Identification Problems in Economic Model Construction

Author(s): Tjalling C. Koopmans

Source: *Econometrica*, Vol. 17, No. 2 (Apr., 1949), pp. 125-144
Published by: The Econometric Society
Stable URL: https://www.jstor.org/stable/1905689
Accessed: 06-07-2019 18:32 UTC

IDENTIFICATION PROBLEMS IN ECONOMIC MODEL CONSTRUCTION[1]

By TJALLING C. KOOPMANS

Statistical inference, from observations to economic behavior parameters, can be made in two steps: inference from the observations to the parameters of the assumed joint distribution of the observations, and inference from that distribution to the parameters of the structural equations describing economic behavior. The latter problem of inference, described by the term "identification problem," is discussed in this article in an expository manner, drawing on other more original work for concepts and theorems, and using a number of examples drawn partly from econometric literature.

1. INTRODUCTION

THE CONSTRUCTION of dynamic economic models has become an important tool for the analysis of economic fluctuations and for related problems of policy. In these models, macro-economic variables are thought of as determined by a *complete system of equations*. The meaning of the term "complete" is discussed more fully below. At present it may suffice to describe a complete system as one in which there are as many equations as endogenous variables, that is, variables whose formation is to be "explained" by the equations. The equations are usually of, at most, four kinds: equations of economic behavior, institutional rules, technological laws of transformation, and identities. We shall use the term structural equations to comprise all four types of equations.

Systems of structural equations may be composed entirely on the basis of economic "theory." By this term we shall understand the combination of (a) principles of economic behavior derived from general observation—partly introspective, partly through interview or experience—of the motives of economic decisions, (b) knowledge of legal and institutional rules restricting individual behavior (tax schedules, price controls, reserve requirements, etc.), (c) technological knowledge, and (d) carefully constructed definitions of variables. Alternatively, a structural equation system may be determined on the dual basis of such "theory" combined with systematically collected statistical data for the relevant variables for a given period and country or other unit. In this article we shall discuss certain problems that arise out of model construction in the second case.

[1] This article will be reprinted as Cowles Commission Paper, New Series, No. 31. I am indebted to present and former Cowles Commission staff members and to my students for valuable critical comments regarding contents and presentation of this article. An earlier version of this paper was presented before the Chicago Meeting of the Econometric Society in December 1947.

Where statistical data are used as one of the foundation stones on which the equation system is erected, the modern methods of statistical inference are an indispensable instrument. However, without economic "theory" as another foundation stone, it is impossible to make such statistical inference apply directly to the equations of economic behavior which are most relevant to analysis and to policy discussion. Statistical inference unsupported by economic theory applies to whatever statistical regularities and stable relationships can be discerned in the data.[2] Such purely empirical relationships when discernible are likely to be due to the presence and persistence of the underlying structural relationships, and (if so) could be deduced from a knowledge of the latter. However, the direction of this deduction cannot be reversed—from the empirical to the structural relationships—except possibly with the help of a theory which specifies the form of the structural relationships, the variables which enter into each, and any further details supported by prior observation or deduction therefrom. The more detailed these specifications are made in the model, the greater scope is thereby given to statistical inference from the data to the structural equations. We propose to study the limits to which statistical inference, from the data to the structural equations (other than definitions), is subject, and the manner in which these limits depend on the support received from economic theory.

This problem has attracted recurrent discussion in econometric literature, with varying terminology and degree of abstraction. Reference is made to Pigou [**16**], Henry Schultz [**17**, especially Chapter II, Section IIIc], Frisch [**4, 5**], Marschak [**15**, especially Sections IV and V], Haavelmo [**6**, especially Chapter V]. An attempt to systematize the terminology and to formalize the treatment of the problem has been made over the past few years by various authors connected in one way or another with the Cowles Commission for Research in Economics. Since the purpose of this article is expository, I shall draw freely on the work by Koopmans and Rubin [**14**], Wald [**18**], Hurwicz [**7, 8**], Koopmans and Reiersöl [**13**], without specific acknowledgement in each case. We shall proceed by discussing a sequence of examples, all drawn from econometrics, rather than by a formal logical presentation, which can be found in references [**14**], [**7**] and [**13**].

