Comparative Efficiency of Maximum Likelihood and *Ex Ante* Reduced Form Methods of Solution of Behaviouristic Equations for Forecasting and Policy-making

by

R. C. Geary

The famous issue of maximum likelihood (ML) versus least squares (LS) in the solution of a behaviouristic equation system flares up from time to time but seems as yet unresolved. Accordingly, as the writer is about to embark on a possibly large model for his own country based on time series, he himself has to face the issue now. The present investigation leaves him convinced that *ex ante* reduced form (RF) with individual equation LS is the better way.

Of course, "it all depends on what one wants the model for", to quote the too familiar cliché. One objective the writer has *not* in mind is individual coefficient estimation. He was vehement some years ago in the assertion that in multivariate regression (and, *a fortiori*, in equation systems) individual coefficients are meaningless: the only coefficients possibly economically significant are those of simple regression [1]. The writer is not aware of any serious attempt to rebut his views; nonetheless, economic interpretations of individual coefficients (usually interpreted as "elasticities" or such like), with their implicit untenable *ceteris paribus* assumption, are still rife.

The only use the writer can find in solving large or small equation systems is forecasting (of the endogenous variables) and policy-making; for what follows, however, it will suffice to assume that forecasting is *an* objective. This objective requires the calculation of y_c, the vector for some specified time t of endogenous values, given the values of the predetermined variables. Of course the estimation of the coefficients is involved, but to be used only as a set and not individually.

Original and Reduced Form

Let to original form (OF) of the model (in matrix notation) be

(1) $y\beta = x\alpha + u$

There are T sets of observations, p endogenous variables y, q exogenous variables x and an error matrix u about which the usual assumptions are made, including non-autoregression and a population var-covar matrix, the same for all times t, β and α are the population coefficient population matrices. The dimensions of the five matrices involved are accordingly as follows:- y: T x p, β: p x p, x: T x q, α : q x p, u: T x p. β is a square matrix, usually with principal diagonal unities. We assume for simplicity that x is pure exogenous, i.e. it contains no lagged endogenous variables, not an issue here. Of course, x need not be linear, though y must. In accordance with the usual convention, the stochastic properties of the model enter solely through u, x being the same for each realisation, of which we have, in practice, only one. The expected value E is the mean of a hypothetically indefinitely large number of realisations. For the comparative efficiency purpose of the present paper, the population values, x β, α and the var-covar matrix are supposed known.

We shall concern ourselves with *ex ante* RF. We assume that (1) has been set up on theoretical considerations: usually one equation is designed to explain each endogenous variable, the explanatory, or causal, variables in each equation being other endogenous and exogenous variables. These explanatory variables are customarily few in number, at most four or five. The coefficients β and α are still in the form of symbols, unestimated. From this form we may derive *ex ante* RF as follows

(2) $y = x\alpha\beta^{-1} + u\beta^{-1}$,

or

(3) $y = x\Upsilon + v$

where $\Upsilon = \alpha\beta^{-1}$ and $v = u\beta^{-1}$. The object of this transformation is to pick out, on the right side, the exogenous variables with non-zero coefficients, which theory, enshrined in (1), ordains. One hopes that, as in the case of OF, the right side exogenous variables will be few in each equation. As the var-covar matrix of u is $Eu'u/T$, the corresponding matrix for v is $E(\beta')^{-1} u'u\beta^{-1}/T$, a fact of considerable importance for what follows.

We do not consider LS applied to the individual equations in OF in non-recursive models as, following the well-known work of Haavelmo [2], we regard this method as invalid. In fact, asymptotically it yields inconsistent estimates of y.

Suppose, now, that it is possible to estimate, by ML or otherwise, the coefficients β and α by b and a respectively and residual u by $\hat{u}$ in a consistent way, i.e. so that each element tends in probability (as T increases) towards its population value, a property which may be written

(4) $b \sim \beta$; $a \sim \alpha$; $\hat{} \sim u$.

It is necessary to have recourse to *ex post* RF to estimate y by y_c;

(5) $y_c = xab^{-1} \sim x\alpha\beta^{-1} = \eta$; $y = y_c + \hat{v}$; $\hat{v} \sim u\beta^{-1}$.

