FRANKLIN M. FISHER

CAUSATION AND SPECIFICATION IN ECONOMIC THEORY AND ECONOMETRICS

I. INDIVIDUAL DECISION-MAKERS

In the micro-economics of perfect competition – at the level of the individual firm and the individual consumer – problems of causation seem straightforward to the economist. An individual consumer, for example, is assumed to come equipped with a good deal of information and a fully developed set of tastes. He takes prices as given and then decides how to allocate his expenditure so as to gain the most satisfaction. Similarly, a firm in perfect competition is assumed to know its technological possibilities. It takes the prices of factors and products as given and makes its production decision so as to gain the most profit. In both cases, the information provided by prices as to the possible opportunities is taken in by the decision-making unit as a primary stimulus; the consumption or production decisions are then the result.

It is true that in this simple picture, prices are not the only cause of the decision-maker's action. His tastes or his production opportunities are formed in some way by personal, social, and technological forces. All this, however, is assumed to have gone on in the past and not to be influenced by the current state of the market. If such forces change, they change sufficiently slowly and independently to allow constancy to be a satisfactory approximation in the analysis of short-run individual decisions.

The matter becomes less simple when the perfectly competitive assumption that prices can be taken as given is dropped. Even so, a pure monopolist presents no great problem. Here what is taken as given is not the price at which the product can be sold but rather the demand schedule facing the firm. Given that schedule, the monopolist decides simultaneously on output and price.

Oligopoly, however, leads to complications. In essence, two rational decision-makers with opposing interests and symmetrically placed must decide on prices and outputs in a situation in which the outcome for each depends heavily on the actions of the other. It is apparent that analysis

Synthese **20** (1969) 489–500; © *D. Reidel Publishing Co., Dordrecht-Holland*

is impossible without some assumption as to what each firm thinks about the behavior of the other. Moreover, since the behavior of A will depend on what he thinks B will do and vice versa, one is faced with the problem that A's action depends on his ideas about B's action which in turn depends on B's ideas about A's action which depends on A's ideas about B's ideas, and so forth. Either one assumes that the two rivals are less than fully aware of the situation and that each forms ideas about the other which a little experimentation can readily prove to be false or else there appears to be no solution. Even if A's ideas about B and B's ideas about A are not shown to be false in the very process of acting thereon, there will remain the fact that B has no reason to act in the way in which A believes him to act unless A continues to act on that belief.

Faced with this problem, economists have tended to assume each rival to have more or less complicated but specific ideas about the other or else have pointed to situations in which interests are not diametrically opposed and some mutually profitable action is possible. The pure problem is (and probably must be) unresolved and it is clear that here is a case in which the simple notion that causation at the individual level takes the form of stimulus and response in the internal processes of a decision-maker tends to break down.

Yet this, I think, is the only case of such a breakdown at that level. While it is possible to complicate the other models by the introduction of expectations, uncertainty, advertising, and so forth, in every case the decision-maker can reasonably be assumed to act on information provided by the external market and then to set the variables under his control in response.

II. MARKETS AND EQUILIBRIUM

The situation is not so straightforward when we consider how the information from the market is itself generated. Here the problem is immediately presented in the perfectly competitive case (monopoly, indeed, is easier to handle). Every firm and every consumer takes prices as given and makes purchasing and selling decisions accordingly. Yet how do those prices come to be set?

A standard answer in this regard is that price in each market takes on

just that value which will clear the market – that value at which producers can sell just what they wish to sell and consumers buy just what they wish to buy. At a more sophisticated level, prices of all goods and all factors in the economy are assumed to simultaneously take on such values that *all* plans on *all* markets are simultaneously fulfilled and such that the income generated in production for each household is just that which the household took as given in making its expenditure decisions. It has been shown that such general equilibrium prices exist under fairly general conditions.[1]

Obviously, this is not a satisfactory answer either in a single market or in the entire economy. Since not all prices are equilibrium prices, unless such equilibrium prices just happen to occur we are left with the question of how they come about. Here we have no very satisfactory theory. Nevertheless, we have a reasonable picture of what occurs. Suppose, for example, that demand falls short of supply in a particular market. Goods will then pile up on sellers' shelves. Sellers will realize that the costs of these inventories could be avoided by shading the price a little bit and some of them will do so. Price will then move downward reducing intended supply and increasing intended demand until equilibrium is reached. A similar thing happens when demand exceeds supply and buyers bid up prices. What is unsatisfactory about this is that we have no good model of the process which explains which sellers change the price or by how much, or why they do it rather than some other sellers. Indeed, we have no satisfactory analysis of how much, if any, goods get traded at disequilibrium prices or what effect such trades have on the market. Whereas the theory of outputs and purchases given prices operates at the individual level, theories of price adjustment are descriptive of market behavior and the fact that such adjustments must have an individual origin tends to be conveniently overlooked.

