COWLES COMMISSION DISCUSSION PAPER: STATISTICS: 336

August 1, 1949

The bias in single-equation methods of estimating behavior equations relating to a small sector of the economy.+

by

T. C. Koopmans

In a previous article,++ it was demonstrated by simple examples that the estimation of an economic relationship by methods based on the study of a single equation generally leads to biased estimates if the relationship studied is actually part of a system of relations describing the simultaneous formation of economic variables. The validity of the other relations of the system restricts the freedom of the variables entering into the relationship studied, and the nature of this restriction needs to be taken into account in constructing unbiased estimates of the parameters of that relationship.

In the present article we shall discuss the question, how far this consideration should affect studies that are concerned with only a small section of the economy, like the market for a single commodity, or a local labor market. It might be thought that such studies could safely be confined to the equations describing the formation of the variables specific to that section, like, for instance, the price and quantity of the commodity studied. One might argue that if broader economic categories like consumer's income and the general level of prices enter into the explanation of these specific variables, there is no need to add equations explaining the simultaneous formation of these general

+This is an unchanged reproduction of a paper written late in 1944 for the first CC Conference. It would probably be phrased differently if it were to be written afresh now.

++T. Koopmans, "Statistical estimation of simultaneous economic relations," Journal of the American Statistical Association, Vol. 40, 1945, CC New Series Papers, No. 11.

factors, because the small weight of the specific variables in the economy as a whole would permit us to neglect the reaction they have on the general factors through other equations. There is, however, a fallacy in this argument, as we shall show by the example of an equation system in which the particular equation on which interest is centered is simplified to the utmost.

Consider a system I of K equations of which the first connects just two "general" variables, money expenditure X on consumers' goods and consumers' income Y,

(1) $$X = \alpha Y + \beta + u, \quad ,$$

With proper units of measurement, the coefficient α approximately equals the income elasticity of expenditure. The term u represents a disturbance, subject to a stable probability distribution, as in the previous article. The remaining K-1 equations of the system need not be specified except that we assume that they are linear and involve both X and Y in such a way that (a) the single-equation least-squares estimate $a^{(X)}$ of α (with X as dependent variable) is biased by a certain percentage of the true value and (b) the system permits the identification of equation (1) as the consumption equation.[+]

Besides this system, consider another system II of K+P-1 equations obtained by splitting up expenditure X according to a number of expenditure items

(2) $$X = X_1 + X_2 + \ldots + X_P \quad ,$$

and replacing (1) by a set of P equations

(3) $$X_p = \alpha_p Y + \beta_p + u_p \,, \qquad p = 1,2,\ldots,P \,,$$

in which, of course,

(4) $$\alpha_1 + \ldots + \alpha_P = \alpha, \qquad \beta_1 + \ldots + \beta_P = \beta, \qquad u_1 + \ldots + u_P = u.$$

[+]This assumption means that it is not possible to derive from the remaining K-1 equations, by combination and elimination, another equation between X and Y which has the same form as (1). For a discussion of identification problems see T. C. Koopmans, "Identification Problems in Economic Model Construction", Econometrica, Vol. 17, 1949, CC New Series Papers, No. 31.

An analysis of this kind (without reference to the additional K-1 equations of the system) has been carried out recently by McCracken,[+] apparently by the single-equation least-squares method. It is well known that, if in the latter method the precaution is taken to include in the right hand member of each equation (3) the same variables[++] as occur in the right hand member in(1) and no others, the estimates[+++] $a_p^{(X)}$ of the α_p will be compatible with the estimate $a^{(X)}$ of α in that

5) $$a_1^{(X)} + a_2^{(X)} + \dots + a_p^{(X)} = a^{(X)} \quad .$$

We shall now compare the estimates $a^{(X)}$, $a_p^{(X)}$ with the corresponding "equation system estimates," i.e. the maximum likelihood estimates a, a_p , obtained by applying the simultaneous equations method to the systems I and II respectively.[++++] It is proved in an appendix to this article that the equation system estimates are also compatible in the sense of equation (5), provided no new conditions[+++++] are imposed on the variances and covariances $\sigma_{u_p u_q}$ of the disturbances u_p in the individual expenditure equation. In particular, in order to obtain compatible estimates, it will not be permissible to assume a priori that the disturbances u_p in expenditures for different items are independent.

