15

JOINT EUROPEAN CONFERENCE OF THE ECONOMETRIC SOCIETY

AND THE INSTITUTE OF MANAGEMENT SCIENCES

(Zurich, 7-11 September 1964)

The Inefficiency of the Von Neumann Ratio in Time Series Regression

by

R. C. Geary
The Economic Research Institute, Dublin.

****

The Inefficiency of the Von Neumann Ratio in Time Series Regression

by

R. C. GEARY

The Economic Research Institute, Dublin.

The Von Neumann ratio in the form

$$(1) \qquad Q = \sum_{t=2}^{T} (v_t - v_{t-1})^2 / \sum_{t=1}^{T} v_t^2$$

is much used for assessing completeness of representation (or goodness of fit) of the regression estimate to equally-spaced data in time series analysis. The v_t, $t = 1, 2, \ldots T$, are the regression residuals pursuant to fitting a $K + 1$ - term regression to data, i.e. the regression is

$$(2) \qquad y_t = b_o + \sum_{k=1}^{K} b_k x_{kt} + v_t, \quad t = 1, 2, \ldots, T,$$

the y_t being the observations, the x_{kt} a given set of independent variables, all strictly functions of t. The problem is to determine whether (2) affords an adequate representation of a relationship of the form

$$(3) \qquad y_t = \beta_o + \sum_{k=1}^{K} \beta_k x_{kt} + u_t, \quad t = 1, 2, \ldots, T,$$

where the u_t are entirely random (or, in fact, non-autocorrelated) residuals with mean zero and variance constant for all t. The point is that at the start we do not know the number of terms K to be used in the regression (2) though we assume that we have available

for use an indefinitely large set of functions x_{kt}, polynomials, Fourier series or the like. Of course, the number of these functions required is less than $T - 1$: if equal to $T - 1$ the fitted curve would pass through all the observed points, which is not helpful. Usually one hopes that only a few terms (not more than, say, five or six) will be required and that we have sound a priori reasons for selecting these. We cannot attach much value to, or have much confidence in, the representational power of a formula (2) in which it is necessary to use many terms, even if the coefficients of these are significantly different from zero. We may infer in such case that the series x_{kt} we are using is unsuitable.

The characteristic feature of time series is that the successive terms are autocorrelated i.e. that the correlation coefficient for the pairs $(y_{t-1} > y_t)$, $t = 2, 3, \ldots, T$, is significantly greater than zero. The systematic procedure on Von Neumann (VN) lines would therefore be first to establish that the original y_t are autocorrelated at the .05, .01 etc probability levels using tables prepared by J. Durbin and G. S. Watson [1] and by H. Theil and A. L. Nagar [5]. If the calculated value of Q given by (1) with y_t substituted for v_t is not significantly different from its expected value on the nul-hypothesis, i.e.

$$(4) \qquad BQ = 2(T - 1)\ \alpha^2/(T - 1)\ \alpha^2 = 2$$

when the v_t (or, in this initial case, the y_t) are normally distributed, the successive y_t are presumed

to be like those which would be found on successive drawings at random from a normal universe. In such case the y_t would be regarded as independent of t. Its only representation could be $y_t = \bar{y}$.

If from the tables the value of Q is deemed significantlylow at some probability level one selects the first term x_{1t} of the predetermined series, establish the simple regression of y_t on x_{1t} and hence derive the residuals v_t. The value of Q according to (1) is then calculated and the significance determined from the probability table, now with two degrees of freedom (d.f.). If not significant the process is ended; if significant, another function x_{2t} is added, and so on.

What we try to discover in the present paper is the sensitivity of the Von Neumann ratio as applied in the manner outlined above. The method used is deliberately to falsify the hypothesis of residual non-autoregression and try to determine how large the falsification has to be before we find ourselves out, given the number of observations T and the probability level of non-acceptance of the nul-hypothesis.

