Journal of Econometrics 2 (1974) 111–120.

SPURIOUS REGRESSIONS IN ECONOMETRICS

C.W.J. GRANGER and P. NEWBOLD

University of Nottingham, Nottingham NG7 2RD, England

Received May 1973, revised version received December 1973

1. Introduction

It is very common to see reported in applied econometric literature time series regression equations with an apparently high degree of fit, as measured by the coefficient of multiple correlation R^2 or the corrected coefficient $\bar{R}^2$, but with an extremely low value for the Durbin–Watson statistic. We find it very curious that whereas virtually every textbook on econometric methodology contains explicit warnings of the dangers of autocorrelated errors, this phenomenon crops up so frequently in well-respected applied work. Numerous examples could be cited, but doubtless the reader has met sufficient cases to accept our point. It would, for example, be easy to quote published equations for which $R^2 = 0.997$ and the Durbin–Watson statistic (d) is 0.53. The most extreme example we have met is an equation for which $R^2 = 0.99$ and $d = 0.093$. However, we shall suggest that cases with much less extreme values may well be entirely spurious. The recent experience of one of us [see Box and Newbold (1971)] has indicated just how easily one can be led to produce a spurious model if sufficient care is not taken over an appropriate formulation for the autocorrelation structure of the errors from the regression equation. We felt, then, that we should undertake a more detailed enquiry seeking to determine what, if anything, could be inferred from those regression equations having the properties just described.

There are, in fact, as is well-known, three major consequences of autocorrelated errors in regression analysis:

(i) Estimates of the regression coefficients are inefficient.
(ii) Forecasts based on the regression equations are sub-optimal.
(iii) The usual significance tests on the coefficients are invalid.

The first two points are well documented. For the remainder of this paper, we shall concentrate on the third point, and, in particular, examine the potentialities for 'discovering' spurious relationships which appear to us to be inherent in a good deal of current econometric methodology. The point of view we intend

to take is that of the statistical time series analyst, rather than the more classic econometric approach. In this way it is hoped that we might be able to illuminate the problem from a new angle, and hence perhaps present new insights. Accordingly, in the following section we summarize some relevant results in time series analysis. In sect. 3 we indicate how nonsense regressions relating economic time series can arise, and illustrate these points in sect. 4 with the results of a simulation study. Finally, in sect. 5, we re-emphasize the importance of error specification and draw a distinction between the philosophy of time series analysis and econometric methodology, which we feel to be of great importance to practitioners of the latter.

2. Some results in time series analysis

Let W_t denote a time series which is stationary (it could represent deviation from some deterministic trend). Then, the so-called mixed autoregressive moving average process,

$$W_t - \phi_1 W_{t-1} - \ldots - \phi_p W_{t-p} = a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}, \tag{1}$$

where a_t represents a sequence of uncorrelated deviates, each with the same variance, is commonly employed to model such series. The sequence a_t is referred to as 'white noise'. For brevity, eq. (1) can be written as

$$\phi(B) W_t = \theta(B) a_t, \tag{2}$$

where $\phi(B)$ and $\theta(B)$ are polynomial lag operators with appropriate roots to ensure stationarity of W_t and uniqueness of representation.

Suppose, now, that one has a given time series X_t. Box and Jenkins (1970) urge that, while this series itself may not be stationary, it can often be reduced to stationarity by differencing a sufficient number of times; that is, there exists an integer d such that

$$\nabla^d X_t = W_t \tag{3}$$

is a stationary time series. Combining eqs. (2) and (3), the series X_t can be represented by the model,

$$\phi(B) \nabla^d X_t = \theta(B) a_t. \tag{4}$$

Eq. (4) is said to represent an autoregressive integrated moving average process of order (p, d, q), denoted as A.R.I.M.A. (p, d, q).

As regards economic time series, one typically finds a very high serial correlation between adjacent values, particularly if the sampling interval is small, such as a week or a month. This is because many economic series are rather 'smooth', with changes being small in magnitude compared to the current level. There is thus a good deal of evidence to suggest that the appropriate value for d in

eq. (4) is very often one. [See, for example, Granger (1966), Reid (1969) and Newbold and Granger (1974).] Alternatively, if $d = 0$ in eq. (4) we would expect $\phi(B)$ to have a root $(1-\phi B)$ with ϕ very close to unity. The implications of this statement are extremely important, as will be seen in the following section.

