Ex-Post Determination of Significance in Multivariate Regression when the Independent Variables are Orthogonal

by

R. C. Geary

The Economic Research Institute, Dublin

In studying a well-known application of analysis of variance to multiple regression using time series the following problem arose : given a set of observations $y_t (t = 1, 2, \ldots, T)$ and an orthogonal set of K independent variables (which may be indefinitely extendable) $x_{it} (i = 1, 2, \ldots, K)$, how does one ex post identify the variables (i.e. the i) which make a significant contribution to the regression? The problem is a classical one in harmonic analysis and has been solved in different conditions by A. Schuster, G. Walker and R. A. Fisher (in [1] where the theorems of the other two authors are conveniently summarized). In the harmonic case the independent variables are Fourier terms and the fact that they can be grouped in pairs (to give the amplitude squared, the statistic discussed by all three authors), facilitates the solution.

The problem is to select k from the full series of K independent variables so that the coefficients of the k variables are significantly different from zero. While a large choice K of independents is presumed available, the writer has in mind that the significant regression will contain not more than 5 or 6 (=k) independent variables. For any of the purposes to which the ultimate formula may be put, e.g. in economic forecasting, the view is taken that increase in the number of variables in any single equation is markedly subject to the law of diminishing returns in usefulness. It would be quite easy to evolve a formula exactly representing the values of the dependent variable, (and ostensibly with no error term at all) but the question is :

for extrapolation purposes can any more confidence be reposed in it than in a formula with but a few terms? The writer's reply, based on experience, would be no.

The assumption of orthogonality means that

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

We may also assume that the means of the x_{it} and y_t are zero. If the full model be

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

where u_t is assumed $N(o,\sigma^2)$ and the u_t independent, then the estimates b_i of β_i are given by

$$(3) \qquad b_i = \beta_i + \Sigma_t x_{it}u_t / \Sigma_t x_{it}^2, i = 1, 2, \ldots, K.$$

Hence the $b_i - \beta_i$ is a random normal sample of K with variance $\sigma^2/\Sigma x_{it}^2$, or the statistics

$$(4) \qquad z_i = b_i\sqrt{\Sigma_t x_{it}^2}, i = 1, 2, \ldots, K,$$

if all the β_i are zero, are a random sample of K from $N(o,\sigma^2)$. In what follows we shall find it convenient to use the notation $N_+(o,\sigma^2)$ for the positive half of $N(o,\sigma^2)$.

As we assume the extreme position of having no <u>a priori</u> theory of relationship, our selection must be guided entirely by the absolute magnitudes of the z_i, i.e. $/z_i/$. Our approach must therefore be that of order statistics. Logically we must assume that, if a particular variable j be deemed significant, each variable i with $/z_i/ \geqslant /z_j/$ must also be significant. This assumption seems justified by the independence of the z_i. It will not necessarily be so in the more general (and more difficult) case of non-orthogonal independent variables.

Suppose that k of the variables be numbered in descending order, i.e. so that $/z_1/ > /z_2/ \ldots > /z_k/$, where the number k is arbitrary but $\geqslant k$, to be determined by some test as a number such that $/z_1/, \ldots, /z_k/$ are

significant at the given probability level α but $/z_{k+1}/$ is not significant.

The writer found it unexpectedly difficult to conceive of a null-hypothesis. At first it appeared that one should assume once for all a random sample of K from $N_+(o,\sigma^2)$ and decide that if the kth ordered value were significant but the (k + 1)th not significant then the first k were the variables required : let this be test A. An apparently insurmountable theoretical objection to this approach is that, tested individually in the particular sample, one might find, say, the 4th ordered value, significant but, say, the 3rd value not significant, despite the fact that $/z_3/>/z_4/$, which seems illogical; at any rate it is contrary to the hypothesis assumed above.