Nevertheless, this is a matter of the state of the art rather than of problems of causation. We have no good theory of how a particular seller is prompted to offer at a particular price different from the one he was assumed to take as given, but we know it is the inability to fulfill plans made at the original price which causes some seller to do so. The difficulty in a causal sense arises elsewhere. It arises because, having a good equilibrium theory of markets and a poor disequilibrium theory, empirically oriented econometric models tend to assume that price adjustments take

place very quickly and that many competitive markets only the average or cumulative results of which are observed are in equilibrium. This means that the equations describing such markets are (in the simplest case) taken to be: a demand curve, giving quantity demanded as a function of price; a supply curve, giving quantity supplied as a function of price; and a market-clearing identity stating that quantity demanded equals quantity supplied. Price is supposed to make these three equations hold for every observation, but no adjustment mechanism is supplied for it. Clearly, in this situation, so far as the model is concerned, price cannot be said to determine quantity or the other way round. The underlying causal structure may be straightforward, but at this level of aggregation it has been blurred, perhaps irretrievably so.

III. SIMULTANEOUS EQUATIONS IN ECONOMETRICS: NORMALIZATION RULES

This is by no means the only case in which the possession of a good equilibrium theory and a poor disequilibrium one leads to an econometric model in which causal relations are blurred by equilibrium conditions. To take an example from macro-economics, savings and investment decisions are made independently. Yet because income generated is income earned, observed saving is identically equal to observed investment. Aside from the confusion which these facts historically produced, most econometric models (but not all theoretical models) today assume that savings and investment plans will be simultaneously fulfilled and that income or some other variable will adjust so that they are.

At all levels, then, it has often proved convenient to write econometric models in terms of simultaneously holding equations in which the dependent variables from one equation appear as independent in others. Those dependent or 'endogenous' variables obviously are determined in a fairly complicated way in such models. Whatever the underlying mechanism, at this level of aggregation the nature of causation of such variables is at least blurred in the model. The usual terminology is to regard such endogenous variables as 'jointly determined' by the other variables in the model, but while this fairly describes the arithmetic involved, it is not a fair description of the underlying causal structure.

It is possible to argue about this in two ways. The first of these is on a

purely abstract plane, to ask whether such simultaneous models are or can be correct and to draw literal conclusions from the way in which such models are stated. The second way is to ask what difference it makes and to inquire what properties such models must have in order to be reasonable approximations to an underlying more straightforward causal process.

An example of the first approach is afforded by the question of the symmetric or asymmetric treatment of the endogenous variables in the estimation of the model. As just pointed out, if one takes the model literally, the causal statement which it permits is that the endogenous variables are jointly determined by the remaining 'predetermined' variables.[2] In the supply and demand example, quantity and price are jointly determined by consumer income, among other things. The model does not give price as a cause of quantity or quantity as a cause of price, or, if it does, it does so symmetrically. This view – that all endogenous variables are simultaneously determined – has been confused, however, with a somewhat similar-sounding but really different position, namely, that any equation of the model must treat all such variables appearing in it in symmetric fashion, that no natural normalization of any equation (by solving it for a particular endogenous variable in terms of the others) exists. Since some techniques in use for estimating the parameters of the equations require the imposition of a normalization rule while others impose symmetric treatment, this has been argued to be a valid criterion for choosing the latter estimators in preference to the former.[3]

There are two things wrong with this view. In the first place, what matters are the properties which an estimator has, not its meshing with causal notions in the abstract. Not enough is known about the properties of such estimators to be able to tell whether the imposition of a normalization rule is a gain or a loss.

Second, and more important for our purposes, the view that no natural normalization rule exists for an equation in a simultaneous model overlooks the genesis of such models and confuses joint determination with absolute symmetry. There are two ways to see this. First, suppose that we agree that simultaneous models in which equilibrium is assumed to hold are really only approximations to non-simultaneous models which reach equilibrium very quickly (we shall return to this below). In such non-simultaneous models, each equation gives the response of a set of

decision-makers to a pre-existing stimulus. That response serves as a new stimulus in other equations, and so forth, the process taking place with very short lags. There is clearly no question as to the existence of natural normalization rules in such equations; they are self-evident. Thus the equation which represents the purchase decisions of consumers is normalized for consumption, that representing the investment decisions of firms is normalized for investment, and so forth. It seems natural to keep those normalization rules when ignoring the short time lags and passing to the limit; indeed, it seems unnatural to do anything else.