The simplest example of the kind of series we have in mind is the random walk,

$$\nabla X_t = a_t.$$

This model has been found to represent well certain price series, particularly in speculative markets. For many other series, the integrated moving average process,

$$\nabla X_t = a_t - \theta a_{t-1},$$

has been found to provide good representation.

A consequence of this behaviour of economic time series is that a naive 'no change' model will often provide adequate, though by no means optimal, forecasts. Such models are often employed as bench-marks against which the forecast performance of econometric models can be judged. [For a criticism of this approach to evaluation, see Granger and Newbold (1973).]

3. How nonsense regressions can arise

Let us consider the usual linear regression model with stochastic regressors:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5}$$

where $\boldsymbol{Y}$ is a $T \times 1$ vector of observations on a 'dependent' variable, $\boldsymbol{\beta}$ is a $K \times 1$ vector of coefficients whose first member β_0 represents a constant term and $\boldsymbol{X}$ is a $T \times K$ matrix containing a column of ones and T observations on each of $(K-1)$ 'independent' variables which are stochastic, but distributed independently of the $T \times 1$ vector of errors $\boldsymbol{\varepsilon}$. It is generally assumed that

$$E(\boldsymbol{\varepsilon}) = \boldsymbol{o}, \tag{6}$$

and

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \boldsymbol{I}. \tag{7}$$

A test of the null hypothesis that the 'independent' variables contribute nothing towards explaining variation in the dependent variable can be framed in terms of the coefficient of multiple correlation R^2. The null hypothesis is

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{K-1} = 0, \tag{8}$$

and the test statistic

$$F = \frac{T-K}{K-1} \frac{R^2}{1-R^2} \tag{9}$$

is compared with tabulated values of Fisher's F distribution with $(K-1)$ and $(T-K)$ degrees of freedom, normality being assumed.

Of course, it is entirely possible that, whatever the properties of the individual time series, there does exist some $\boldsymbol{\beta}$ so that

$$\boldsymbol{\varepsilon} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$$

satisfies the conditions (6) and (7). However, to the extent that the Y_t's do not constitute a white noise process, the null hypothesis (8) *cannot* be true, and tests of it are inappropriate.

Next, let us suppose that the null hypothesis is correct and one attempts to fit a regression of the form (5) to the *levels* of economic time series. Suppose, further, that, as we have argued in the previous section is often the case, these series are non-stationary or, at best, highly autocorrelated. In such a situation the test procedure just described breaks down, since the quantity F in eq. (9) will *not* follow Fisher's F distribution under the null hypothesis (8). This follows since under that hypothesis the residuals from eq. (5),

$$\varepsilon_t = Y_t - \beta_0; \qquad t = 1, 2, \ldots, T,$$

will have the same autocorrelation properties as the Y_t series.

Some idea of the distributional problems involved can be obtained from consideration of the case:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t,$$

where it is assumed that Y_t and X_t follow the *independent* first order auto-regressive processes,

$$Y_t = \phi Y_{t-1} + a_t, \quad X_t = \phi^* X_{t-1} + \alpha_t. \tag{10}$$

In this case, R^2 is simply the square of the ordinary sample correlation between Y_t and X_t. Kendall (1954) gives:

$$\text{var}\,(R) = T^{-1}(1+\phi\phi^*)/(1-\phi\phi^*).$$

Since R is constrained to lie in the region $(-1, 1)$, if its variance is greater than $\frac{1}{3}$ then its distribution cannot have a single mode at zero. The necessary condition is $\phi\phi^* > (T-3)/(T+3)$. Thus, for example, if $T = 20$ and $\phi = \phi^*$, a distribution which is not unimodal at the origin will arise if $\phi > 0.86$, and if $\phi = 0.9$, $E(R^2) = 0.47$.

Thus a high value of R^2 should not, on the grounds of traditional tests, be regarded as evidence of a significant relationship between autocorrelated series. Also a low value of d strongly suggests that there does not exist a $\boldsymbol{\beta}$ such that $\boldsymbol{\varepsilon}$ in eq. (5) satisfies eq. (7). Thus, the phenomenon we have described might well arise from an attempt to fit regression equations relating the levels of independent time series. To examine this possibility, we conducted a number of simulation experiments which are reported in the following section.

