sequence of Xs to bring into the OLS regression, so avoiding an immense series of such regressions.

With time series when, with one's OLS regression, one has obtained an $\bar{R}^2$ near 1 and DW or tau indicating probable residual randomness (or white noise), one may go ahead with extrapolation in time, such extrapolated estimates being subject to known probabilistic ranges of error. One may be wrong, perhaps due to new variables (i.e., other than the X_i) affecting the relationship, but one may at least state that as far as past experience goes the extrapolations should be as stated.

The multiple regression computer prints-out have the silly habits of producing DWs, even when time series are <u>not</u> involved. Such values are meaningless, a remark which would apply also to the tau value. To repeat, the essence of time series is that time automatically orders the data (Y, X_i) in a particular way. In the non-time series case (say cross-section) the data should also be ordered <u>before</u> computation of DW or tau. In simple regression (i.e.

one indvar, say X_1) the obvious course would be to reorder the residuals according to the magnitude of X_1 and then compute DW or tau; if Y or X_1 are related this will result in Y being ordered, generally increasing or decreasing, i.e. autoregressed like time series. In the case of multiple regression the best course might be to reorder the residuals according to the magnitude of the principal component of the indvars. Could not the computer be programmed to do this?

At one time I thought that this problem of indvar intercorrelation could be bypassed so as to make the coefficients meaningful by substituting for the matrix of indvars the matrix of components. These are linear functions of the original variables X_i and number also k, and have the precious property that each pair is exactly uncorrelated (i.e. for each (i, j) $j \neq i$ $r_{ij} = 0$). This procedure might have the added bonus of reducing the number of indvars, i.e. only the first one, two or three having statistically significant coefficients.

The trouble about using their principal components instead of original indvars is that the latter have identity and the former have not. Thus, if one were studying the effect of a change in social welfare payments on unemployment, one indvar might be B.M. Walsh's percentage of s.w. payments to wages together with other indvars, the depvar the unemployment rate. Suppose that the coefficient of the ratio X_1 was significant, its value b_1. It would be perfectly sensible to ask "What would be the effect on Y of an increase of 1 in X_1?", even if the answer were not b_1. But if the first <u>component</u> were, say, X_1' and its significant coefficient b_1', it would simply be meaningless to ask what would be the effect on the unemployment rate of an increase of 1 in X_1', because, in general, we don't know what X_1' is; we can't describe it. The same remark applies to other components with even greater force since, while the principal component in a sense synthesizes all the indvars, the other components are much more difficult to identify.

To some minds the statistically significant individual coefficients can be regarded as having a meaning because of the Frisch-Waugh theorem which states that the b_1 is <u>exactly</u> the value which would be found from a simple regression -

$$(3) \qquad RY = a_1 + b_1 RX_1 + e_1,$$

R being a symbol for residue, in fact the residues when Y and X_1 are each OLS - regressed on the remaining indvars ($X_2 X_3 \ldots X_k$). So it would be right to state that b_1 is the effect of a change of 1 in the first variable on the depvar when each of those two variables have been corrected for the effects of the other indvars.

Geary generalized Frisch - Waugh to the following effect. Write (1) in the form -

$$(4) \qquad Y = a + \sum_{i=1}^{k_1} b_i X_i + \sum_{j=k_1+1}^{k_1+k_2} b_j X_j + e$$

with $k_i + k_2 = k$, the variables having been divided arbitrarily into two groups of k_1 and k_2. Then -

$$(5) \qquad RY = a_1 + \sum_{i=1}^{k_1} b_i RX_i + e_1,$$

the Rs indicating the residues when the variables Y, X_1, X_2, ... X_{k_1}, and each OLS - regressed on $X_{k_1+1}, \ldots, X_{k_1+k_2}$. The b_i in (4) and (5) would be identical. This would have been an efficient way to calculate the coefficients of (1) using a primitive calculating machine but has little practical point with the advent of the computer, apart from algebraic interest, in providing a meaning for individual regression coefficients.

The statement, based on the original regression, that a change of 1 in X_1 would result in a change of b_1 in Y would be valid were it not for the necessary condition which is that the other indvars remain unchanged. In general the latter condition is unacceptable because if there is a significant correlation between X_1 and any other variable, say X_2, X_2 cannot be presumed to remain unchanged when X_1 changes by 1.

The previous paragraph hints at a procedure which might yield an answer to the question of the effect on the depvar of an increase of 1 in a particular indvar, say X_1. Regress all the other indvars on X_1 -

(6) $X_i = a_i + b_{i1} X_1, \; i = 2, 3, \ldots, k.$

One infers that a rise of 1 in X_1 would entail a rise of b_{i1} in X_i. Hence the total effect of a rise of 1 in X_1 will be found by substituting the values 1, b_{21}, b_{31}, ... b_{k1} for the k indvars in (1), ignoring e.

All practitioners agree that it is a sound principle in multiple regression to use as few indvars as possible consistent, of course, with high values of $\bar{R}^2$ and residual randomness. This may be the place to remark that the statistical process of OLS regression is anything but an exact science. Wise judgment is of the essence, such judgment being based on a plentiful supply of computer data derived from ample routine computering. One must be very careful about elimination of indvars.

One's original set of indvars are presumably those which theory or simple ratiocination indicates might be related to the depvar and whose (depvar, indvar) simple correlation is statistically significant. But here it might be argued that it might be prudent to retain the variable in one's set even if uncorrelated with the depvar, since with other indvars it may be correlated. Anyway, assume that one has a large set of indvars to start with.

