THE ECONOMIC RESEARCH INSTITUTE

Memorandum No. 19

The Nature of Residual Error in the Time Series Context.

By R. C. Geary

The remarks which follow were inspired by the well-known illustration in <u>Statistical Tables for Biological, Agricultural and Medical Research</u> by R.A. Fisher and F. Yates (F-Y).[1] The data in this illustration 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. To these data the authors fit the first 5 orthogonal polynomials, i.e. a polynomial of the 5th degree in t, time (in years). The analysis of variance, given by the authors, is summarized in Table 1.

Table 1. Analysis of Variance For F-Y Illustration

Term	Degrees of Freedom	Sum of Squares	Mean Square	F
1	1	157.94	157.94	7.21
2	1	267.56	267.56	12.21
3	1	3.60	3.60	
4	1	6.01	21.91 { 6.01	
5	1	2.44	2.44	
Remainder	24	579.44	24.14	
Total	29	1,016.99	-	

[1] Fifth Edition (Oliver and Boyd Ltd, Edinburgh and London, 1957).

The final column in Table 1 is mine. Reference to the authors' Table V shows that, with (1,27) degrees of freedom (d.f.) the first term F is significant at the .05 probability level and the second term F is significant at the .01 probability level. Though the point is not important, I do not quite agree with the authors' version of the analysis of variance in their combining the first five terms' contribution for the purpose of establishing mean square with 5 d.f., because the constituents are so different in value. My main concern is with the authors' general inference from their exercise :-

> "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".[2]

On commonsense grounds alone the last sentence of the quotation is of doubtful validity. We note, in fact, that while the contributions of the "negligible" 3rd, 4th, 5th terms are respectively 4, 6, 2 the remainder mean square averages 24. We must suspect - and our suspicion will be proved to be correct - that the remainder contains terms whose contribution to SS is sizable, in fact of the same order of magnitude of the significant contributions of the 1st and 2nd terms. The F with (24, 3) d.f. for (remainder, terms 3 - 5) is 6.01 (= 3 x 24.14/(3.60 + 6.01 + 2.44)) which is significant at the .01 probability level. The authors' idea of "adequacy" will not coincide with that of most workers in this field if only because the R^2 of the first two terms regression has a value of only 0.4184(=(157.94 + 267.56)/1016.99).

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

While it is easy to criticise the authors' treatment it is much more difficult to suggest a remedy which is satisfactory in stochastic terms, or even, indeed, to propound the problem at all. We shall try to do so by continuing to study the F-Y illustration. In the first place, it may be remarked that, using 29 orthogonal polynomials, i.e. deriving a polynomial of degree 29 in t a function may be derived which will pass through all the observed points. A glance at the appended diagram showing the vast dispersion of the observations indicates that this would not be a useful exercise, if what we have in mind is the derivation of a law of relationship between the observations and time t. It would, however, be revealing to set out the contributions of each of the 29 orthogonal polynomials to the aggregate sum of squares. Twenty of these with a a remainder are shown in Table 2 [3].

Table 2. Contribution to Sum Squares of Each of Twenty Orthogonal Polynomial Terms to Total Sum Squares in F-Y Illustration.

Term No.	Contri-ution to SS	Term No.	Contri-ution to SS
1	158.0	11	14.1
2	267.5	12	17.5
3	3.6	13	1.5
4	6.0	14	0.2
5	2.5	15	73.2
6	3.5	16	124.0
7	0.1	17	48.8
8	1.9	18	3.1
9	2.6	19	35.1
10	46.9	20	14.5
Remainder (9 d.f.)			192.4
Total Sum Squares (29 d.f.)			1,017.0

[3] These were produced on the Elliott 803 Computer of the Agricultural Institute, by courtesy of the Director, Dr T. Walsh, and with the cooperation of Mr D. Harrington.