2. CONCEPTS AND EXAMPLES

The *first example*, already frequently discussed, is that of a competitive market for a single commodity, of which the price p and the quantity q are determined through the intersection of two rectilinear schedules, of demand and supply respectively, with instantaneous response of quantity

[2] See T. C. Koopmans [**12**].

to price in both cases. For definiteness' sake, we shall think of observations as applying to successive periods in time. We shall further assume that the slope coefficients α and γ of the demand and supply schedules respectively are constant through time, but that the levels of the two schedules are subject to not directly observable shifts from an equilibrium level. The structural equations can then be written as:

$$(1)\quad \begin{cases} (1d) \qquad q + \alpha p + \epsilon = u \quad \text{(demand)} \\ (1s) \qquad q + \gamma p + \eta = v \quad \text{(supply)}. \end{cases}$$

Concerning the shift variables u and v we shall assume that they are random drawings from a stable joint probability distribution with mean values equal to zero:

$$(2)\qquad \phi(u, v), \qquad \mathcal{E}u = 0, \qquad \mathcal{E}v = 0.$$

We shall introduce a few terms which we shall use with corresponding meaning in all examples. The not directly observable shift variables u, v are called *latent variables,* as distinct from the *observed variables,* p, q. We shall further distinguish *structure* and *model.* By a structure we mean the combination of a specific set of structural equations (1) (such as is obtained by giving specific numerical values to α, γ, ϵ, η) and a specific distribution function (2) of the latent variables (for instance a normal distribution with specific, numerically given, variances and covariance). By a model we mean only a specification of the form of the structural equations (for instance their linearity and a designation of the variables occurring in each equation), and of a class of functions to which the distribution function of the latent variables belongs (for instance, the class of all normal bivariate distributions with zero means). More abstractly, a model can be defined as a set of structures. For a useful analysis, the model will be chosen so as to incorporate relevant a priori knowledge or hypotheses as to the economic behavior to be described. For instance, the model here discussed can often be narrowed down by the usual specification of a downward sloping demand curve and an upward sloping supply curve:

$$(3)\qquad \alpha > 0, \qquad \gamma < 0.$$

Let us assume for the sake of argument that the observations are produced by a structure, to be called the "true" structure, which is contained in (permitted by) the model. In order to exclude all questions of sampling variability (which are a matter for later separate inquiry), let us further make the unrealistic assumption that the number of observations produced by this structure can be increased indefinitely. What inferences can be drawn from these observations toward the "true" structure?

A simple reflection shows that in our present example neither the "true" demand schedule nor the "true" supply schedule can be determined from any number of observations. To put the matter geometrically, let each of the two identical scatter diagrams in Figures 1A and 1B represent the jointly observed values of p and q. A structure compatible with these observations can be obtained as follows: Select arbitrarily "presumptive" slope coefficients α and γ of the demand and supply schedules. Through each point $S(p, q)$ of the scatter diagrams draw two straight lines with slopes given by these coefficients. The presumptive