If both the equation system estimates and the single equation estimates are compatible, it follows that the biases (in large samples) in the latter are also

---

[+] P. W. McCracken, "A hypothetical projection of expenditure for commodity groups based on past relationships to gross national product," U.S. Department of Commerce, Washington, 1943.

[++] In this connection, the constant term should be considered as the coefficient of a variable that assumes the constant value 1 for each time interval.

[+++] The notations $a^{(X)}$, $a_p^{(X)}$ are chosen to indicate least squares regression coefficients with X, X_p respectively as dependent variables.

[++++] In the system II, of course, X is wherever it occurs replaced by the right hand member of (2).

[+++++] Certain "old" conditions may arise from whatever conditions were imposed in the system I on the variance of the total disturbance u, or on its covariances with the disturbances in other equations of that system.

compatible. A given percentage bias in $a^{(X)}$ must in the average -- in some sense of the word average -- be reflected by a percentage bias of the same order of magnitude in the estimates $a_p^{(X)}$ for individual expenditure items. It follows that the mere smallness of an item is no ground for expecting a small percentage bias in the corresponding $a_p^{(X)}$. On the whole, the single equation method is as biased for the explanation of expenditure on the individual commodity as it is for total consumption expenditure. If there is any reason for expecting the single equation method to do better for any particular commodity, it can only be on grounds in some way specific to that commodity.

It is therefore of interest to study how the bias in the elasticity estimate $a^{(X)}$ of total expenditure is apportioned by items of expenditure. It is proved in the appendix that if u and u_p are not correlated with distrubances in the other K-1 equations common to the systems I and II, (equations which we have not specified), we have, in the limit for an infinitely large sample,

$$(6) \qquad Ea_p^{(X)} - \alpha_p = \frac{\sigma_{u_p u}}{\sigma_u^2} \left(Ea^{(X)} - \alpha\right),$$

if E denotes taking the mathematical expectation. This equation connects the biases in the single-equation least-squares estimates of total and individual income elasticities of expenditure. It satisfies the compatibility condition because

$$(7) \qquad \sum_p \sigma_{u_p u} = \sigma_{uu} = \sigma_u^2 .$$

In order to interpret (6) let us write it in a form involving relative instead of absolute biases in estimated elasticities,

$$(8) \qquad \frac{Ea_p^{(X)} - \alpha_p}{\alpha_p} = \frac{\sigma_{u_p} / \sigma_u}{\alpha_p / \alpha} \rho_{u_p u} \frac{Ea^{(X)} - \alpha}{\alpha} .$$

Here a " σ " with a single subscript denotes as before a standard deviation, " ρ " a correlation. The ratio σ_{u_p} / σ_u measures the degree of disturbance of the p-th individual expenditure equation relative to that of the total expenditure equation. The ratio α_p / α expresses what might be called the relative "marginal size" of the p-th expenditure item. What matters in this ratio is not the fraction of the total consumption expenditure devoted in the average to the p-th item, but the fraction of a change in the total expenditure (produced by a change in income) that is absorbed by the corresponding change in the p-th item.

It appears that if the single-equation least-squares method is applied, the percentage bias in the p-th individual elasticity (relative to the percentage bias in the total elasticity), is proportional to the relative disturbance in the p-th expenditure equation divided by the relative marginal size of the p-th item. An expenditure item that is highly disturbed in relation to its marginal size will be estimated with a large bias -- one other thing being equal. This other thing is the correlation $\rho_{u_p u}$ between the individual and the total disturbance. The higher this correlation, the greater the bias.