Alternatively, one can look at the way in which simultaneous models are constructed. Typically, they are not seamless webs. Rather they are put together equation by equation. One can readily imagine a situation in which one equation of the model was different and the rest the same; one can imagine all equations but one different and that one the same. In particular, one can imagine experiments in which all equations but one are suppressed and all but one of the variables set by government fiat. The remaining variable would then be determined by the remaining equation. It does not take much to realize that there is generally a natural pairing of equations and variables in such experiments. Investment, for example, would be set by the equation describing investment behavior, and so forth. Indeed, the specification of the equations is arrived at by considering just such experiments: what would investment be if output, prices, and interest rates, say, were given? Yet the naturalness of a normalization rule cannot depend on the other equations in the model.

To this view it is sometimes objected that certain equations appear to have no natural normalization rules. Thus, while it is clear that the consumption equation ought to be normalized for consumption and the investment equation for investment, ought demand and supply equations to be normalized for quantities or prices? In one way or another, the demand and supply example is the only one in which ambiguity arises, and in that example what is at issue is not the existence of a normalization rule but ignorance as to its nature. We have already seen that we lack a satisfactory theory of disequilibrium price formation. Lacking that theory, it is often convenient to assume demand and supply in equilibrium. This leads to a situation in which both equations appear to be normalized for the same variable, namely quantity, but the difficulty arises merely because a great deal has been implicitly suppressed. Quantity demanded and

quantity supplied are different variables and there is no difficulty in saying that the demand equation determines the first and the supply equation the second, provided that we add an equation saying how price comes to be set to make them equal. It is the absence of a price-formation equation which is the misspecification, not the statement that normalization rules exist.[4]

IV. SIMULTANEOUS MODELS AND CAUSATION

If the argument over normalization rules just discussed stems from a literal reading of simultaneous models as indivisible wholes, a rather more important argument has concerned whether such models can be literally true. It is clear that, taking such models as written, the nature of causation in them is at best obscure. It can be argued that the statement that all endogenous variables are jointly determined by the predetermined variables is an evasion and that such models show the endogenous variables simultaneously causing each other, a situation which is at best hard to understand. This view, pressed energetically by Wold, in particular[5], has led in two related directions.

The first such direction is the view that simultaneous models cannot be correct, that true models must of necessity involve a unidirectional causal flow. In imposing simultaneity, this view holds, econometricians make a serious error. All models ought to be formulated in a non-simultaneous, recursive fashion. This view seems to me to be correct as regards the underlying nature of economic processes, which are certainly not simultaneous; nevertheless, it seems to me not to address the correct question. That question is not whether simultaneous models can be taken literally but rather whether they can be appropriately considered as limiting approximations to underlying non-simultaneous models.

The second such direction is the reinterpretation of simultaneous models so that causation is once again unidirectional. Put rather too simply, that reinterpretation[6] takes the form of stating that the variables on the right-hand side of a given equation should be interpreted not as the variables themselves but as the values of those variables forecast by the model. Thus, quantity demanded, for example, reacts not to price but to predicted price, the prediction being itself generated from the model. This view has a consequence for estimation; Wold [15] has proposed an esti-

mator with the property that the values of the parameters estimated lead to a fixed point in the following sense. Take the equations of the model and consider the values of the endogenous variables predicted, given the predetermined variables. Now take any single equation of the model in a naturally normalized form. Insert the predicted values of the right-hand side endogenous variables and generate the value of the normalized variable. The fixed-point property is that the values so generated should be the same as those predicted by all equations together.

The properties of that fixed-point estimator are not yet fully understood. Some of them seem desirable and there are some serious problems.[7] Whether that estimator is a useful one, however, seems to me to turn on its properties and not on its genesis in terms of a reinterpretation of the seemingly inconsistent causal structure of simultaneous models. That reinterpretation is not one with which most econometricians agree in any case. It is not necessary if one regards simultaneous models as approximations.