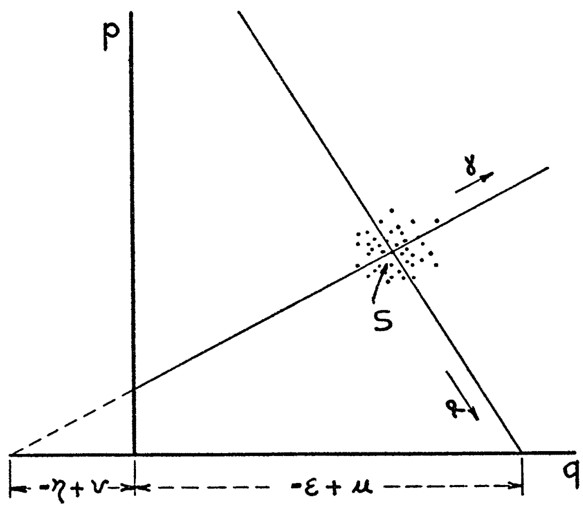

FIG. 1A

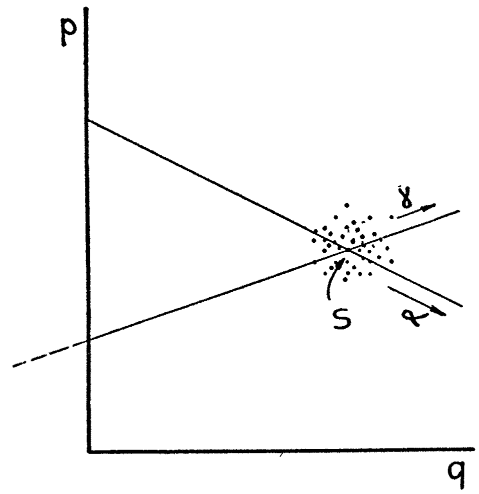

FIG. 1B

demand and supply schedules will intersect the quantity axis at distances $-\epsilon + u$ and $-\eta + v$ from the origin, *provided* the presumptive slope coefficients α and γ are the "true" ones. We shall assume this to be the case in Figure 1A. In that case the values of ϵ and η can be found from the consideration that the averages of u and v in a sufficiently large sample of observations are practically equal to zero.

However, nothing in the situation considered permits us to distinguish the "true" slopes α, γ (as shown in Figure 1A) from any other presumptive slopes (as illustrated in Figure 1B). Any arbitrary set of slope coefficients α, γ (supplemented by corresponding values ϵ, η of the intercepts) represents another, statistically just as acceptable, hypothesis concerning the formation of the observed variables.

Let us formulate the same remark algebraically in preparation for further examples in more dimensions. Let the numerical values of the "true" parameters $\alpha, \gamma, \epsilon, \eta$ in (1) be known to an individual who, taking delight in fraud, multiplies the demand equation (1d) by 2/3, the supply equation (1s) by 1/3, and adds the result to form an equation

$$(4d) \qquad q + \frac{2\alpha + \gamma}{3} p + \frac{2\epsilon + \eta}{3} = u',$$

which he proclaims to be the demand equation. This equation is actually different from the "true" demand equation (1d) because (3) implies $\alpha \neq \gamma$. Similarly he multiplies the same equations by 2/5 and 3/5 respectively, say, to produce an equation

$$\text{(4s)} \qquad q + \frac{2\alpha + 3\gamma}{5} p + \frac{2\epsilon + 3\eta}{5} = v',$$

different from the "true" supply equation (1s), but which he presents as if it were the supply equation. If our prankster takes care to select his multipliers in such a manner as not to violate the sign rules (3) imposed by the model, the deceit cannot be discovered by statistical analysis of any number of observations.[3] For the equations (4), being derived from (1), are satisfied by all data that satisfy the "true" equations (1). Moreover, being of the same form as the equations (1), the equations (4) are equally acceptable a priori.

Our *second example* differs from the first only in that the model specifies a supply equation containing in addition an exogenous variable. To be definite, we shall think of the supply of an agricultural product as affected by the rainfall r during a critical period of crop growth[4] or crop gathering. This variable is called exogenous to our model to express the plausible hypothesis that rainfall r, while affecting the market of the commodity concerned, is not itself affected thereby. Put in mathematical terms, this hypothesis specifies that the disturbances u and v in

$$\text{(5)} \quad \begin{cases} \text{(5d)} \qquad q + \alpha p \qquad\quad + \epsilon = u \quad \text{(demand)} \\ \text{(5s)} \qquad q + \gamma p + \delta r + \eta = v \quad \text{(supply)} \end{cases}$$

are statistically independent[5] of the values assumed by r.

It will be seen at a glance that the supply equation still cannot be determined from a sample of any size. If, starting from "true" structural equations (5) we multiply by $-1/2$ and $3/2$, say, and add the results to obtain a pretended supply equation,

$$\text{(6s)} \qquad q + \frac{3\gamma - \alpha}{2} p + \frac{3\delta}{2} r + \frac{3\eta - \epsilon}{2} = v'$$

[3] The deceit could be discovered if the model were to specify a property (e.g., independence) of the disturbances u and v, which is not shared by $u' = (2u + v)/3$ and $v' = (2u + 3v)/5$. We have not made such a specification.