To illustrate this last factor, let us consider two extreme cases, in both of which there are P = 100 expenditure items, all of the same marginal size,

$$\alpha_p / \alpha = 1/100 \quad , \qquad p = 1,2,\ldots,100 \ , \tag{9}$$

and subject to the same degree of disturbance as measured by σ_{u_p}. In the first case we shall assume all individual disturbances to be actually independent (although such independence has not been assumed in the construction of the estimates a_p, a, by the maximum likelihood method!). The correlation $\rho_{u_p u}$ is then due only to the fact that the total disturbance u is the sum of individual disturbances u_p, and, because now $\sigma_{u_p u} = \sigma^2_{u_p}$ and therefore $\sum_p \sigma^2_{u_p} = \sigma^2_u$, we have

$$\sigma_{u_p} / \sigma_u = 1/10 \quad , \quad \rho_{u_p u} = 1/10 \ . \tag{10}$$

In the second case we assume perfect correlation between each individual disturbance and the total disturbance. Since then $\sum_p \sigma_{u_p} = \sigma_u$,

$$\sigma_{u_p} / \sigma_u = 1/100 \ , \quad \rho_{u_p u} = 1. \tag{11}$$

In both cases, therefore, each individual elasticity is estimated with the same percentage bias as the total elasticity. In the first case, however, the individual item is subject to a disturbance which is 10 times as large in relation to its marginal size. This large disturbance must be reflected in a large sampling standard deviation of the estimate $a_p^{(X)}$ as compared (in relative terms) with that of $a^{(X)}$. In the first case, it can therefore be said that going down to smaller items, although it does not decrease the amount of the percentage bias in $a_p^{(X)}$, nevertheless decreases its importance as compared with the sampling variation of $a_p^{(X)}$. Even this cannot be said in the second case, in which one common factor determines all disturbances in individual expenditure items.

It is likely that, in reality, the finer the breakdown of expenditure items, the less individual disturbances are correlated with each other and with the total disturbance, although there is no good reason to believe that this correlation tends to zero with increasing fineness of subdivision. Since, from (7),

(12) $$\frac{1}{\sigma_u^2} \sum_p \sigma_{u_p u} = \sum_p \frac{\sigma_{u_p}}{\sigma_u} \rho_{u_p u} = 1 ,$$

the tendency of $\rho_{u_p u}$ to decrease with finer subdivision must be compensated by a tendency of the relative degree of disturbance to increase in relation to the relative marginal size of the items (which also adds up to unity for all items), and it is due to this compensation that the "average" bias of the estimated elasticity for the individual item remains unaffected by the fineness of subdivision. Nevertheless, there is scope for great variation between items, and, to resume, one should expect a larger bias according as the disturbances in the expenditure for a certain item are (a) relatively larger or (b) more typical of the disturbances in consumption expenditure as a whole, and conversely.

The question arises in what amount of detail it is necessary to study the other relations of a complete system if the immediate purpose is the unbiased estimation of the coefficient of just one equation. This question indicates an important field for further research. In cases like that just discussed, where the individual expenditure equation contains no variables specific to the explanation of that item (e.g., prices of that and competing commodities) it may perhaps be sufficient if the particular equation studied is grafted on to what has been called a macro-economic system of equations, much like the systems developed by Tinbergen, which connect indices or totals of broad economic categories. This grafting can be done, for instance, by replacing (1) by a set of two equations, one referring to the one item studied, the other to the sum of all other items. Furthermore, it may not be necessary in all cases to perform a full-fledged statistical estimation of all parameters in the "other" equations. If the sole purpose is to correct for bias in the equation studied, it may be sufficiently accurate in some cases to have crude guesses of the other parameters, based on general economic information or expectation.

For an illustration of that possibility, let us go back once more to the single system of the previous article, consisting of a supply equation (13a) and a demand equation (13b) of a commodity

$$(13)\quad \begin{cases} (13a) & x_1 + \alpha_{s2}x_2 + \alpha_{s3}x_3 \qquad\qquad + \alpha_s = u_s \\ (13b) & x_1 + \alpha_{d2}x_2 \qquad\qquad + \alpha_{d4}x_4 + \alpha_d = u_d \, , \end{cases}$$

connecting quantity x_1 and price x_2 of a commodity with other factors x_3 and x_4 affecting supply and demand respectively. For the sake of argument let us suppose that this were a complete system. One might think, for instance, of x_3 as rainfall, and of x_4 as temperature, and make the unlikely assumption that no economic variables besides x_1 and x_2 enter the demand and supply equations (13) for this particular commodity. Alternatively, one might suppose that x_3 and x_4 are predetermined by earlier values of economic variables, and therefore strictly independent of the process of formation of x_1 and x_2 described by (13).