V. SIMULTANEOUS MODELS AS APPROXIMATIONS

As has already been indicated, the latter view seems the most natural one. We observe economic variables as sums or averages over relatively large periods such as years or, at best, weeks or days. Even if the true underlying disequilibrium process is not simultaneous, if its adjustment to equilibrium is sufficiently rapid, the model framed in terms of observed variables may differ insignificantly from simultaneity. Thus, demand and supply are not always in equilibrium at every instant; if prices adjust sufficiently rapidly, however, only a small error is committed in assuming that demand and supply balance over the course of a year.[8] If one takes this view, there is no difficulty in reconciling the apparent multi-directional causation of simultaneous models with the stimulus-response models of individual decision-makers which economists find most familiar; simultaneous models are merely approximations.

The matter clearly cannot end there, however, for the issue arises of whether this view of simultaneous equation models has any consequence for the specification and estimation thereof.

For estimation, the question is the following. Suppose we could observe variables at very finely divided points of time. Consider the esti-

mators which we would then construct. Now pass to the limit and ask whether those estimators approach those which in fact we use for simultaneous models. This question has never been fully formally answered for the kind of approximation which we have just been discussing in which the observations are sums or averages, but there seems little doubt that for reasonable specifications of the limiting process, the answer is in the affirmative.[9] For a related problem in which the observations are taken only at discrete points of time, we know this to be true if the properties of the random disturbances (which economists put in models to account for the many independent and, one hopes, small effects inevitably left out) are also smooth as the limit is approached.[10]

A rather different recent development concerns the admissibility of a given simultaneous model as such a limit.[11] If a non-simultaneous process is to generate a simultaneous model in the time averages of the variables as the time intervals involved approach 0, it is evident that some restriction as to the stability of that process is involved. If one also requires that the simultaneous approximation be reasonably robust against small changes in the exact structure of the omitted lags, then rather strong conditions on the model itself can be derived. Thus, to take a linear example, suppose that the true model is:

$$(1) \qquad Y_{t+k\Delta\theta} = BY_{t+(k-1)\Delta\theta} + H_t,$$

where Y is a vector of endogenous variables; B is a matrix of parameters; H_t represents the influence of the predetermined variables; t is the observation period; and $\Delta\theta$ is the length of the true time interval involved in the reactions of the model. The observed variables are:

$$(2) \qquad \bar{Y}_t = \sum_{k=1}^{n} Y_{t+k\Delta\theta}\Delta\theta,$$

where $n = 1/\Delta\theta$. Then:

$$(3) \qquad \bar{Y}_t = B\bar{Y}_t + H_t + B(Y_t - Y_{t+n\Delta\theta})\,\Delta\theta.$$

As $\Delta\theta$ goes to 0 and n goes to infinity, this will approach the simultaneous model:

$$(4) \qquad \bar{Y}_t = B\bar{Y}_t + H_t$$

if and only if

$$\text{(5)} \qquad \lim_{n\to\infty} \frac{Y_t - Y_{t+n\Delta\theta}}{n} = 0.$$

It is not hard to show, however, that this occurs if and only if the matrix B has all its eigenvalues in or on the unit circle, with plus one not an eigenvalue. A simultaneous model in which this turns out not to be the case cannot be such a limit.

Moreover, even stronger restrictions can be generated by such considerations if we take the view, as earlier, that simultaneous models are made up of individual equations which can be suppressed in thought-experiments. If we suppress one or more equations and imagine the corresponding endogenous variables set by fiat, then the remaining part of the model becomes itself a simultaneous model. Since the parameters of that remaining submodel would be unaltered in such a case, such an experiment could only be valid if that submodel satisfied the same conditions as did the full model, that is, if the submodel were itself capable of being the limit of a non-simultaneous process. Thus, in the linear case, for example, we obtain the result that not only the matrix B itself, but also every principal submatrix thereof must satisfy the eigenvalue condition already mentioned.

Clearly, this appears to be a very strong requirement[12], although how restrictive it is in practice remains to be discovered. If it is strong, the possibility exists that it will prove a useful tool in testing and specifying simultaneous equation models. Thus, if a particular submodel fails the test, attention ought to be paid to the specification of the equations of that submodel, because there is something inconsistent therein. Such tests are particularly interesting because they are the only ones internal to the model (that is, not involving forecasting experiments) which relate directly to the interrelations of the various equations rather than to each equation separately.

How useful all this is in practice is an empirical matter on which work is just beginning. To the extent that it is, it represents an attempt to ensure that models built in a now standard form are consistent with the notions of causation which underlie them.