[4] With respect to this example, the assumption of a linear relationship can be maintained only if we think of a certain limited range of variation in rainfall. Another difficulty with the example is that for most agricultural products, the effect of price on supply is delayed instead of instantaneous, as here assumed. A practically instantaneous effect can, however, be expected in the gathering of wild fruits of nature.

[5] It is immaterial for this definition whether the exogenous variable is regarded as a given function of time—a concept perhaps applicable to a variable set by government policy—or as itself a random variable determined by some other

of the same prescribed form as (5s), any data will satisfy this equation (6s) as well as they satisfy the two equations (5).

A similar reasoning can *not* be applied to the demand equation in the present model. Any attempt to construct another pretended demand equation by a linear combination involving the supply equation (5s) would introduce into that pretended demand equation the variable r which by the hypotheses underlying the model does not belong in it.

It might be thought that, if r has the properties of a random variable, its presence in the pretended demand equation might be concealed because its "contribution" cannot be distinguished from the random disturbance in that equation. To be specific, if $4/3$ and $-1/3$ are arbitrarily selected multipliers, the disturbance in the pretended demand equation might be thought to take the form

$$u' = \frac{4u - v}{3} - \frac{\delta}{3} r.$$

This, however, would violate the specification that r is exogenous and that therefore r and u' are to be statistically independent as well as r and (u, v). The relevance of the exogenous character of r to our present discussion is clearly illustrated by this remark.

Our analysis of the second example suggests (and below we shall cite a theorem establishing proof) that a sufficiently large sample does indeed contain information with regard to the parameters α, ϵ of the demand equation (it being understood that such information is conditional upon the validity of the model). It can already be seen that there must be the following exception to the foregoing statement. If in fact (although the model does not require it) rainfall has no influence on supply, that is, if in the "true" structure $\delta = 0$, then any number of observations must necessarily be compatible with the model (1), and hence does not convey information with regard to either the demand equation or the supply equation.

As a *third example* we consider a model obtained from the preceding one

---

structure involving probability distributions—a concept applicable particularly to weather variables. It should further be noted that we postulate independence between r and (u, v), not between r and (p, q), although we wish to express that r "is not affected by" p and q. The meaning to be given to the latter phrase is that in other equations explaining the formation of r the variables (p, q) do not enter. Precisely this is implied in the statistical independence of r and (u, v), because (p, q) is, by virtue of (5), statistically dependent on (u, v), and any role of (p, q) in the determination of r would therefore create statistical dependence between r and (u, v). On the other hand, the postulated statistical independence between r and (u, v) is entirely compatible with the obvious influence, by virtue of (5), of r on (p, q).

by the inclusion in the demand equation of consumers' income i as an additional exogenous variable. We assume the exogenous character of consumers' income merely for reasons of exposition, and in full awareness of the fact that actually price and quantity on any market do affect income directly to some extent, while furthermore the disturbances u and v affecting the market under consideration may well be correlated with similar disturbances in several other markets which together have a considerably larger effect on consumers' income.

The structural equations are now

$$(7)\quad \begin{cases} (7\text{d}) & q + \alpha p + \beta i \qquad\quad + \epsilon = u \quad \text{(demand)} \\ (7\text{s}) & q + \gamma p \qquad\quad + \delta r + \eta = v \quad \text{(supply).} \end{cases}$$

Since each of the two equations now excludes a variable specified for the other equation, neither of them can be replaced by a different linear combination of the two without altering its form. This suggests, and proof is cited below, that from a sufficiently large sample of observations, the demand equation can be accurately determined provided rainfall actually affects supply ($\delta \neq 0$), and the supply equation can be determined provided consumers' income actually affects demand ($\beta \neq 0$).