It is quite clear that the var-covar matrix of

$$(y - y_c) \sim E(\beta')^{-1} u'u\beta^{-1}/T.$$

We can also estimate a calculated value of y, call it y_{1c}, from the *ex ante* RF version of the model (3):-

(6) $y_{1c} = xc \sim xy = \eta$; $y = y_{1c} + \hat{w}$, $\hat{w} \sim u\beta^{-1}$

Obviously the var-covar matrix of $(y - y_{1c}) \sim E(\beta')^{-1} u'u\beta^{-1}/T$, as in the case of OF. It is to be noted for comparative purposes, that, at (5) (OF) and (6) (RF), y(involved in the estimation of y_c and y_{1c}), x and π are identical, y and x because they are data and π is the value of y, given x wh u is zero, so that $\pi\beta = x\alpha$ implies $\pi = x\alpha\beta^{-1}$.

Given our criterion based on the difference between the actual and calculated value of the endogenous variables $(y - y_c)$ or $(y - y_{1c})$ and our objective (forecasting and policy-making), the identity of the population var-covar matrices means that there is no asymptotic ($T \to \infty$) difference in efficiency between OF (with ML) and RF (with LS).

Desiring to examine the issue to a closer approximation, we decided to compare the values (from now on using non-matrix notation) of $E\,(y_t - y_{tc})^2$ and $E\,(y_t - y_{t1c})^2$, the mean square of an indefinitely large number of replications of the deviations for given values of x_t, for a particular simple model. We prefer the criterion we have adopted to, say, $E\,(y_{tc} - \pi_t)^2$, mainly because, in any relization, the latter is not estimable, whereas the former is. Our method is to expand the criteria to terms in T^{-1}, the terms in T^o being the same in both cases, as we have seen in the general case.

The Simple Recursive Model

As our object is measurement, we have recourse to the only case in which the OF (ML) solution is algebraically manageable, which is the recursive system of equations. In this case, as is well-known, the ML solution is found by individual equation LS in OF, when u in (1) is normally distributed, now assumed. We select the simplest possible recursive model, as follows

$$(7)\quad \left.\begin{array}{ll} \text{(i)} & y_{t1} = \alpha_1 x_{t1} + u_{t1} \\ \text{(ii)} & y_{t2} = \beta_1 y_{t1} + \alpha_2 x_{t2} + u_{t2} \end{array}\right\} t = 1, 2, \ldots, T$$

The estimates of α_1, α_2, and β_1 are a_1, a_2 and b_1 respectively. There is no issue with regard to 7 (i): the OF and RF estimates of α_1 (by a_1) and of y_{t1} (by y_{t1c}) are identical. Investigation will therefore be confined to 7 (ii) of which the ML solution is found by ordinary LS, yielding the equations:-

(8) $$(b_1 - \beta_1)\Sigma(\alpha_1 x_{t1} + u_{t1})^2 + (a_2 - \alpha_2)\Sigma(\alpha x_{t1} + u_{t1})\, x_{t2} = \Sigma\, u_{t2}\, (\alpha x_{t1} + u_{t1})$$

$$(b_1 - \beta_1)\Sigma(\alpha_1 x_{t1} + u_{t1})\, x_{t2} + (a_2 - \alpha_2)\,\Sigma\, x^2_{t2} = u_{t2}\, x_{t2}$$

The Σ indicates summation with regard to t. It will now be convenient to deem (without loss of generality) x_1 and x_2 as standardized, i.e.,

(9) $$\Sigma x_{t1} = 0;\ \Sigma x_{t2} = 0;\ \Sigma x^2_{t1} = T;\ \Sigma x^2_{t2} = T;$$
$$\Sigma x_{t1} x_{t2} = T\rho$$

Then (3) becomes

(10) $$(b_1 - \beta_1)(\alpha_1^2 + 2\alpha_1 e_1 + e_5) + (a_2 - \alpha_2)(\alpha_1\rho + e_2) = \alpha_1 e_3 + e_6$$

$$(b_1 - \beta_1)(\alpha_1\rho + e_2) \qquad + (a_2 - \alpha_2) = e_4,$$

with

$$Te_1 = \Sigma\, u_{t1} x_{t1};\ Te_2 = \Sigma\, u_{t1} x_{t2};\ Te_3 = \Sigma\, u_{t2} x_{t1};$$
$$Te_4 = \Sigma\, u_{t2} x_{t2};$$

(11) $$Te_5 = \Sigma\, u^2_{t1}; \qquad Te_6 = \Sigma\, u_{t1} u_{t2}.$$

As, for the purpose of comparison of efficiency, we are entitled to assume knowledge of α_1, α_2, β_1 and the variances of the error terms u_{t1} and u_{t2}, namely σ_1^2 and σ_2^2, we can go further and assume, without loss of generality, that σ_1 and σ_2 are both* unity and u_{t1} and u_{t2} independent. Then the variances of all the e's (except e_5) at (11) are 1/T, while it will suffice to note that $Ee_5 = 1$.