We consider a simple case. Suppose the true relation is

$$(5) \qquad y_t = \beta_o + \beta_1 x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t,$$
$$t = 1, 2, \ldots, T, \ \beta_1, \ \beta_2, \ \beta_3 \neq 0,$$

where the u_t are randomly distributed with mean zero and standard deviation independent of t which, without loss of generality, may be assumed to be unity. The

reason for introducing the extraneous term in x_{3t} will be apparent later. Also without loss of generality the functions x_{1t}, x_{2t} and x_{3t} may be regarded as orthogonal, i.e.,

$$(6) \qquad \sum_{t=1}^{T} x_{kt}x_{k't} = 0; \quad \Sigma x_{k't} = 0, \; k, \; k' = 1, \; 2, \; 3, \; k \neq k'.$$

We require the following additional notation :-

$$(7) \quad \begin{cases} (a) \; u'_t = u_t - u_{t-1}; \; v'_t = v_t - v_{t-1}; \; x'_{kt} = x_{kt} - x_{kt-1} \\ (b) \; X_k = \Sigma x_{kt}^2 \; ; \; X'_k = \Sigma x'^2_{kt}, \; t = 2, \; 3, \; \ldots, \; T. \\ (c) \; Y_k = X'_k/X_k \end{cases}$$

As the set of x_{kt} is known so also are the X_k, X'_k and Y_k.

In error we now try to approximate (5) by the regression

$$(8) \qquad y_t = b_o + b_1 x_{1t} + v_t.$$

From (5), (6) and (8), using (7), we then have

$$(9) \quad \begin{cases} b_o - \beta_o = \Sigma u_t/T \\ b_k - \beta_k = \Sigma x_{kt}u_t/X_k. \end{cases}$$

Formulae (9) are quite general. For the moment we require only k = 1. The deviation v_t is given by

$$(10) \qquad v_t = (\beta_o - b_o) + (\beta_1 - b_1)x_{1t} + \beta_2 x_{2t} + \beta_3 x_{3t} + u_t,$$

which, using (9), is

$$(11) \qquad v_t = (u_t - \bar{u}) - x_{1t}\Sigma_t, x_{1t}, u_t, / X_1 + \beta_2 x_{2t} + \beta_3 x_{3t};$$

whence

$$(12) \qquad v'_t = u'_t - x'_{1t}\ \Sigma_t, x_{1t}, u_t, / X_1 + \beta_2 x'_{2t} + \beta_3 x'_{3t}.$$

We now introduce the notion of the <u>representative</u> value of Q, namely $\bar{Q}$, given by

$$(13) \qquad \bar{Q} = E \Sigma v_t'^2 / E \Sigma v_t^2,$$

where E is the expected value pursuant to random variation in variables u_t. $\bar{Q}$ is not necessarily equal to EQ though it will be close to it unless T is small. Using the assumed properties $Eu_t = 0$, $Eu_t^2 = 1$, $Eu_t u_{t'}$ $(t \neq t') = 0$ it can easily be shown that

$$(14) \qquad E \Sigma v_t^2 = T - 2 + \beta_2^2 X_2 + \beta_3^2 X_3$$

$$(15) \qquad E \Sigma v_t'^2 = 2(T - 1) - X_1'/X_1 + \beta_2^2 X_2' + \beta_3^2 X_3'.$$

It may be noted that, in deriving (14) and (15), and hence (13), it has not been necessary to use the property that the u_t are normally distributed. To use the probability points table which are based on the hypothesis of universal normality we have to pretend that we think that the residual v_t are independently and normally distributed.

In [5] the probability points are shown for .01 and .05 for certain values of T up to 100 and for 2, 3 ... 6

coefficients adjusted (two in our case). Let λ be the value appropriate to the probability level to which we are working. The critical value of β_2, namely B_2, is then found by setting

$$\bar{Q} = \lambda \tag{16}$$

or, using (13), (14) and (15),

$$B_2^2 X_2 = \frac{(2 - \lambda)T + 2(\lambda - 1) - Y_1 - \beta_3^2 X_3 (\lambda - Y_3)}{(\lambda - Y_2)}, \tag{17}$$

where the X_k and Y_k are given by (7). If the true value of β_2 is less than as shown by (17) we shall wrongly decide that the value is zero, that the process should stop. The value shown is therefore a measure of the sensitivity of the VN ratio in this simple case, It is clear that as $\beta_3^2 X_3$ increases the efficiency, given T, of VN increases since the Y_k will usually be small. This is to be expected since the larger $\beta_3^2 X_3$ is, the less the relative influence of u_t in the residual v_t, given by (10), which tends to assume functional form with a necessarily small value of the VN ratio. Since $B_2^2 X_2$ is necessarily positive a zero or negative value for a high value of $\beta_3^2 X_3$ will be interpreted as equation (16) having no solution in B_2, a case of no practical importance.

