Structural parameter estimation with omitted variables

by C. E. V. Leser

1. Theoretical considerations.

The problem of specification bias arising out of the omission of relevant variables in econometric relationships has been considered by various writers, notably by Griliches (1957, 1961), by Theil (1957, 1958) and by Wold and Faxer (1957). Some of the discussion is in terms of autocorrelation of residuals, whilst alternatively the effect of correlation between regressors and omitted variables forming part of the residual has been examined.

Geary (1963) points out that the classical least squares assumption of non-correlation between relevant variables included in and excluded from an actual regression is unlikely to be satisfied. Specification bias in the regression coefficients is thus likely, and in order to eliminate or at least reduce the bias, a systematic search for omitted variables appears to be called for.

Alternatively, a transformation of variables by trend elimination may largely remove the bias in the regression coefficients under certain assumptions. It will be suggested here that these assumptions may be taken as realistic in many cases encountered in practice.

Take the simple regression model

$$y = \beta x + z \qquad (1)$$

where y, x and z are variables measured as deviations from their means. A vector of observations for x and y is available from time series; there would be no difficulty

in extending the model to multiple regression in which x is represented by a matrix of observations and β by a vector of coefficients. z is the unknown omitted variable which may be a function of numerous economic variables. The full relationship is taken as exact, and true linearity in the relationship between x and y is implied.

Now assume for the moment that both x and z followed a linear trend and that in both cases the deviations from the trend were independently distributed random variables and thus not autocorrelated. Then orthogonality between the two sets of trend deviations could also be assumed. It follows immediately that if the trend was removed with the help of regression on time or by taking first differences, regression of the y residual on the x residual would give an unbiased estimate of β.

In practice, a linear term is hardly ever appropriate to describe long-term movements in economic time series, as evidenced from the fact that first differences are generally autocorrelated and may not legitimately be treated as random. As a result, regression coefficients derived from first differences will normally contain an element which reflects the changes in trend direction which are common to both the regressor and the omitted variable, since many economic variables will undergo rapid changes in some periods, whilst other periods will see a general slowing down of changes. This festure is a help for prediction but a hindrance for structural parameter estimation.

We may, however, assume that with a correctly specified trend, randomness of the trend deviations

will hold. Thus model (1) may be supplemented by the equations

$$\begin{aligned} x &= t_1 + u_1 \\ z &= t_2 + u_2 \end{aligned} \qquad (2)$$

where t_1, t_2 are the trends and u_1, u_2 the random components of the variables. Then y will also have a trend and a random component. Writing v for the latter we have

$$v = \beta u_1 + u_2 \qquad (3)$$

Assuming that we can correctly specify and eliminate the trend from x and y and that u_1 and u_2 are orthogonal, regression of v on u_1 will yield an unbiased estimate of β.

It should be noted that since u_2 cannot at the same time be orthogonal to v, some asymmetry in the relationship between x and y is implied. There must be a clear causal direction or some other reason why the regression of y on x is chosen rather than the regression of x on y.

The problem is, of course, to find the trends t_1 of x and $\beta t_1 + t_2$ of y. There does not seem to be an operational trend construction method which can be relied upon to produce non-autocorrelated residuals but a moving-average type of trend would generally seem more promising than a low-degree polynomial. The quasi-linear trend method developed by Leser (1961) is theoretically founded and may, with some qualifications, be considered as suitable.

2. A practical experiment.

To illustrate the foregoing consideration national accounts data at current prices for Ireland from 1947 to 1964 published by the Central Statistics Office (1966) have been analysed. The variables are as follows

C personal expenditure
G government current expenditure
I gross fixed capital formation
B net stockbuilding
X exports of goods and services
M imports of goods and services
Y gross national product

so that

$$C + G + I + B + X = M + Y \qquad (4)$$

For each variable, the 17 first differences for the observation period have been calculated. Furthermore, the quasi-linear trend has been constructed and eliminated, thus yielding for each variable 18 errors or temporary disturbances, which have zero-sum and zero-correlation with time. The Durbin-Watson d-statistic has been evaluated for each series and the results are given in Table 1.

Table 1. Value of d-statistic for national accounts data at current prices, Ireland 1947-64.

Series	First differences	Quasi-linear trend errors
C	1.24	3.11
G	0.94	2.59
I	0.45	2.00
B	3.07	3.05
X	1.44	2.38
M	1.88	2.61
Y	0.82	2.23

Positive first-order serial correlation thus seems to be a feature of all first differences except for stockbuilding and perhaps imports. In contrast, the quasi-linear trend errors have a tendency towards negative serial correlation; this tendency is inherent in the trend construction method though it disappears asymptotically. In most cases the d-values for the trend errors differ less from 2 than those for the first differences.