Each of the terms has one d.f. The term number represents the degree in t of the polynomial so that, in effect, a polynomial of degree 20 in t has been fitted to the 30 observations. Apart from rounding-off deviations the first 5 terms of Table 2 are, of course, identical with those of F-Y given in Table 1. As anticipated, the contributions of some of the subsequent terms are large, the largest being that of term no. 16, namely 124.0.

Stochastic Interlude

At this point it may be appropriate to make a few general remarks on testing for significance in curve fitting to time series using the regression method. The whole exercise is based on the assumption that there is an inherent relation between the sequence of observations and time t, disturbed in greater or lesser degree by an error term which initially or ultimately (i.e. after certain transformations) is assumed to be a random variable (i.e. random as regards t) with certain stochastic characteristics, e.g. that the sequence is a normal sample with mean zero and estimable variance. In an earlier paper [4] the writer has given his opinion that the whole object of regression is to enable one to estimate the value of the dependent variable from given values of the independent variables, in the present case the known values of the orthogonal polynomials adjudged significant : we may, for instance, be

[4] "Some Remarks About Relations Between Stochastic Variables : A Discussion Document" by R. C. Geary, Review of The International Statistical Institute, Vol 31 : 2, (1963).

interested in forecasting by extrapolation. The practical value of the operation will, therefore, depend on the magnitude of the residual, or error, standard deviation in relation to the changes which one is trying to forecast; and experience has shown that values of R^2 of even the .99 variety may result in confidence limits of uselessly wide range. Of course, as an exercise in analysis for its own sake there may be some theoretical interest in being able to state that a given time series is e.g. a "(2, 8, 16; .96)" meaning that it is significantly and completely explained by terms 2, 8, 16 of a specified orthogonal series with $R^2 = .96$ and recourse must be had to sophisticated statistical procedures to enable one to make such a statement. The whole object of statistical science is to describe possibly very numerous sets of figures and their relationships in terms of a few estimable paramteters.

By "adequacy" in the foregoing quotation F-Y may mean what the author has called <u>completeness</u> [5] of relationship, (whereby in time series analysis all the significant independent variables have been identified and the residual is non-autoregressive). Even if the R^2 is small (say .4 as in the F-Y illustration), circumstances can be envisaged in which the result would have some practical value. Imagine a manufacturer of a highly perishable, even ephemeral, product (ice-cream ?!) working on a day-to-day basis, manufacturing his day's supply in the early morning. He cannot keep

[5] "Determination Of Linear Relations Between Systematic Parts of Variables With Errors Of Observation The Variances Of Which Are Unknown" by R. C. Geary, <u>Econometrica</u>, Vol. 17 No. 1 (1949).

stocks overnight without deterioration of his product. He notices that demand varies considerably from day-to-day. From the supply side he is reasonably satisfied with his annual production. Initially unable to anticipate daily sales, he produces the same quantity of his product each day, philosophically accepting his losses. He may ruefully calculate the difference between what his profit (given total annual production) would be if he had been able to forecast exactly each day's demand and what his profit actually has been. He consults a statistician who finds a significant correlation between his actual daily sales and some factor, say temperature at the time daily manufacture starts. This correlation need not be very high (say .6) for him to base his daily production policy on the regression formula with improvement in profits. Of course, the statistician will recognise that sum squares of deviations is not the function which he should minimize but the sum of the absolute value of the deviations. He is sustained by the conviction that his least square regression procedure will give him an answer which will improve profits, even if he does not know the optimal formula.

We must now distinguish between what may be termed (a) specific and (b) general hypotheses. All the well-known theory of regression, including estimation of testing for significance of coefficients, is based on specific hypotheses. This is the situation in which on past experience, the results of other researches etc. or from plain commonsense or knowledge we may write down a plausible relationship and estimate and discuss the estimates of the parameters involved. With general hypotheses we have no such guide; here we set out to discover the terms (or series) which are significant with no prior knowledge of the forces at work, painfully aware of the hazards of nonsense correlation [6], especially rife in time series : for example any two economic series increasing in time will, on crude analysis, be found to be highly correlated. It goes without saying that general hypotheses are far more difficult to deal with than are special hypotheses.