The *fourth example* is designed to show that situations may occur in which some but not all parameters of a structural equation can be determined from sufficiently many observations. Let the demand equation contain both this year's income i_0 and last year's income i_{-1}, but let the supply equation not contain any variable absent from the demand equation:

$$(8)\quad \begin{cases} (8\text{d}) & q + \alpha p + \beta_0 i_0 + \beta_{-1} i_{-1} + \epsilon = u \\ (8\text{s}) & q + \gamma p \qquad\qquad\qquad\quad + \eta = v. \end{cases}$$

Now obviously we cannot determine either α or ϵ, because linear combinations of the equations (8) can be constructed which have the same form as (8d) but other[6] values α' and ϵ' for the coefficients α and ϵ. However, as long as (8d) enters with some nonvanishing weight into such a linear combination, the ratio β_{-1}/β_0 is not affected by the substitution of that linear combination for the "true" demand equation. Thus, if the present model is correct, the observations contain information with respect to the relative importance of present and past income to demand, whereas they are silent on the price elasticity of demand.

The *fifth example* shows that an assumption regarding the joint distribution of the disturbances u and v, where justified, may open the door to a determination of a structural equation which is otherwise indeterminate. Returning to the equation system (5) of our second example, we shall now make the model specify in addition that the

[6] As regards ϵ' this is true whenever $\epsilon \neq \eta$. As regards α' it is safeguarded by (3).

disturbances u in demand and v in supply are statistically independent. Remembering our previous statement that the demand equation can already be determined without the help of such an assumption, it is clear that in attempting to construct a "pretended" supply equation, no linear combination of the "true" demand and supply equations (5), other than the "true" supply equation (5s) itself, can be found which preserves the required independence of disturbances in the two equations. Writing λ and $1 - \lambda$ for the multipliers used in forming such a linear combination, the disturbance in the pretended supply equation would be

$$v' = \lambda u + (1 - \lambda)v. \tag{9}$$

Since u and v are by assumption independent, the disturbance v' of the pretended supply equation is independent of the disturbance u in the demand equation already found determinable, if and only if $\lambda = 0$, i.e., if the pretended supply equation coincides with the "true" one.

We emphasize again the expository character of the foregoing examples. It has already been indicated that the income variable i is not truly exogenous. By assuming it to be so, we have held down the size of the equation system underlying our discussion, and we may as a result have precluded ourselves from seeing indeterminacies that could come to light only by a study of all relationships participating in the formation of the variables involved. It will therefore be necessary to develop criteria by which indeterminacies of the coefficients of larger equation systems can be detected. Before discussing such criteria for linear systems, we shall formalize a few of the concepts used or to be used.

3. THE IDENTIFICATION OF STRUCTURAL PARAMETERS

In our discussion we have used the phrase "a parameter that can be determined from a sufficient number of observations." We shall now define this concept more sharply, and give it the name *identifiability* of a parameter. Instead of reasoning, as before, from "a sufficiently large number of observations" we shall base our discussion on a hypothetical knowledge of the probability distribution of the observations, as defined more fully below. It is clear that exact knowledge of this probability distribution cannot be derived from any finite number of observations. Such knowledge is the limit approachable but not attainable by extended observation. By hypothesizing nevertheless the full availability of such knowledge, we obtain a clear separation between problems of statistical inference arising from the variability of finite samples, and problems of identification in which we explore the limits to which inference even from an infinite number of observations is subject.

A *structure* has been defined as the combination of a distribution of latent variables and a complete set of structural equations. By a *complete* *set of equations* we mean a set of as many equations as there are endogenous variables. Each endogenous variable may occur with or without time lags, and should occur without lag in at least one equation. Also, the set should be such as to permit unique determination of the nonlagged values of the endogenous variables from those of the lagged endogenous, the exogenous, and the latent variables. Finally, by *endogenous variables* we mean observed variables which are not exogenous, i.e., variables which are not known or assumed to be statistically independent of the latent variables, and whose occurrence in one or more equations of the set is necessary on grounds of "theory."

It follows from these definitions that, for any specific set of values of the exogenous variables, the distribution of the latent variables (i.e., one of the two components of a given structure) entails or generates, through the structural equations (i.e., the other component of the given structure), a probability distribution of the endogenous variables. The latter distribution is, of course, conditional upon the specified values of the exogenous variables for each time point of observation. This conditional distribution, regarded again as a function of all specified values of exogenous variables, shall be the hypothetical datum for our discussion of identification problems.