*If the original values in (7) be indicated by primed symbols, (except x_{t1} and x_{t2} unchanged) transition to the form with residual variances unity is effected by:-

$$u_{t1} = u'_{t1}/\sigma_1;\ u_{t2} = u'_{t2}/\sigma_2;\ \alpha_1 = \alpha'_1/\sigma_1;\ y_{t1} = y'_{t1}/\sigma_1;$$
$$y_{t2} = y'_{t2}/\sigma_2;\ \beta_1 = \beta'_1\,\sigma_1/\sigma_2.$$

The OF forecasting formula from (7) is,

(12) $y_{t2c} = a_1 b_1 x_{t1} + a_2 x_{t2}$,

where a_1 is found by LS from (7) (i) as

(13) $a_1 = \alpha_1 + \Sigma u_{t1}/T = \alpha_1 + e_1$.

The decision function is

(14) $X_t = E(y_{t2} - y_{t2c})^2$.

It is to be noted that in the hypothetically indefinitely large number of replications implicit in (14), t, x_{t1} and x_{t2} are the same in all. From (7) and (12),

(15) $y_{t2} - y_{t2c} = (u_{t2} + \beta_1 u_{t1}) - (a_1 b_1 - \alpha_1 \beta_1) x_{t1} - (a_2 - \alpha_2) x_{t2}$,

From which it is evident that the leading term in (14) is

(16) $E(u_{t2} + \beta_1 u_{t1})^2 = 1 + \beta_1^2$.

Two Cases

Two views may be taken about the explicit error term, $(u_{t2} + \beta_1 u_{t1})$, in (15), according to whether one is concerned with (i) measuring goodness-of-fit of estimate y_{t2c} to observation y_{t2} or (ii) using the formula for forecasting and policy-making. In case (i) only the T sets of observations are involved: u_{t1} and u_{t2} in (15) are the error terms involved in the estimation of α_1, α_2 and β_1 so that the error term is not statistically independent of the estimates of the coefficients of x_{t1} and x_{t2}. In case (ii) we are concerned with future time, actual as regards forecasting and hypothetical as regards policy-making; the u_{t1} and u_{t2}, being errors pertaining to future time, are independent of the errors in the estimated coefficients of x_{t1} and x_{t2} which are functions of the errors in past time. The result is different values of X_t, given by (14), in cases (i) and (ii). We consider approximations to both.

Case (i): Goodness-of-fit

For the present purpose expansion of (14) to the term in 1/T only is required: the right side of (15) is squared after substitution of a_1 from (13) and a_2 and b_1 from (10):- $b_1 - \beta_1 = [(\alpha_1 e_3 + e_6) - (\alpha_1^2 \rho + e_2)] e_4 / d$

(17) $a_2 - \alpha_2 = [(\alpha_1^2 + 2\alpha_1 e_1 + e_5) e_4 - (\alpha_1^2 \rho e_3 + \alpha_1 \rho e_6 + \alpha_1 e_2 e_3 + e_2 e_6)] / d$,

where d is given by

(18) $d = (\alpha_1^2 + 2\alpha_1 e_1 + e_5) - (\alpha_1^2 \rho^2 + 2\alpha_1 e_2 + e^2_2$;

so that (using (11)),

(19) $\dot{=} \Delta = \alpha^2_1 (1 - \rho^2) + 1$

and

(20) $Ed^2 \doteq \Delta^2$.

The symbol "$\doteq$" means "equals, to the approximation required". The actual or approximate values of the six terms (T_i)* in the expansion of the right side of X (14) using (15) are, after much algebra, given by:-

$$T_1 = 1 + \beta_1^2$$
$$-\Delta T.\ T_2 \doteq 2(\alpha_1^2 x^2_{t2} + x^2_{t2} - \alpha^2_1 \rho x_{t1} x_{t2})$$
$$-\Delta T.\ T_3 \doteq 2(\alpha^2_1 x^2_{t1} - \alpha^2_1 \rho x_{t1} x_{t2}) + 2\Delta \beta^2_1 x^2_{t1}$$
(21) $$\Delta T.\ T_4 \doteq x^2_{t2} (\alpha^2_1 + 1)$$
$$-\Delta T.\ T_5 \doteq 2\alpha^2_1 \rho x_{t1} x_{t2}$$
$$\Delta T.\ T_6 \doteq (\alpha^2_1 + \Delta \beta^2_1) x^2_{t1}$$

whence

(22) $$X_t \doteq (1 + \beta^2_1) - \frac{1}{T\Delta}\{(\alpha^2_1 + \beta^2_1 \Delta) x^2_{t1} - 2\rho\alpha^2_1 x_{t1} x_{t2} + (1 + \alpha^2_1) x^2_{t2}\}$$