It may be worthwhile placing on record a generalised version of formulae (14) and (15) and hence of $\bar{Q}$. Suppose that, instead of having one coefficient (apart from the constant) in the regression (i.e. b_1) we had K', and therefore K - K', in the "false residual" v_t. Then it may be shown that

$$(18) \qquad E\Sigma v_t^2 = T - K' - 1 + \sum_{K'+1}^{K} \beta_k^2 X_k,$$

$$(19) \qquad E\Sigma v_t^2 = 2(T-1) - \sum_{1}^{K'} Y_k + \sum_{K'+1}^{K} \beta_k^2 X_k Y_k.$$

Of course the orthogonal property is assumed in all the functions of x_{kt} whether in the regression or not. The use of orthogonal functions imparts a great simplicity to work of this kind.

Analysis of Variance (AV)

This method is also much used for assessing completeness of representation in time series analysis. We set up a regression

$$(20) \qquad y_t = b_o + \sum_{k=1}^{K'} b_k x_{kt} + \sum_{k=K'+1}^{K} b_k x_{kt} + v_t.$$

All the x_{kt} are still orthogonal functions. The problem is to establish the sequential randomness of the residuals v_t. This is done by identifying the K' significant functions in the first Σ on the right. The K - K' terms in the second Σ, all deemed insignificant, is arbitrary. The analysis of variance schema [2] is on the following lines :-

Group	Degrees of freedom (d.f.)	Sum squares	Mean square	F-ratio
First (K') terms	K'	S_1	$M_1=S_1/K'$	$F_1=M_1/M_3$
Second (K-K') terms	K-K'	S_2	$M_2=S_2/(K-K')$	$F_2=M_2/M_3$
Residual terms	T-K-1	S_3	$M_3=S_3/(T-K-1)$	-
Total	T-1	S	-	-

If F_1 is significant and F_2 is not significant at the selected probability level the K' function regression will be regarded as complete and the remaining K - K' terms ignored as adding nothing to our knowledge of the relationship. There are conceptual difficulties with this approach which will be briefly referred to later. In the final section of the paper, it is proposed (as in [2] but on a larger scale) that in the first and second groups the contribution of each term should be set out individually, i.e. each with one d.f.

In the three function (5) case, how large has the coefficient β_2 to be to enable us to reject the hypothesis that (5) is represented by the regression

$$(21) \qquad y_t = b_o + b_1 x_{1t} + b_2 x_{2t} + v_t \ ?$$

we set up the analysis on the lines indicated in the foregoing schema with K' = 1 and K = 2. From (9) the sum squares and hence the mean square M_2 (since there is one d.f.) is

$$(22) \qquad M_2 = (\beta_2 X_2 + \Sigma x_{2t} u_t)^2 / X_2,$$

from which

$$(23) \qquad EM_2 = \beta_2^2 X_2 + 1.$$

from (18), (with K' = 2 and K = 3)

$$(24) \qquad EM_3 = (T - 3 + \beta_3^2 X_3)/(T - 3).$$

Analogous to the procedure in the VN case we equate the quoteint of (22) by (23) (to form a typical value of F_2 in the schema) to the appropriate probability point λ' in the F-probability table [5] from which the critical value B_2' of β_2 is found from

$$(25) \qquad B_2'^2 X_2 = [(\lambda' - 1)(T - 3) + \lambda' \beta_3^2 X_3] / (T - 3).$$

Other things being equal, the larger β_3^2 the larger B_2' and therefore the less efficient the AV method, in direct contrast with the VN test. It has long been recognised as a weakness in the Av method as applied to multivariate regression that a low value of F_2 (see schema) may be due as much to a high value of M_3 as a low value of M_2. As will presently be shown the AV method is, however, far more efficient than the VN method for dealing with the present problem. At the same time we must not be blind to the hazards of AV : this is why the term in x_{3t} has been introduced into (5). It may be added that, since when β_3 is zero $Eb_3^2 X_3$ is unity; consequently $\beta_3^2 X_3$ may be regarded as O(1) in regard to T when β_3 is not zero.

Von Neumann versus Analysis of Variance

The decision with which we are faced is, having set up a simple regression, i.e. of form (8), do we stop or do we go on ? To stop would be a wrong decision (since it has been assumed that β_2 is different from zero): to go on would be the right decision. At a given significant probability level the test must be favoured which yields the smaller

regression estimate of β_2. The statistics B_2 and B_2' (given by (17) and (25)) are near-average estimates of what has been termed the critical value of β_2, i.e. if β_2 were less than these respective values in absolute terms, β_2 would be deemed zero and a wrong decision made.