4. Some simulation results

As a preliminary, we looked at the regression

$$Y_t = \beta_0 + \beta_1 X_t,$$

where Y_t and X_t were, in fact, generated as *independent* random walks each of length 50. Table 1 shows values of

$$S = \frac{|\hat{\beta}_1|}{\widehat{S.E.}(\hat{\beta}_1)},$$

the customary statistic for testing the significance of β_1, for 100 simulations.

Table 1
Regressing two independent random walks.

S:	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8
Frequency:	13	10	11	13	18	8	8	5
S:	8–9	9–10	10–11	11–12	12–13	13–14	14–15	15–16
Frequency:	3	3	1	5	0	1	0	1

Using the traditional t test at the 5% level, the null hypothesis of no relationship between the two series would be rejected (wrongly) on approximately three-quarters of all occasions. If $\hat{\beta}_1/\widehat{S.E.}(\hat{\beta}_1)$ were distributed as $N(0, 1)$, then the expected value of S would be $\sqrt{2/\pi} \simeq 0.8$. In fact, the observed average value of S was 4.5, suggesting that the standard deviation of $\hat{\beta}_1$ is being underestimated by the multiple factor 5.6. Thus, instead of using a t-value of approximately 2.0, one should use a value of 11.2, when attributing a coefficient value to be 'significant' at the 5% level.

To put these results in context, they may be compared with results reported by Malinvaud (1966). Suppose that X_t follows the process (10) and the error series obeys the model

$$\varepsilon_t = \phi \varepsilon_{t-1} + a_t,$$

so that, under the null hypothesis, Y_t will also follow this process, where a_t and α_t are independent white noise series. In the case $\phi = \phi^* = 0.8$, it is shown that the estimated variance of $\hat{\beta}_1$ should be multiplied by a factor 5.8, when the length of the series is $T = 50$. The approximations on which this result is based break down as both ϕ and ϕ^* tend to unity, but our simulation indicates that the estimated variance of $\hat{\beta}_1$ should be multiplied by $(5.6)^2 \simeq 31.4$ when $T = 50$ and random walks are involved.

Our second simulation was more comprehensive. A series Y_t was regressed on m independent series $X_{j,t}; j = 1, 2, \ldots, m$, with m taking values from one to

five. Each of the series involved obey the same model, the models being

(i) random walks,
(ii) white noises,
(iii) A.R.I.M.A. (0, 1, 1),
(iv) changes in A.R.I.M.A. (0, 1, 1), i.e., first order moving average.

Table 2

Regressions of a series on m independent 'explanatory' series.

Series either all random walks or all A.R.I.M.A. (0, 1, 1) series, or changes in these. $Y_0 = 100$, $Y_t = Y_{t-1} + a_t$, $Y_t' = Y_t + kb_t$; $X_{j,0} = 100$, $X_{j,t} = X_{j,t-1} + a_{j,t}$ $X_{j,t}' = X_{j,t} + kb_{j,t}$; $a_{j,t}, a_t, b_t, b_{j,t}$ sets of independent $N(0, 1)$ white noises. $k = 0$ gives random walks, $k = 1$ gives A.R.I.M.A. (0, 1, 1) series. H_0 = no relationship, is true. Series length = 50, number of simulations = 100, $\bar{R}^2$ = corrected R^2.

		Per cent times H_0 rejected[a]	Average Durbin–Watson d	Average $\bar{R}^2$	Per cent $\bar{R}^2 > 0.7$
			Random walks		
Levels	$m = 1$	76	0.32	0.26	5
	$m = 2$	78	0.46	0.34	8
	$m = 3$	93	0.55	0.46	25
	$m = 4$	95	0.74	0.55	34
	$m = 5$	96	0.88	0.59	37
Changes	$m = 1$	8	2.00	0.004	0
	$m = 2$	4	1.99	0.001	0
	$m = 3$	2	1.91	−0.007	0
	$m = 4$	10	2.01	0.006	0
	$m = 5$	6	1.99	0.012	0
			A.R.I.M.A. (0, 1, 1)		
Levels	$m = 1$	64	0.73	0.20	3
	$m = 2$	81	0.96	0.30	7
	$m = 3$	82	1.09	0.37	11
	$m = 4$	90	1.14	0.44	9
	$m = 5$	90	1.26	0.45	19
Changes	$m = 1$	8	2.58	0.003	0
	$m = 2$	12	2.57	0.01	0
	$m = 3$	7	2.53	0.005	0
	$m = 4$	9	2.53	0.025	0
	$m = 5$	13	2.54	0.027	0