(23)

The *ex ante* RF model of the system is

(i) $y_{t1} = \alpha_1 x_{t1} + u_1$

(ii) $y_{t2} = \Upsilon_1 x_{t1} + \alpha_2 x_{t2} + v_{t2}$,

where

(24) $$\Upsilon_1 = \alpha_1 \beta_1;\ v_{t2} = u_{t2} + \beta_1 u_{t1}.$$

The RF (LS) expression corresponding to (22) is found to be

(25) $$X'_t = E(y_{t2} - y'_{t2c})^2 = (1 + \beta^2_1) - \frac{(1 + \beta^2_1)}{T(1 - \rho^2)} (x^2_{t1} - 2\rho x_{t1} x_{t2} + x^2_{t2}),$$

---

*If the coefficients on the right of (15) are A_1, A_2 and A_3, then $T_1 = EA^2_1$, $T_2 = 2EA_1 A_2$, $T_3 = 2EA_1 A_3$, $T_4 = EA^2_2$, $T_5 = 2EA_2 A_3$, $T_6 = EA_3^2$.

From (22) and (25),

(26) $DT\,(X_t - X'_t) \doteq (1 + \rho^2\phi)x^2_{t1} - 2\rho(1 + \phi)x_{t1}x_{t2} + (\rho^2 + \phi)x^2_{t2}$

where

(27) $\Delta = \alpha^2_1(1 - \rho^2) + 1;\ D = \Delta(1 - \rho^2);\ \phi = \beta^2\Delta$.

As the discriminant of the right side of (26),

$= -4\beta^2_1(1 - \rho^2)^2$, is negative, this right side is always positive. Hence $X_t \geq X'_t$. Hence, as regards goodness of fit, RF with LS is at least as efficient as OF with ML. It is surely remarkable that this property holds for each set of the exogenous variables and not merely for the sum squares differences $E\Sigma(X_t - X'_t)$.

Case (ii) - Forecasting and Policy-making

As already remarked, the error term in (15), $(u_{t2} + \beta_1 u_{t1})$ is now independent of the coefficients of x_{t1} and x_{t2}, which means that $E(y_{t2} - y_{t2c})^2$ is the sum $T_1 + T_4 + T_5 + T_6$ (see (21))in the OF (ML) situation:-

(28) $$Y_t = E(y_{t2} - y_{t2c})^2 \doteq (1 + \beta^2_1) + \frac{1}{\Delta T}\{(\alpha^2_1 + \beta^2_1\Delta)x^2_{t1} - 2\rho\alpha^2_1 x_{t1}x_{t2} + (1 + \alpha^2_1)\}x^2_{t2}$$

The corresponding RF (LS) impression is

(29) $$Y'_t = E(y_{t2} - y_{t2c})^2 \doteq (1 + \beta^2_1) + \frac{1 + \beta^2_1}{T(1 - \rho^2)}(x^2_{t1} - 2\rho x_{t1}x_{t2} + x^2_{t2})$$

We now see that the expressions for Y and Y' at (28) and (29) differ respectively from X and X' given by (22) and (25) only in the sign following $(1 + \beta^2_1)$. Hence

(30) $$X_t - X'_t \doteq Y'_t - Y_t.$$

The situation is therefore now reversed: OF with ML is now more efficient than RF with LS. In both cases the relative superiority arises only in the term in T^{-1}. The T^0 term is identical throughout, namely $(1 + \beta^2_1)$, so that asymptotically the two approaches are equally efficient.

The Value of $E(y_{t2c} - \pi_t)^2$

Here the population value π_t of y_{t2c} (or y'_{t2c}) is given by

(31) $\eta_t = \alpha_1 \; \beta_1 \; x_{t1} + \alpha_2 \; x_{t2} = \Upsilon_1 \; x_{t1} + \alpha_2 \; \alpha_{t2}$.