Continuation of Study of F-Y Illustration :

We note at once from Table 2 that the residual mean square after 20 terms is 21.38 almost identical with the 21.91 (= 192.4/9) after the removal of 2 terms ! We begin to suspect that the inherent error variance of the system may be of about this magnitude despite the proliferation of quite small numbers : not fewer than 9 of the 20 terms have a contribution less than 4. The problem confronting us appears to be this : can we discover any clear break in the series which will enable us to state confidently that certain specified terms should be included in the regression while the rest are to be deemed included in the error term ?

The foregoing remarks as to the remainder after 5 terms applies to the remainder after 20 terms, namely that, if only we knew them , we might find one or more sizable contributions to sum squares for terms 21 - 29. In default of this information - the computer

---

6 "Why Do We Sometimes Get Nonsense-correlation Between Time Series - A Study in Samping And the Nature of Time Series" By G. U. Yule, Journal of the Royal Statistical Society 89 : 1 (1926).

had a programme for only 20 terms - the best course appears to be to pretend that we are dealing with a problem of 20 (and not 29) terms. So total sum squares is now deemed to be 824.6(=1,017.0 - 192.4) instead of the original 1,017.0.

In Table 4 the 20 contributions are arrayed in descending order of magnitude with term number indication.

Table 3. Data of Table 2 in Descending Order of Magnitude with Standard Deviations (SD)

Term No.	Contribution to SS	√ = SD	Term No.	Contribution to SS	√ = SD
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 d.f.) 824.6

In the null-hypotheses case, when the 20 original observations are arrayed in random order, each of the 20 terms is an estimate of the population variance, the population mean being zero. The √'s (with + sign) would, therefore, be a random sample of 20 from the positive side of the population frequency distribution. As is usual we make the assumption for what follows that the populations from which samples are drawn are normal, an aspect dealt with later.

We shall now try systematically to find a break in the sequence of Table 3 enabling us to identify stochastically the terms which are significant. The method will be to study points in the sequence starting at the bottom at which the jumps are improbable on the null-hypotheses. We shall first have to study the <u>Distribution of the Highest Value in a Normal Sample</u>.

We are concerned only with the positive side of the standard normal distribution table [7]. If the cumulative frequency from 0 to x of any continuous distribution is F(x) the cumulative frequency of the largest member - we deal only with non-negative measures - of a sample of n is $[F(x)]^n$. A particular probability level is selected, say .95, and the following equation is solved for x :-

$$(1) \qquad [F(x)]^n = 0.95 .$$

If the top sample value at any stage is greater than the solution x we shall infer significance for this term and all terms with greater values : at the .95 probability level we shall have succeeded in breaking the sequence of 20 terms into two parts, a significant part and a residual error part.

The .95 normal probability points (population SD unity) for top sample values x_{hn} for a certain range of values of n are shown in Table 4.

---

[7] "Biometrika Tables for Statisticians" (Ed. E.S. Pearson and H.O. Hartley) Vol. I, Second Edition, 1958.

Table 4. Values of Normal .95 Probability Points x_{hn} and Median Values x_{mn} for Top Elements in Samples of n.

n	x_{hn}	x_{mn}	n	x_{hn}	x_{mn}
2	2.24	1.05	15	2.93	2.01
3	2.39	1.26	16	2.95	2.03
4	2.49	1.41	17	2.97	2.05
5	2.57	1.52	18	2.98	2.08
6	2.63	1.61	19	3.00	2.10
7	2.68	1.67	20	3.02	2.12
8	2.73	1.73			
9	2.76	1.79	25	3.09	2.21
10	2.80	1.83	30	3.14	2.28
11	2.83	1.87	35	3.18	2.34
12	2.86	1.91	40	3.22	2.38
13	2.89	1.94	45	3.25	2.42
14	2.91	1.98	50	3.28	2.46

Attention is now directed to the arrows in Table 3. These mark the suspect breaks in the sequence, as indicated by the jumps between consecutive values of the variances (or S.D.'s) : the arrows are placed above the suspect values. Thus the problem poses itself : in a normal sample of 3, consisting of 0.32, 0.45 and 1.22, is the top value of 1.22 stochastically acceptable ? An analogous problem presents itself in the variance jump from 3.6 to 6.0, the sample size now being 10.