[a]Test at 5% level, using an overall test on $\bar{R}^2$.

All error terms were distributed as $N(0, 1)$ and the A.R.I.M.A. (0, 1, 1) series was derived as the sum of a random walk and independent white noise. The results of the simulations, with 100 replications and series of length 50 are shown in table 2.

It is seen that the probability of accepting H_0, the hypothesis of no relationship, becomes very small indeed for $m \geqslant 3$ when regressions involve indepen-

dent random walks. The average $\bar{R}^2$ steadily rises with m, as does the average d, in this case. Similar conclusions hold for the A.R.I.M.A. (0, 1, 1) process. When white noise series, i.e., changes in random walks, are related, classical regression yields satisfactory results, since the error series will be white noise and least squares fully efficient. However, in the case where changes in the A.R.I.M.A. (0, 1, 1) series are considered – that is, first order moving average processes – the null hypothesis is rejected, on average twice as often as it should be.

It is quite clear from these simulations that if one's variables are random walks, or near random walks, and one includes in regression equations variables which should in fact not be included, then *it will be the rule* rather than the exception to find spurious relationships. It is also clear that a high value for R^2 or $\bar{R}^2$, combined with a low value of d, is *no indication of a true relationship.*

5. Discussion and conclusion

It has been well known for some time now that if one performs a regression and finds the residual series is strongly autocorrelated, then there are serious problems in interpreting the coefficients of the equation. Despite this, many papers still appear with equations having such symptoms and these equations are presented as though they have some worth. It is possible that earlier warnings have been stated insufficiently strongly. From our own studies we would conclude that if a regression equation relating economic variables is found to have strongly autocorrelated residuals, equivalent to a low Durbin–Watson value, the *only conclusion that can be reached is that the equation is mis-specified,* whatever the value of R^2 observed.

If such a conclusion is accepted, the question then arises of what to do about the mis-specification. The form of the mis-specification can be considered to be either (i) the omission of relevant variables or (ii) the inclusion of irrelevant variables or (iii) autocorrelated residuals. In general, the mis-specification is best considered to be a combination of these possibilities. The usual recommendations are to either include a lagged dependent variable or take first differences of the variables involved in the equation or to assume a simple first-order autoregressive form for the residual of the equation. Although any of these methods will undoubtedly alleviate the problem in general, it is doubtful if they will completely remove it.

It is not our intention in this paper to go deeply into the problem of how one should estimate equations in econometrics, but rather to point out the difficulties involved. In our opinion the econometrician can no longer ignore the time series properties of the variables with which he is concerned – except at his peril. The fact that many economic 'levels' are near random walks or integrated processes means that considerable care has to be taken in specifying one's equations. One

method we are currently considering is to build single series models for each variable, using the methods of Box and Jenkins (1970) for example, and then searching for relationships between series by relating the residuals from these single models. The rationale for such an approach is as follows. In building a forecasting model, the time series analyst regards the series to be forecast as containing two components. The first is that part of the series which can be explained in terms of its own past behaviour and the second is the residual part [a_t in eq. (4)] which cannot. Thus, in order to explain this residual element one must look for other sources of information–related time series, or perhaps considerations of a non-quantitative nature. Hence, in building regression equations, the quantity to be explained is variation in a_t – not variation in the original series. This study is, however, still in its formative stages. Until a really satisfactory procedure is available, we recommend taking first differences of all variables that appear to be highly autocorrelated. Once more, this may not completely remove the problem but should considerably improve the interpretability of the coefficients.