Now suppose that the value of $-\alpha_{s2} = -\alpha_s$ [+] (the price elasticity of supply) were known a priori, and that we wish to make use of this information in deriving those estimates of the other parameters that maximize the likelihood function under this restriction. By straightforward mathematical operations,[++] it is found that the so restricted maximum likelihood estimate $a^*_{d2} = a^*_d$ of the price elasticity of demand α_d satisfies

$$(14) \qquad a^*_d = -\frac{m_{11.4} + m_{12.4}\alpha_s}{m_{12.4} + m_{22.4}\alpha_s} ,$$

where

$$(15) \qquad m_{k\ell.4} = m_{k\ell} - \frac{m_{k4}m_{\ell 4}}{m_{44}} .$$

From equation (14) we read the following corroboration and extension of a result obtained in Section II of the previous article. If supply is completely inelastic (i.e. independent of simultaneous price, $\alpha_s = 0$), the (asymptotically unbiased) equation system estimate of the elasticity of demand coincides with the single equation estimate $a_d^{(2)}$ obtained by taking price (x_2) as dependent variable. If supply is infinitely elastic (i.e., the supply side permits only of one price, which may be associated with any supply, $\alpha_s = \infty$), the equation system estimate of the elasticity of demand coincides with the single equation estimate $a_d^{(1)}$ in which demand (x_1) is the dependent variable. If the supply curve has a finite positive slope (the normal case, $\alpha_s < 0$), the equation system estimate of the elasticity of supply is located <u>between</u> the two aforementioned estimates.

[+] Since all quantities "α" and "a" in what follows have "2" as second index, this index can conveniently be omitted.

[++] The simplest method to obtain this result is to change the scales of the equation (13) by the transformation

$$\alpha_{i1}' = 1/\sigma_i , \quad \alpha_{i2}' = \alpha_{i2}/\sigma_i , \quad \sigma_i' = 1, \quad i = s,d,$$

to maximize the likelihood function with respect to α_{d1}', α_{d2}', α_{d4}'; and to eliminate the estimate of α_{d4}'.

Since here demand and supply curves are only convenient names for what might be any two simultaneous economic relations, it is of interest to follow through the case where the a priori known value of α_s is positive. Then, supposing that both $a_d^{(1)}$ and $a_d^{(2)}$ are also positive, with $a_d^{(1)} < a_d^{(2)}$, we have[+]

$$
(16)\quad \begin{cases}
\text{if } 0 < \alpha_s < a_d^{(1)} & \text{then } a_d^{(2)} < a_d^* < \infty , \\
\text{if } \alpha_s = a_d^{(1)} & \text{then } a_d^* = \infty , \\
\text{if } a_d^{(1)} < \alpha_s < a_d^{(2)} & \text{then } -\infty < a_d^* < 0 , \\
\text{if } \alpha_s = a_d^{(2)} & \text{then } a_d^* = 0 , \\
\text{if } a_d^{(2)} < \alpha_s < \infty & \text{then } 0 < a_d^* < a_d^{(1)} .
\end{cases}
$$

Any assumption that limits the possible values of α_s can thus be translated into limits on the maximum likelihood estimate a_d^* of α_d. In the example of section II of the previous article, if $-\alpha_s = -\alpha_{s2}$ were known to fall between 0.4 and 1.2, it would follow in Case I that $a_d^* = a_{d2}^*$ fell between 0.37 and 0.48. Procedures of this kind may be found useful generally in assessing the effect of additional relations on the estimation of the parameters of one particular relation.

[+]The complete duality between α_s and a_d^* can be seen by writing (14) in the form

$$a_d^* + \alpha_s - \frac{a_d^* \alpha_s}{a_d^{(1)}} - a_d^{(2)} = 0 .$$

Of course the situation represented by the third line in (16) will have only a small probability of occurrence if the true value α_d sufficiently exceeds 0.