We accordingly introduce the efficiency ratio H defined by

$$H^2 = B_2'^2/B_2^2, \tag{26}$$

H being non-negative.

Asymptotic efficiency will be considered in the first place. All the $\beta_k^2 X_k$ and $\beta_k^2 X_k'$ involved are ordinary magnitudes i.e. O(1) in T as are λ and λ'. It is therefore evident from (17) that $B_2^2 X_2$ is O(T) while, from (25), $B_2'^2 X_2$ is O(1). Accordingly H, from (26), is $O(T^{-\frac{1}{2}})$. When T is indefinitely large the efficiency of VN, compared with AV is zero. However, we have ordinarily to deal with T in something like the range 10 - 100 where, as will appear, the advantage in all circumstances is by no means so overwhelmingly in favour of AV. A few particular applications will now be considered.

Orthogonal polynomials. The following Table shows the value of relative efficiency H for a series of values of T (assumed odd for arithmetical convenience) and for five values of $\beta_3^2 X_3$, namely 0, 2, 5, 7, 10. The polynomials involved, x_{1t}, x_{2t} (see (5)) are respectively of degree one, two and three in t. As regards $B_2^2 X_2$ we require only

the ratios $Y_k = X'_k/X_k$, k = 1, 2, 3 - see (17). The X_k were derived from Table XXIII of [2] and the X'_k were calculated by the writer from data in [2]. As the Y_k are $O(T^{-2})$ they become very small as T increases. Indeed they may be ignored for T > 20. The λ were derived from [5], Table 2 and the λ' from [2], Table V, each for probability .05.

Table. Comparative Efficiency E of VN and AV for Certain Values of T and of $\beta_3^2 X_3$ at Probability Level .05 using Orthogonal Polynomials for Fitting

T	$\beta_3^2 X_3$					1% AV prob. point for T-3 d.f.
	0	2	5	7	10	
11	.696	.843	1.086	1.277	1.650	11.3
21	.584	.688	.906	1.146	2.123	8.3
31	.528	.602	.764	.934	1.504	7.6
41	.485	.542	.658	.772	1.078	7.3
51	.458	.505	.599	.688	.904	7.2
101	.390	.417	.467	.509	.594	6.9
∞	0	0	0	0	0	6.6

As expected the larger the value of the undetected disturbance in the residual (i.e. the $\beta_3 x_{3t}$ term) the less favourable the comparison for AV. However, the general verdict must be overwhelmingly against VN in this, the polynomial, case. Indeed, the apparent superiority of VN as indicated by the values of E greater than 1 in the top right part of the Table is illusory. As shown by the 1% points in the final column, an analysis would be incompetent which failed to detect so very large a term in β_3 in the error residual. As indicated later, proper AV

procedure can guard against this eventuality.

Fourier terms. Here the analysis can be much more general than in the preceding case because of the well-known summation properties of the Fourier terms. Let the given Fourier series be

$$x_{kt} = \cos \frac{2\Pi \alpha_k}{T}(t + \gamma_k), \tag{27}$$

where the α_k are given positive integers less than T and the γ_k are any real constants, also given. We then have

$$X_k = \sum_{t=1}^{T} x_{kt}^2 = T/2$$

$$X_k' = \sum_{t=2}^{T} x_{kt}'^2 = (1 - \cos \frac{2\Pi \alpha_k}{T})[T - 1 + \cos \frac{2\Pi \alpha_k}{T}(2\gamma_k + 1)] \tag{28}$$

or

$$Y_k = X_k' / X_k \sim 2(1 - \cos \frac{2\Pi \alpha_k}{T})$$

for T not small. As in the orthogonal polynomial case H is $O(T^{-\frac{1}{2}})$ so that the asymptotic efficiency of VN compared with AV is zero. The main difference between the polynomial and Fourier cases is that, as indicated by (28), the fundamental ratios Y_k with Fourier independents can no longer be regarded as tending to zero with T. Furthermore, from (17) a situation can arise where it may be impossible to equate, as in (16) the representative value of the VN ratio to the probability point (given T). In fact, (28) shows that the value of the denominator squared in (17), namely $(\lambda - Y_2)$, need not necessarily be positive. It will be positive