Next, the variables have been correlated in pairs and Table 2 shows the result. It may be noted that on account of (4) we cannot reasonably expect a priori all pairs of variables to be uncorrelated. If, for example, all variables on the left-hand side were orthogonal to each other, the remaining pairs would have a positive expectation for their correlation.

Table 2. Correlation between national accounts data, Ireland 1947-64

Variables	First differences		Trend errors	
	r^2	Sign of r	r^2	Sign of r
C G	.7307	+	.4206	+
C I	.4129	+	.0541	+
G I	.4885	+	.0962	+
C B	.0622	+	.0664	+
G B	.0298	+	.0374	+
I B	.0034	+	.0132	+
C X	.2680	+	.0400	+
G X	.2892	+	.0544	+
I X	.1789	+	.0675	-
B X	.0001	-	.0544	-
C M	.4510	+	.2696	+
G M	.3660	+	.1520	+
I M	.3604	+	.2055	+
B M	.2737	+	.2968	+
X M	.1838	+	.0019	+
C Y	.6416	+	.2390	+
G Y	.6353	+	.2301	+
I Y	.3700	+	.0235	-
B Y	.0398	+	.0145	+
X Y	.4768	+	.2589	+
M Y	.1120	+	.0576	-

For the first differences, all correlations with one exception are positive; and furthermore, the correlation is substantial not only between most left-hand and right-hand side variables of (4) but also between all pairs of left-hand side variables except those including B. For the trend errors, positive correlation coefficients still predominate over negative ones but the correlation is low between pairs of left-hand variables except C and G. This association between C and G is probably meaningful but not of outstanding theoretical interest, indicating merely the effect of wage and salary rises.

On both counts of the values obtained for d and r^2, the quasi-linear trend errors are better qualified to be treated as random variables than the first differences, and will be so treated notwithstanding their imperfections in this context.

To apply these considerations, take it that we are interested in the effect of exports on gross national product. Conditions appear favourable for a regression of Y on X, on account of the relatively low correlations between the trend errors of X and the other variables. Implied assumptions are: a) causal direction is from X to Y, i.e. export-led growth rather than growth-led exports; b) linearity of both variables in the relationship; c) the current term for X to be relevant, though a lagged term could also appear, as part of the omitted variable.

Our being ignorant of the precise nature of the omitted variable, various hypotheses may be investigated. In their formulation, an error term denotes a term which

is uncorrelated with exports and any specified component of the omitted variable. Alternative hypotheses for the omitted variables are then: an error term; fixed capital formation and an error term; personal expenditure and an error term; fixed capital formation, personal expenditure and an error term. According to the hypothesis the regression coefficient β of Y on X is estimated by regressing Y on X alone, on X and I, on X and C, or on X, I and C. The partial regressions will yield additional regression coefficient estimates which, however, are of no direct interest here. Table 3 shows the results of this exercise.

Table 3. Results of regressing Y on X.

Additional regressors	b	s_b	R^2
First differences:			
None	1.428	0.386	.477
I	1.091	0.386	.599
C	0.779	0.326	.746
I and C	0.753	0.333	.752
Trend errors:			
None	0.839	0.366	.259
I	0.829	0.392	.259
C	0.706	0.344	.415
I and C	0.621	0.369	.437

The relatively low values of R^2 obtained by analysing the trend errors may be noted but they do not furnish an argument against using the method, any more than the lower value of R^2 obtained with first differences than with original data argues against first differences. The real point of Table 3 is the relative insensitivity of b to the specification when derived from the quasi-linear trend errors. The result suggests that, say,

a £1 mill. increase in exports brings about a rise in gross national product by somewhat less, and not more, than £1 mill. The trend removal carried out here appears, if not to eliminate, at least to reduce the risk of specification bias.

References

Central Statistics Office (1966), National income and expenditure 1964, Dublin.

Geary, R.C. (1963), "Some remarks about relations between stochastic variables: a discussion document", Review of the International Statistical Institute, Vol. 31, pp.163-181.

Griliches, Z. (1957), "Specification bias in estimates of production functions", Journal of Farm Economics, Vol. 39, pp.8-20.

do. (1961), "A note on serial correlation bias in estimates of distributed lags", Econometrica, Vol. 29, pp.65-73.

Leser, C.E.V. (1961), "A simple method of trend construction", Journal of the Royal Statistical Society, Ser. B, Vol. 23, pp.91-107.

Theil, H. (1957), "Specification errors and the estimation of economic relationships", Review of the International Statistical Institute, Vol. 25, pp.41-51.

do. (1958), Economic forecasts and policy. Amsterdam: North-Holland Publishing Co.

Wold, H. and Faxer , P. (1957), "On the specification error in regression analysis", Annals of Mathematical Statistics, Vol. 28, pp.265-267.