Normal Theory Probability Points

The writer prefers the following approach (test B) which avoids the foregoing difficulty. If at stage j(j<k) a variable is being tested, the null-hypothesis will be that in a random sample of K-j+1 from $N_+(o,\sigma^2)$ the largest member $/z_j/<\lambda_\alpha$, the α-probability point when $\sigma^2=1$. This point is found as the solution in x of

$$p^n = 1 - \alpha \tag{5}$$

with

$$p = \sqrt{\frac{2}{\Pi}} \int_o^x du\ e^{-u^2/2} \tag{6}$$

where n = K-j+1. p^n in (5) is, of course, the distribution function of the largest element in a random sample of n. The procedure is systematic. One starts with $/z_1/$ in a sample of K. If this value be deemed significant, $/z_2/$ in a sample of K-1 is tested, and so on. The process stops when $/z_k/$ is significant but $z_{k+1}/$ is not.

Table 1 displays the .05 and .01 probability points for the first six ordered statistics for certain sample sizes from $N_+(0,1)$. Despite his

preference, for the reason given above, for test B, which requires only 1st order significance points, those for the first six orders are given in Table 1, for those who prefer test A. An application given later will show that the two tests result in quite different decisions. It would be well to have this matter clarified further. The 1st order probability points for sample sizes n=10-50, by units, are shown in Table 1A.

Estimate s^2 of σ^2

We make no attempt here to develop a Studentized theory, so that, to apply normal theory, we require to estimate the variance σ^2.

We have now to make the crucial decision that we are not interested in any regression with more than k independent variables, where $K \geq k \geq k$. Hence, since k variables may be under test, we exclude the descending ordered values of $/z_i/$, $i = 1, 2, \ldots, k$, actually found in the sample from the computation of s^2, the estimate of σ^2. In the null-hypothesis case the sum square analysis of $\Sigma(y_t-\bar{y})^2$ is

$$(7) \qquad \sum_{t=1}^{T}(y_t-\bar{y}) = z_1^2 + z_2^2 + \ldots + z_k^2 + S^2,$$

where S^2 is the residual sum squares. The expectations of each of the first k terms on the right of (7) on the null-hypothesis will be substituted for the $z_1^2, \ldots, z_k^2$ as s^2E_i, where the E_i are the values for $N_+(0,1)$ and the actual values from the sample for the last term (to give S^2), so that

$$(8) \qquad (T-1)\, s^2 = s^2 \sum_{i=1}^{k} E_i + S^2$$

or

$$(9) \qquad s^2 = S^2/(T-1-\Sigma E_i)$$

The values of E_i for $i = 1, 2, \ldots, 6$ for certain values of T are given in Table 2.

TABLE 1

Probability Points (.05 and .01) for Order Statistics x_i (i = 1, 2, 3, 4, 5, 6) from $N_+(0,1)$ for Samples 10 to 50

Sample size n	Decending order i					
	1	2	3	4	5	6
	Probability .05					
10	2.80	2.09	1.71	1.44	1.22	1.03
15	2.93	2.25	1.90	1.66	1.47	1.31
20	3.02	2.36	2.03	1.80	1.63	1.48
30	3.14	2.52	2.20	1.99	1.82	1.69
40	3.22	2.63	2.31	2.11	1.95	1.83
50	3.28	2.72	2.40	2.21	2.04	1.93
	Probability .01					
10	3.29	2.41	1.98	1.68	1.44	1.23
15	3.40	2.57	2.16	1.89	1.67	1.50
20	3.48	2.67	2.29	2.02	1.82	1.66
30	3.59	2.81	2.45	2.21	2.01	1.86
40	3.66	2.91	2.55	2.31	2.13	1.99
50	3.72	2.98	2.62	2.40	2.21	2.08

NOTE

Found by equating the first i terms in the binomial expansion $(p+q)^n$ to $1-\alpha(=.95,.99)$, to give solution p which equated to the positive normal distribution gives upper limit x, the figure tabled. This is derived from the standard table [3] for the full range normal probability P (i.e. from $-\infty$ to /x/) from the relation $P = (1+p)/2$. Figures for order 1 were derived directly from $p^n=1-\alpha$. Figures for orders 2, ..., 6 were derived by inverse graphical interpolation from [4]. The figures for these latter orders may be out by not more than one unit in the second decimal place; quite accurate enough for the present purpose, the figures will require recalculation for definitive tabulation.