Estimate of Population Variance

We are now confronted with the difficulty that, to apply normal theory, we require to know the population variance which, of course, will be different at each arrow stage. Having selected the break points from observation of the sample values themselves the appropriate variance at the first test break cannot be estimated as

$$s_3^2 = [(0.32)^2 + (0.45)^2 + (1.22)^2]/3 = 1.8/3 = 0.6.$$

Since the top value is suspect of being too high this estimate is biassed upwards. Neither can the variance be estimated as the sum of the last two values divided by 2 since this would be an under-estimate : one cannot leave out the top value of a sample and estimate the variance from the remaining values simply by omitting it! The simplest course would appear to be to substitute for the top suspect value the median top to be expected from a normal sample of given size.

Let s_n^2 be the estimate of the population variance for sample size n and x_{mn} the median top value to be estimated as the solution of

(2) $$[F(x)]^n = .50,$$

where F(x) is, as before, the cumulative one-sided normal frequency, population variance unity.
Then set

(3) $$\sum_{i=1}^{n-1} x_i^2 + x_{mn}^2 s_n^2 = n\, s_n^2$$

or

$$(4) \qquad s_n^2 = \left(\sum_{i=1}^{n-1} x_i^2 \right) / (n - x_{mn}^2),$$

where the x_i^2 are the actual values shown in Table 3. The values of x_{mn} are shown in Table 4.

The final stages of the calculation are shown in Table 5.

Table 5. Test of Significance of Apparent Breaks in Sequence in the F-Y Illustration

n	s_n	x_n	x_n/s_n	a
1	2	3	4	5
3	0.461	1.22	2.65	0.977
10	1.693	2.45	1.45	0.898
11	1.803	3.75	2.08	0.896
14	2.656	5.92	2.23	0.848
17	3.972	8.56	2.16	0.802
18	4.486	11.14	2.48	0.792
19	5.230	12.57	2.40	0.772
20	5.994	16.36	2.73	0.759

Notes

Col.2 : From formula (4); e.g. n = 10; Σ = 19.0 (count of last 9 items in SS column, Table 4)

Col.3 : E.g. n = 10, x_{10} = 2.45 is 10th value from end of SD column, Table 4.

Col.5 : a is test of normality[8] (or ratio mean deviation to standard deviation) applied to "residuals" at each n stage; e.g. n=10, sample is last 9 items in SD column of Table 4 together with x_{mn} (=1.83) from Table 5.

[8] Tests of Normality by R. C. Geary and E. S. Pearson, (1938).

Comparison of the column 4 figures in Table 5 with the appropriate x_{hn}'s in Table 4 shows that, at the 95% probability level only the value, 2.65, is significant for n = 3 while x_{hn} = 2.39; and this finding is more than dubious since the estimate of the variance for the application of normal theory is based on a sample of 2 !; in any case, no interest attaches to a regression which allegedly contains 17 significant terms. In the discussion that follows, no reference is made to the n = 3 entries in Table 5.

At the other end of Table 5 for n = 20 we are testing whether the single quadratic orthogonal polynomial affords a complete representation of relationship, the remaining 19 terms being collectively a random residual. While the column 4 value of 2.73 falls short of the .95 probability value of 3.02 it is the highest in the table and, if a break is to be identified in the sequence, this is it. There is no good reason for making a break after the second (or linear term in t) (as F-Y do) than there is in including also the third term, in the orthogonal polynomial of degree 16 in t (see Table 3), however repugnant to our habits of thought and procedure in time series regression. Of course no claim can be made that the technique developed here is in any sense the most efficient for logically dividing the sequence of terms into the two classes significant and residual. A technique of greater sensitivity might identify the second term as significant; but in such case one might fairly surmise that it would also include the third term, however _a priori_ unlikely.