One begins with an OLS regression on this whole set. The F or $\bar{R}^2$ test will indicate its significance. If insignificant there is no point in going ahead. If DW or tau indicates residual autocorrelation new indvars should be sought. If the $\bar{R}^2$ is close to unity on the original set of indvars the OLS regression may be, useful for extrapolation even with "bad" DWs and taus, on the assumption that the missing indvars will not affect the extrapolated value much. One notes that some of the variables have insignificant values. Leave these out and repeat the OLS regression. If this omission does not alter the value of $\bar{R}^2$ much the omission is justified: the trouble here is (and where judgment enters) is that we have no way of knowing how much is "much". If the value of $\bar{R}^2$ is lowered one must experiment further as to which of the low-coefficient indvars to retain in the set. One must not automatically reduce one's original set of indvars because on whole-set regression their coefficients are insignificantly small. Leser-Geary have shown that one can have significant OLS regression (by the F test) with all coefficients not significant in a multiple regression; it is only in simple regression (one indvar) that the F and t (coefficient) tests are absolutely consistent with one another, as regards probabilistic inference. Multivariate regression is the OLS regression process of the depvar on the whole set of indvars in which it is impossible to isolate individual indvars. In this sense all regression is "simple".

It is suggested that the following be added to routine computer processes for OLS regression -

(i) in non-time OLS regression, reorder residuals according to magnitude of principal component of indvars (or of the magnitude of the single indvar in the case of simple regression) before calculation of DW or tau.

(ii) for assessing the effect of an increase of 1 in each indvar on the depvar, allow for the effect of all other indvars by according the variable in question, say X_i, the value 1 and other variables the value b_{ji} given by the simple OLS regression of X_j on X_i.

(7) $$X_j = a_j + b_{ji}X_i + ej \ \forall j \neq i.$$

The print-out would give the values of b_{ji} and the actual values on the depvar of an increase of 1 for each indvar.

The computer can select the "best" set of indvars from a large initial set when it is given the rules of selection. In this and other application the computer may be too good; on producing the near perfect answer and withholding intermediate information, valuable aids to interpretation may be lost. In the present application, for instance, all pairs of c.c.s in (Y, X_i) should be given, perhaps the b_{ji} suggested here as well.

A weakness in the b_{ji} proposal is that (7) (and indeed all OLS regressions) is a cause-effect statement: in (7) X_i is the cause of X_j. This may not be the case. A statement neutral to cause - effect might be better: "a rise of 1 m X_i will be accompanied by a rise of c_{ji} in X_j." This issue of cause-effect v. functional was much discussed years ago and techniques evolved for dealing with the neutral case. These techniques are difficult, indecisive and generally unsatisfactory, so OLS procedure may remain if as a pis aller.

Significance in OLS regression is nearly always assessed by reference to a null-hypothesis table for the F test using degrees of freedom: with number of sets of data T and number of indvars k, these are k (numerator) and (T - k -1) (denominator). One is confronted with the problem of deciding by F which of a set of regressions based on different sets selected from the large original set is the best. Now the present computer programmes of which I am aware simply yield the value of F. It would be better that the NHP* should be given. It is admitted that this would be much to expect since it implies the store's containing all the null-hypothesis frequency distributions of F for two dimensions of degrees of frequency. The F table I use has simply the critical NH values of F for each pair of d.f. and a number of two-ended probability levels, the lowest .005: for instance, with d.f.s 7 (= k) and 24 (i.e. number of sets of observations is 32 = T), F = 3.99. One assesses that NHP is less than .005 if one obtains a value of F greater than 3.99 but one does not know what this probability is.

Now the regression is not much use unless the actual value of F does not greatly exceed the lowest NHP tabled value. It also happens that the tabled NHP values vary greatly with the number of indvars k. Thus for probabilities .005 and .05 and T = 32,

* NHP = null hypothesis probability.

	Critical value of F for NHP =		
k	.005	.05	Ratio
1	9.18	4.17	.45
2	6.40	3.33	.52
3	5.32	2.95	.55
4	4.74	2.73	.58
5	4.38	2.59	.59
6	4.15	2.49	.60
7	3.99	2.42	.61

It is inferred from these figures that whatever NHP one is working to, the value of F must be the greater the fewer the number of indvars. In the practical case we may have to choose between two regressions with different number of indvars with F values very much larger than those of the tables and hence with infinitesimal (but unknown) NHP. I would base my preference on the regression with the lower NHP, if I knew it, with a very strong hanker for the regression with fewer indvars as the better one may be able to rationalize.

To take a specific example suppose that, as above, T = 32. In two selections of k = 2 and 7 one finds values of F of 24 and 13 respectively. Which regression is one to prefer? One might argue this way from the foregoing table. Corresponding to the ratios in the last columns (of critical values for NHP .05 to NHP .005) we would have ratios .27 (= 6.40/24) and .31 (=3.99/13). Since one expects a rise in the ratio with size of k one would assess the regressions as about equal in significance and one would prefer that with the two indvars.

All this turgid ratiocination would be unneccessary if, as well as F values, the computer would supply the NHPs. Also, with its vast and increasing sophistication, should it not supply the best set of indvars from a large original set. But give us prints - out of intermediate steps as well for our better interpretation of the results.

Would ESRI computer experts interest themselves in devising a programme incorporating the points in this note?

As a short footnote to the foregoing assessment of - the aggregate effect of a rise of 1 in X_i on y, it might be asked "Should not insignificantly valued b_{ji}s be omitted (i.e. regarded a having value zero)?"I have no strong views on the subject and would value those of my colleagues. I am inclined to favour leaving all values in because of (i) simplicity for the computer, (iii) the values may be genuine (e.g. "sign right" according to theory), (iii) values will be small and with different signs, so that effect of inclusion or exclusion on answer to "increase of 1" will lie within the confidence limits of the latter.

18 January 1980 R.C. Geary.