TABLE 1A

Probability Points (.05 and .01) for 1st Order Statistics for Samples of 10 to 50 by Units from $N_+(0,1)$

Sample size n	Probability .05	Probability .01	Sample size n	Probability .05	Probability .01
10	2.800	3.289	30	3.137	3.587
11	2.830	3.316	31	3.146	3.595
12	2.858	3.339	32	3.155	3.603
13	2.883	3.362	33	3.164	3.611
14	2.906	3.382	34	3.173	3.619
15	2.928	3.401	35	3.182	3.626
16	2.948	3.419	36	3.190	3.634
17	2.966	3.435	37	3.197	3.641
18	2.984	3.452	38	3.205	3.648
19	3.000	3.465	39	3.213	3.655
20	3.016	3.479	40	3.220	3.661
21	3.031	3.492	41	3.227	3.667
22	3.045	3.504	42	3.234	3.673
23	3.058	3.517	43	3.241	3.679
24	3.071	3.527	44	3.247	3.685
25	3.083	3.538	45	3.253	3.691
26	3.095	3.549	46	3.260	3.697
27	3.106	3.559	47	3.266	3.702
28	3.117	3.568	48	3.272	3.707
29	3.127	3.577	49	3.277	3.713
			50	3.283	3.718

TABLE 2

Value of Ex_i^2 for Random Samples of n from $N_+(0,1)$

Sample size n	Descending order i							
	1	2	3	4	5	6	7	8
10	3.799621	2.171462	1.426472	0.970990	0.660253	0.437538	0.275135	0.155713
20	4.916871	3.216540	2.410593	1.897055	1.528207	1.245702	1.020668	0.836765
30	5.599340	3.867966	3.037613	2.502189	2.112625	1.809929	1.564854	1.360810
40	6.093230	4.343362	3.498975	2.951316	2.550458	2.237010	1.981502	1.767200
50	6.480929	4.718344	3.864523	3.308782	2.900577	2.580232	2.318119	2.097405
60	6.800321	5.028251	4.167506	3.605907	3.192432	2.867188	2.600425	2.375213
70	7.072022	5.292497	4.426376	3.860271	3.442774	3.113818	2.843555	2.615017
80	7.308510	5.522905	4.652444	4.082727	3.662028	3.330132	3.057110	2.825948
90	7.517919	5.727220	4.853153	4.280453	3.857122	3.522820	3.247552	3.014259
100	7.705850	5.910793	5.033661	4.458440	4.032894	3.696576	3.419431	3.184363

NOTE

Data kindly supplied by F. M. O'Carroll, using an Elliott 803 computer, by courtesy of The British Petroleum Company Limited. To estimate Ex_i^2 for values of n between consecutive pairs shown, log n might be used for graduation. The quasi-constancy of the ratio $Ex_i^2 / \log_e n$ is quite remarkable: with the value of 1.650 for n = 10 it gradually increases with n from 1.641 for n= 20 to 1.673 for n = 100. The phenomenon, however, becomes progressively less marked as the order number i increases and the value of n declines.

In order to obtain some experience about how formula (9) works in practice ten random samples of 20 each were drawn from N(0,1). Nine values of s^2 were obtained for each using ordered statistics and deriving the B_i from Table 2:-

Order										Average
1	.94	1.08	1.49	.44	.70	1.47	.84	1.04	.57	.73	.93
2	.97	.87	1.63	.42	.54	1.70	.85	1.07	.59	.76	.94
3	.97	.88	1.71	.35	.49	1.85	.87	1.01	.60	.78	.95
4	.94	.89	1.73	.37	.48	1.93	.83	.95	.62	.77	.95
5	.97	.92	1.69	.39	.48	1.97	.84	.88	.61	.72	.95
6	1.00	.92	1.73	.42	.47	1.92	.88	.85	.59	.75	.95
7	1.04	.93	1.75	.46	.46	1.75	.89	.82	.57	.77	.95
8	1.05	.89	1.79	.50	.47	1.69	.88	.79	.61	.82	.95
9	1.04	.77	1.93	.52	.48	1.51	.86	.77	.64	.87	.94

The true value of σ^2 is, of course, unity. The first row is the ordinary mean square estimate, the second row is found from (9) using B_1 etc. It is obvious that the column figures generally take their aspect from those in the first row. The last (average) column is reassuring in showing that the process of estimation is unbiased. The vagaries of the figures in the first row point to the need for a Studentized theory of the present problem, when T is as small as 20, though it is some small consolation that we are concerned with the $\sqrt{}$ of the figures shown, and not with the figures themselves.