We have rejected $E(y_{t2c} - \eta_t)^2$ as a valid criterion as assessing the relative merits of ML and LS. Nevertheless it may be interesting to observe that if Z_t and Z'_t be the respective values of this expression under ML and LS conditions $(Z_t - Z'_t)$ is found to be approximately $-(X_t - X'_t)$ given by (26). Hence $(Z_t - Z'_t)$ is a non-positive quantity for all values of the exogenous set (x_{t1}, x_{t2}). This result is a consequence of ML being asymptotically more efficient for estimating the coefficients which alone enter the calculation: the residual errors u_{t1} and u_{t2} as explicit terms are eliminated. Therefore, to complete (30),

(32) $X_t - X'_t \doteq Y'_t - Y_t \doteq Z'_t - Z_t$.

<u>A Constructed Example</u>

Unsure, at the start, that we would be able to cope with the algebra of even the simple recursive model, we set up a constructed illustration using the following population values (see (7)):- $\alpha_1 = 2$, $\beta_1 = 5$, $\alpha_2 = 3$, $T = 30$. x_{t1} and x_{t2} were found from fairly highly correlated (ρ = .83) annual time series; u_{t1}, and u_{t2} were independent random samples from N(0,1). So y_{t1} and y_{t2} were built up, constituting, with x_{t1} and x_{t2}, the "data".

We need not give the details. Following are the estimated values of the coefficients using the two systems:-

Coefficient	Estimation	Population
	Original form (OF)	
α_1	2.20	2
β_1	4.92	5
α_2	3.17	3
	Reduced form (RF)	
α_1	2.20	2
$\alpha_1\beta_1 = \Upsilon_1$	12.18	10
β_1	5.54	5
α_2	1.57	3

Hence, on the showing of these figures, there can be no question about the superiority of OF (ML) as regards individual coefficient estimation, in which, however, we are not interested. The *ex ante* RF (with LS) yields bizarre values. Yet all the errors of estimate of the coefficients lie within the .95 probability limits. The main reason for the greater accuracy of the OF (ML) estimates is that the residual (population) variance is 1, whereas it is $1 + \beta^{1}{}_{2} = 26$ in the RF (LS) case. Yet the latter affords the better goodness-of-fit to the data for we find:-

$$\text{OF}: \Sigma (y_{t2} - y_{t2c})^2/T = 21.8$$

$$\text{RF}: \Sigma (y_{t2} - y^{1}{}_{t2c})^2/T = 20.9$$

As we have but one realization there is no possibility of calculating the E values of the text. Comparison of the deviations in each of the T = 30 sets of data shows that in 17 cases $(y_{t2} - y^{1}{}_{t2c})^2$ (i.e. RF) is the smaller and in the remaining 13 cases $(y_{t2} - y_{t2c})^2$ (i.e. OF) is the smaller. If we had the E values the RF (LS) value would be smaller in every case. In truth, as far as results go in any single realisation, there seems little to choose between OF (ML) and RF (LS). As stated in the text all the advantage comes from computational simplicity in the ~~general case~~.

Conclusion

From the strictly statistical point of view there is but little difference in efficiency between the OF (with ML) and *ex ante* RF (with individual equation LS) approaches. For forecasting and policy-making, OF (ML) is the more efficient by our criterion; on the goodness-of-fit test, RF (LS) is the more efficient. It is true that these comparisons are based on an examination of the simplest possible recursive system: the writer would be greatly surprised, however, if investigation by algebra or Monte Carlo on a general system yielded a different assessment, for then the problem would remain of explaining away the recursive case.

Even in this simple case (and only to terms in $1/T$ the elementary algebra was formidable, but the outcome pleasing in that quite definite conclusions emerged. That most of the paper is devoted to this special case must not blind us to the fact that these conclusions are far less important than the fact, very easily established at the start, that asymptotically the two approaches are equally efficient, statistically speaking.

Computationally, the argument overwhelmingly favours RF (with LS). In adopting RF we bypass all the problems associated with identification etc. Even as regards theory: in [1] the writer has seriously raised the problem as to whether <u>ex ante</u> RF (see (2)) does or does not represent a more valid cause-effect economic statement than does OF (1).

The first term (($1 + \beta^2_1$) of the error variance in the special and $E(\beta')^{-1}u'u\beta^{-1}/T$ in the general case) is the incubus. It goes far towards showing why forecasts of year-to-year changes are generally so poor (even with impressive R^2s and reassuring DWs). No effort should be spared to make all residual error variances as small as possible.

6 October 1967
Revised 24 November 1967 R. C. Geary

References

[1] R. C. Geary, "Some Remarks about Relations between Stochastic Variables: A discussion Document", <u>Review of the International Statistical Institute</u>, Volume 31: 2, 1963.

[2] T. Haavelmo, "The Statistical Implications of a Set of Simultaneous Equations. <u>Econometrica</u>, Vol. 11, No.1, 1943