Massachusetts Institute of Technology

BIBLIOGRAPHY

[1] Ando, A., F. M. Fisher, and H. A. Simon, *Essays on the Structure of Social Science Models*, M.I.T. Press, Cambridge, Mass., 1963.
[2] Bentzel, R. and B. Hansen, 'On Recursiveness and Interdependency in Economic Models', *Review of Economic Studies* **22** (1954–55) 153–168.
[3] Chow, G. C., 'A Comparison of Alternative Estimators for Simultaneous Equations', *Econometrica* **32** (1964) 532–553.
[4] Debreu, G., *Theory of Value*, John Wiley & Sons, New York, 1959.
[5] Fisher, F. M., 'Dynamic Structure and Estimation in Economy-Wide Econometric Models', in *The Brookings Quarterly Econometric Model of the United States* (ed. by J. S. Duesenberry *et al.*), Rand McNally & Co., Chicago, and North-Holland Publishing Co., Amsterdam, 1965, Chapter 15.
[6] Fisher, F. M., 'A Correspondence Principle for Simultaneous Equation Models', M.I.T. Department of Economics Working Paper No. 9, November 1967 (forthcoming in *Econometrica*).
[7] Gorman, W. M., 'Professor Strotz on a Specification Error', *Discussion Papers* (University of Birmingham, Faculty of Commerce and Social Science), *Series A*, **24** (1960).
[8] Koopmans, T. C., *Three Essays on the State of Economic Science*, McGraw-Hill Book Co., New York, 1957.
[9] Lyttkens, E., 'Non-iterative Estimation of Special Interdependent System by GEID-Specification', paper presented at the Blaricum meetings of the Econometric Society, January 4–6, 1967.
[10] Samuelson, P. A., 'Some Notions on Causality and Teleology in Economics', in *Cause and Effect* (ed. by D. Lerner), The Free Press, New York, 1965, pp. 99–144.
[11] Simon, H. A., 'Causal Ordering and Identifiability', in *Studies in Econometric Method* (ed. by Wm. C. Hood and T. C. Koopmans) (Cowles Commission Monograph 14), John Wiley & Sons, New York, 1953, Chapter 3. Reprinted as Chapter 1 of H. A. Simon, *Models of Man*, John Wiley & Sons, New York, 1957, and as Chapter 2 of [1].
[12] Strotz, R. H., 'Interdependence as a Specification Error', *Econometrica* **28** (1960) 428–442.
[13] Wold, H. O. A. in association with L. Juréen, *Demand Analysis, A Study in Econometrics*, John Wiley & Sons, New York, 1953.
[14] Wold, H. O. A., 'Forecasting by the Chain Principle', in *Econometric Model Building: Essays on the Causal Chain Approach* (ed. by H. O. A. Wold), North-Holland Publishing Co., Amsterdam, 1964, Chapter 1.
[15] Wold, H. O. A., 'A Fix-Point Theorem with Econometric Background', *Arkiv für Matematik* **6** (1965) 209–240.

REFERENCES

[1] See Debreu [4]. A good exposition is given in Koopmans [8], Essay 1.
[2] If the model has a particular 'block triangular' structure, some jointly determined variables are determined causally prior to others which they help to determine. See Simon [11]. Ando *et al.* [1] discusses the consequences of such structures and related ones for the analysis of dynamic systems.

[3] See Chow [3]. For a fuller discussion of simultaneous equation estimation including some of the questions here covered, see Fisher [5].

[4] Naturally, in some models it makes sense to normalize one or the other of the demand and supply equations for price instead of quantity. This depends on the behavior supposed to be represented in those equations. Thus, if the supply equation represents behavior of sellers called upon to supply a certain output and naming a price for so doing, it should be normalized for price.

[5] Wold and Juréen [13], and other writings.

[6] Also due to Wold [14].

[7] See Lyttkens [9].

[8] This view was put forth in Bentzel and Hansen [2]. An alternate view is that we observe variables only at discrete moments in time and that this leads to simultaneity (see Strotz [12]) but this seems less in accord with the nature of the variables generally used.

[9] See Samuelson [10], p. 139.

[10] See Strotz [12] and, especially, Gorman [7]. What appears to be involved is the specification that as time lags approach 0 the correlation between a disturbance and its immediately past value approaches 1.

[11] See Fisher [6] for a complete discussion.

[12] It can be generalized to non-linear models. See [6].