It is necessary to stress the arbitrariness of the number k. Experience shows that the estimate s^2 of residual variance σ^2 is considerably influenced by choice of k. One reason, anyway, for keeping it small is seen in (8) : thereby the quantity of data for estimating σ^2 is the larger than if k were greater.

An Application

The foregoing theory will now be applied to data furnished by R. A. Fisher and F. Yates [2]. These data are the difference in yields (bushels per acre) on two plots of wheat which differ only

in manurial treatment, in the thirty years 1855-1884. The problem is : can these data be satisfactorily represented by a regression on a few orthogonal polynomials?

The x_{it} of the earlier formulae are therefore orthogonal polynomials always with T = 30. These are tabled in [2] for i (i.e. the degree in t of the polynomial) 1 to 5, and T (the authors' n') 3 to 75 by units.

TABLE 3

Analysis of Variance for F-Y Illustration

Term (degree in t)	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean Square (MS)
1	1	157.94	157.94
2	1	267.56	267.56
3	1	3.60	3.60
4	1	6.01	6.01
5	1	2.44	2.44
Remainder	24	579.44	24.14
Total	29	1,016.99	-

The authors' analysis of variance is shown in Table 3. Their general inference from their exercise is:-

> "As will be seen, the first two terms account for a substantial part of the variation, but the mean squares of the remaining three terms are all below the residual mean square. Thus a parabola adequately describes the slow changes."*

The present paper originated in the writer's doubts about this conclusion, especially in the use of the word 'adequately.' As a straightforward point it will be noted that the residual SS is so large that if it were analysed further it might contain constituuents of the same order of magnitude as the SS of the first two terms : as the subsequent investigation

* Op. cit. [2], p.31

shows, this is the case. Comparison of the residual MS of 24.14 with that of terms 3, 4 and 5 suggests that these are abnormally small; in fact the F with (24, 3) DF is 6.01 which is significant at the .10 probability level. Finally, the R^2 of 0.42(=(157.94 + 267.56)/1016.99) is very small.

It would have been revealing to set out the contribution to SS of each of 29 orthopolynomials (in effect to construct a polynomial of degree 29 in t passing through all the 30 time-points). However the computer to which the writer had access had a subroutine for only the first 20 polynomials. These contributions are set out in descending order of magnitude in Table 4, with indication of polynomial degree. In the null-hypothesis case each of these terms would be an estimate of the population variance and, on the population normality presumption, their square roots (with + sign) would be regarded as a random sample of 20 from $N_+(o, \sigma^2)$, the variance σ^2 to be estimated from the data. We shall now apply the foregoing theory to assess in probability the significance of the leading terms.

TABLE 4

Values of $/z_i/$ in Descending Order for Fisher-Yates Application

Term no. (OP degree)	Contribution to SS	$\sqrt{} = /z/$	Term no. (OP degree)	Contribution to SS	$\sqrt{} = /z/$
2	267.5	16.36	4	6.0	2.45
1	158.0	12.57	3	3.6	1.90
16	124.0	11.14	6	3.5	1.87
15	73.2	8.56	18	3.1	1.76
17	48.8	6.99	9	2.6	1.61
10	46.9	6.85	5	2.5	1.58
19	35.1	5.92	8	1.9	1.38
12	17.5	4.18	13	1.5	1.22
20	14.5	3.81	14	0.2	0.45
11	14.1	3.75	7	0.1	0.32

Total Sum Squares (20 DF) above	824.6
Remainder Sum Squares (9 DF)	192.4
Total Sum Squares (29 DF)	1,017.0

NOTE

The SS column kindly supplied by D. Harrington, using an Elliott 803 computer, by courtesy of The Agricultural Institute, Dublin.