Normality. Throughout an attempt has been made to play the game according to the rules and one of these is that, if a break is made, the constituent items in the residual are not only random but normally distributed. From chart A [9] it will be observed that none of the value shown in column 5 of Table 5 are significant of non-normality as the values all lie between the upper and lower 10% probability limits of a, on the hypotheses of universal normality.

Auto-regression. According to the systematic procedure outlined in Memorandum No. 15 time series regression should start with establishing that, in probability, the original series was auto-regressive. The Von Neumann test Q (defined in formula (1) of Memorandum No. 15) affords no such assurance. The original value of Q is 1.46 which, while less than the mean value 2 is not significantly so, since the 95% probability value is about 1.30 [10] for n = 30. Therefore auto-regression cannot be inferred and there is no justification for starting the regression process at all. After removing the principal term (i.e. the quadratic orthogonal polynomial) the value of Q is 1.96, not significant. On removal of the two principal terms the value of Q is 2.47. It is true that arithmetically there is

---

[9] Op. cit [8].

[10] This value is based on randomization procedure (or non-parametric, whereby inferences may be made without the assumption of universal normality). The necessary formula are as follows :-

$$M(Q) = 2$$

$$\text{Var } Q = M(Q^2) - M^2(Q) = 2(2n-\beta_2-3)/n(n-1),$$

effected a regular trend towards the hypotheses of non-autoregression in these three values (1.46, 1.96, 2.47) but none is significantly different in the stochastic sense from the mean value 2. The Von Neumann analysis repeats what is virtually the conclusion of the earlier analysis, namely that there is little but randomness in this material. It is hoped, however, that the technique expounded here for the <u>ex post</u> derivation of significant terms in regression analysis may prove more useful with less recalcitrant material.

<u>Footnote No. 10 continued</u>

where $\beta_2 = M_4/M_2^2$, M_k being the kth moment from the mean of the original data. These formulae were derived from formulae in 'The ContiguityRatio And Statistical Mapping' By R. C. Geary, <u>Incorporated Statistician</u>, Vol. 5 : No. 3, (1952). It is extremely interesting that the coefficient of β_2 is $O(n^{-2})$. When n is not too small β_2 can safely be given its normal value 3 so that

$$\text{Var } Q \sim 4(n-3)/n(n-1)$$

the value used in the test. There would, however, be no difficulty about using the exact value if meticulousness were deemed necessary.

Conclusion

Undeniably *ex post* identification of significant independent variables in time series orthogonal regression presents its particular problems, towards the solution of which the simple techniques outlined in the paper may seem worth trying out. If the writer's submission[11], namely that the object of regression is estimation of the dependent variables, then no effort must be spared in reducing the residual variances. This, in turn, will entail inclusion of a far more numerous set of independent variables in the future than in the past, experimentally to start with. Though we may not realise it, the sparsity of independents has probably been influenced by (a) the amount for computation involved with only desk machines available and (b) our preconceived ideas of the identity of the independents. As to (a), let us realise that the electronic computer, with its subroutines, has arrived. As to (b), let us realise, in humility, that at the start we did not know as much as we thought we knew. At the same time if a significant independent turns up rather unexpectedly in the analysis it will be prudent to try to rationalize its inclusion.

In ordinary regression an indefinitely large series of independents will not be available, in this respect differing from the kind of time series dealt with in this paper. We can, however, be more expansive than we have been prone to be, even if many of the independents are to be rejected later, as insignificant.

---

11 *Op. cit* [4].

They will have served their purpose in helping to establish an estimate of the true residual variance. There, will, therefore, be two hypothetical elements in the hypothetical residual SS (each with its DF) (a) the contribution of the experimental but rejected independents and (b) the final residual. Only when the ratio of the two MS's is indubitably insignificant should the analysis be regarded as completed.

****

Difference in yields (bushels per acre) on two plots of wheat, 1855-1884. Actual difference and fitted regression quadratic in time. Data Source F-Y.