Perhaps at this point we should make it clear that we are not advocating first differencing as a universal sure-fire solution to any problem encountered in applied econometric work. One cannot propose universal rules about how to analyse a group of time series as it is virtually always possible to find examples that could occur for which the rule would not apply. However, one can suggest a rule that is useful for a class of series that very frequently occur in practice. As has been noted, very many economic series are rather smooth, in that the first serial correlation coefficient is very near unity and the other low-order serial correlations are also positive and large. Thus, if one has a small sample, of say twenty terms, the addition of a further term adds very little to the information available, as this term is so highly correlated with its predecessor. It follows that the total information available is very limited and the estimates of parameters associated with this data will have high variance values. However, a simple calculation shows that the first differences of such a series will necessarily have serial correlations that are small in magnitude, so that a new term of the differenced series adds information that is almost uncorrelated to that already available and this means that estimates are more efficient. One is much less likely to be misled by efficient estimates.

The suggested rule perhaps should be to build one's models both with levels and also with changes, and then interpret the combined results so obtained. As an example (admittedly extreme) of the changes that can occur in one's results from differencing, Sheppard (1971) regressed U.K. consumption on autonomous expenditure and mid-year money stock, both for levels and changes for the time period 1947–1962. The regression on levels yielded a corrected R^2 of 0.99 and a d of 0.59, whilst for changes these quantities were -0.03 and 2.21 respectively. This provides an indication of just how one can be misled by regressions involving levels if the message of the d statistic is unheeded.

It has been suggested by a referee that our results have relevance to the structural model – unrestricted reduced form controversy, the feeling being that the structural model is less vulnerable to the problems we have described since its equations are in the main based on well developed economic theory and contain relatively few variables on the right-hand side. There is some force to this argument, in theory at least, although we believe that in practice things are much less clear cut.

When considering this problem the question immediately arises of what is meant by a good theory. To the time series analyst a good theory is one that provides a structure to a model such that the errors or residuals of the fitted equations are white noises that cannot be explained or forecast from other economic variables. On the other hand, some econometricians seem to view a good theory as one that appears inherently correct and thus does not need testing. We would suggest that in fact most economic theories are insufficient in these respects as even if the variables to be included in a model are well specified, the theory generally is imprecise about the lag structure to be used and typically says nothing about the time-series properties of the residuals. There are also data problems in that the true lags need not necessarily be integer multiples of the sampling interval of the available data and there will almost certainly be added measurement errors to the true values of the variables being considered. All of these considerations suggest that a simple-minded application of regression techniques to levels could produce unacceptable results.

If one does obtain a very high R^2 value from a fitted equation, one is forced to rely on the correctness of the underlying theory, as testing the significance of adding further variables becomes impossible. It is one of the strengths of using changes, or some similar transformations, that typically lower R^2 values result and so more experimentation and testing can be contemplated. In any case, if a 'good' theory holds for levels, but is unspecific about the time-series properties of the residuals, then an equivalent theory holds for changes so that nothing is lost by model building with both levels and changes. However, much could be gained from this strategy as it may prevent the presentation in econometric literature of possible spurious regressions, which we feel is still prevalent despite the warnings given in the text books about this possibility.

References

Box, G.E.P. and G.M. Jenkins, 1970, Time series analysis, forecasting and control (Holden-Day, San Francisco).

Box, G.E.P. and P. Newbold, 1971, Some comments on a paper of Coen, Gomme and Kendall, J. R. Statist. Soc. A 134, 229–240.

Granger, C.W.J., 1966, The typical spectral shape of an economic variable, Econometrica 34, 150–161.

Granger, C.W.J. and P. Newbold, 1973, Some comments on the evaluation of economic forecasts, Applied Economics 5, 35–47.

Kendall, M.G., 1954, Exercises in theoretical statistics (Griffin, London).
Malinvaud, E., 1966, Statistical methods of econometrics (North Holland, Amsterdam).
Newbold, P. and C.W.J. Granger, 1974, Experience with forecasting univariate time series and the combination of forecasts, J. R. Statist. Soc. A 137, forthcoming.
Reid, D.J., 1969, A comparative study of time series prediction techniques on economic data, Ph.D. Thesis (University of Nottingham, U.K.).
Sheppard, D.K., 1971, The growth and role of U.K. financial institutions 1880–1962 (Methuen, London).