In the first place we shall have to pretend that T is 20, and not 30, the original number of observations. It is of interest to observe from Table 4 that the residual MS after removal of 20 terms is 21.38(=192.4/9) almost identical with the Table 3 residual MS of 21.91 after removal of only the 2 leading terms. If we could analyse the Table 3 remainder with 9 DF, a few of the terms might be still of the same order of magnitude of the leading terms shown : a new reordering would be necessary. This is why the problem has to be recast to one of 20 DF.

Tables 1 and 2 enable us to deal with the 6 leading terms. Accordingly let $k = 6$. We arbitrarily impose on the solution the assumption that we are interested only in regressions of 6 or fewer terms. As the significance of the 6 leading terms is in doubt the actual values of their contribution to SS is ignored in estimating s^2 which, from Table 2 and (8) is as follows:-

(9) $$s^2 = 106.2/(20-15.2150)=22.19.$$

Note that 20 has been used for T-1 because of the recasting of the original problem from 30 to 20 terms. The 15.2150 is ΣB_i from Table 2, i = 1, ..., 6. From (7)s = 4.71 which, divided into the leading value 16.36 of Table 4 gives 3.47 so that, from Table 1, this term may be regarded as significant at the .01 value of 3.48 from n = 20. The first six ordered values of z_i/s with their significance by reference to Table 1 for each of the tests A and B are as follows:-

	1	2	3	4	5	6
$/z_i//s$	3.47	2.67	2.37	1.82	1.48	1.45
Test A	**	**	**	*		
Test B	**					

** Significant (or almost so) at .01 probability level.

* Significant at .05 probability level.

The test A inferences were drawn from the full series of ordered statistics in Table 1 for n = 20; with test B we are concerned only with first orders for sample sizes 20, 19, 18, 17, 16, 15 in succession,

derivable by graphical interpolation from Table 1. Test B is the much less sensitive, identifying only the first term, in which case A and B are identical.

In this Fisher-Yates application we would accordingly regard the four leading terms as significant by test A at the .05 probability level. In the total sum squares (for 20 terms) of 824.6 they account for 76% or $R^2 = 0.76$.

It may seem somewhat bizarre to regard the 3rd and 4th terms, of degree in t of 16 and 15, respectively, as included in the regression; it would be less strange if harmonic analysis were involved. It is perfectly sensible to fit linear, or even quadratic and linear exponential terms to time series, however disparate, to enable one to make general statements, e.g. about the rate of increase of the series in time, realising that these statements depend on the type of curve fitted, determined therefore ex ante. In such cases it should be recognised that the residual variance after one or two terms may overstate the time residual variance so that stochastic statements about coefficient significance may be very much on the safe side; e.g. if the tables show the coefficient to be barely significant at the .05 level, the true probability may be considerably less than .05.

While the writer has indicated his preference for test B it must be confessed that it does seem to be over-rigorous and he will not quarrel with those who favour test A. In any particular case there will be no difficulty about applying both tests for samples of size encompassed by the tables in the paper. On commonsense grounds one should favour inclusion of a doubtful variable. If this variable is in fact significant (i.e. has a non-zero coefficient in fact) one has made the right decision. If the variable is not significant the estimate of residual variance, used for assessing the confidence limits of the significant variables, is not biased. The main object of the experiment should be to ensure that all significant variables are included, whether or not these have been exactly identified.

Truth to say, the present problem, in its regression aspect, is rather trivial. The real problem is the identification of significant independent variables in a possibly extended series, related to the dependent variable.

REFERENCES

[1] R. A. Fisher : "Tests of Significance in Harmonic Analysis," Proceedings of the Royal Society of London, Series A, Vol. 125 (1929)

[2] R. A. Fisher and F. Yates : Statistical Tables for Biological, Agricultural and Medical Research, (1957)

[3] E. S. Pearson and H. O. Hartley (Editors), Biometrika Tables for Statisticians, Vol. 1, Second Edition, (1958)

[4] K. Pearson (Editor), Tables of the Incomplete Beta-Function, (1934)