only when λ is large or (from (28), second formula) α_k/T is sufficiently small. To set up an efficiency test, given T and therefore λ, α_k must be so small that

$$\lambda \geq 2\left(1 - \cos \frac{2\pi\alpha_k}{T}\right)$$

(30)

$$\cos \frac{2\pi\alpha_k}{T} \geq 1 - \frac{\lambda}{2}$$

when T is not too small. For probability .05 and T ranging from 25 to 40 (see [5].) (the kind of range usually dealt with) λ is about 3/2. Accordingly, from (30), the ratio α_k/T cannot exceed 0.2.

The Fourier case may be studied from a viewpoint different from that of orthogonal polynomials. We consider the value of β_3 (the critical value, namely B_3) required to make VN and AV equally efficient, i.e. H = 1, or from (26), we equate the right sides of (17) and (25) and solve for $\beta_3^2 X_3$. Noting that λ, λ', Y_k (k = 1, 2, 3) are all ordinary magnitudes, i.e. O(1) in T, we find

$$B_3^2 X_3 \sim \frac{(2 - \lambda)T}{\lambda - Y_3} \tag{31}$$

when T is not small. Now $Y_3 \geq 0$ so that the minimum value of $B_3^2 X_3$ is

$$(B_3^2 X_3)_{min} = \frac{(2 - \lambda)T}{\lambda} \sim \frac{T}{3}. \tag{32}$$

In general, in order that VN and AV should be equally efficient, $B_3^2 X_3$ should be O(T), say approximately T/3. A glance at the last column of the Table will show that this condition entirely disqualifies VN when T exceeds, say, 20. Except as regards the last few remarks the discussion in this paragraph is general, and not confined to Fourier independents.

Conclusion

The insensitivity of VN having been established, we must consider some alternative treatment. Clearly this treatment should be on AV lines. Incidental to the main purpose of the paper, it appears that AV can also mask residual autoregression : this was the object of introducing the term in x_{3t}.

The sensible course would appear to be to set up a regression in considerably more individual orthogonal terms than we have been accustomed to do in the past, in fact, the number of terms should be of the order of T. With orthogonal independents and even with only a desk machine there will be little difficulty in calculating the coefficients b_k and hence in setting up an AV for single terms, each term having one d.f. with relatively few terms in the remainder. Then the F_k for the individual terms (see schema) would be arranged in ascending order of magnitude and a selection made of the top terms representing estimates of the variance of u_t. As suggested earlier, unless these be many (i.e. the significant terms few), much confidence cannot be reposed in the representational power of the series x_{ky} selected. Of course, the F-table as it stands cannot be used for assessing significance. It is a nice problem to decide on stochastic lines where the division should be drawn between terms to be deemed respectively insignificant and significant at say, .05 and .01 probability levels. It might be well to prepare a table on such lines for different values of T, number of significant terms and a few

probability levels.

In [4] the author suggested that statisticians should not be satisfied with residual randomness as indicating that their task was completed, even when the test of randomness is efficient, which the VN certainly is not. In that paper it was suggested that specific attention should be given to the reduction of the residual variance as a problem in its own right.

All the foregoing considerations apply to multiple regression generally and not only to time series. There is this important difference, however, that in general regression we may not have an indefinitely extended series of independents available. If, using a short series, as has been so often the case in econometric analysis in the past, residual variance is found to be unsatisfactorily large, other variables or relationships should be sought. The full set of independents should then be orthogonalised using perhaps the latent vector method [3] which has the advantages of being symmetrical and unique. Furthermore, most computer companies have programme sub-routines to do the necessary calculations. One can always transform one's orthogonal solution back to the original independent variables.

****

References

[1] Durbin, J. and Watson, G. S., "Testing for Serial Correlation in Least Squares Regression II," Biometrika, 38 (1951).

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

[3] Geary, R. C., "The Contiguity Ratio and Statistical Mapping" (Appendix), The Incorporated Statistician, 5 : 3 (1954).

[4] Geary, R. C., "Some Remarks about Relations between Stochastic Variables", Review of the International Statistical Institute, 31 : 2 (1963).

[5] Theil, H. and Nagar, A. L., "Testing the Independence of Regression Disturbances", Journal of the American Statistical Association, 56 (1961).

****