We shall call two structures S and S' (observationally) *equivalent* (or indistinguishable) if the two conditional distributions of endogenous variables generated by S and S' are identical for all possible values of the exogenous variables. We shall call a structure S permitted by the model (uniquely) *identifiable* within that model if there is no other equivalent structure S' contained in the model. Although the proof has not yet been completely indicated, it may be stated in illustration that in our third example almost all structures permitted by the model are identifiable. The only exceptions are those with either $\beta = 0$ or $\delta = 0$ (or both). In the first and second examples, however, no structure is identifiable, although in the second example, we have stated that the demand equation by itself is determinate. To cover such cases we shall say that a certain parameter θ of a structure S is uniquely *identifiable* within a model, if that parameter has the same value for all structures S' equivalent to S, contained in the model. Finally, a *structural equation* is said to be *identifiable* if all its parameters are identifiable.

This completes the formal definitions with which we shall operate. They can be summarized in the statement that anything is called identifiable, the knowledge of which is implied in the knowledge of the distribution of the endogenous variables, given the model (which is accepted as valid). We now proceed to a discussion of the application of this concept to linear models of the kind illustrated by our examples.

4. IDENTIFIABILITY CRITERIA IN LINEAR MODELS

In our discussion of these examples, it has been possible to conclude that a certain structural equation is not identifiable whenever we are able to construct a different equation, obtained by linear combination of some or all structural equations, which likewise meets the specifications of the model. In the opposite case, where we could show that no such different linear combination exists, we could not yet conclude definitely that the equation involved is identifiable. Could other operations than linear combination, perhaps be used to derive equations of the same form?

We shall now cite a theorem which establishes that no such other operations can exist. The theorem relates to models specifying a complete set of structural equations as defined above, and in which a given set of endogenous and exogenous variables enters linearly. Any time lags with which these variables may occur are supposed to be integral multiples of the time interval between successive observations. Furthermore the exogenous variables (considered as different variables whenever they occur with a different time lag) are assumed not to be linearly dependent, i.e., in the functional sense.[7] Finally, although simultaneous disturbances in different structural equations are permitted to be correlated, it is assumed that any disturbances operating in different time units (whether in the same or in different structural equations) are statistically independent.

Suppose the model does not specify anything beyond what has been stated. That is, no restrictions are specified yet that exclude some of the variables from specific equations. Obviously, with respect to such a broad model, not a single structural equation is identifiable. However, a theorem has been proved **[14]** to the effect that, given a structure S within that model, any structure S' in the model, equivalent to S, can be derived from S by replacing each equation by some linear combination of some or all equations of S.

It will be clear that this theorem remains true if the model is narrowed down by excluding certain variables from certain equations, or by other restrictions on the parameters. Thus, whenever in our examples we have concluded that different linear combinations of the same form prescribed for a structural equation did not exist, we have therewith established the identifiability of that equation. More in general, the analysis of the identifiability of a structural equation in a linear model consists in a

[7] The criteria of identifiability to be stated would require amended formulation if certain identities involving endogenous variables would be such that each variable occuring in them also occurs, in some equation of the complete set, with a time lag, and if this time lag were the same for all such variables. In this case, a complication arises from linear (functional) dependence among lagged endogenous (and possibly exogenous) variables.

study of the possibility to produce a different equation of the same prescribed form by linear combination of all equations. If this is shown to be impossible, the equation in question is thereby proved to be identifiable. To find criteria for the identifiability of a structural equation in a linear model is therefore a straightforward mathematical problem, to which the solution has been given elsewhere [**14**]. Here we shall state without proof what the criteria are.

A *necessary condition* for the identifiability of a structural equation within a given linear model is that the number[8] of variables excluded from that equation (more generally: the number of linear restrictions on the parameters of that equation) be at least equal to the number (G, say) of structural equations less one. This is known as the *order condition* of identifiability. A *necessary and sufficient condition* for the identifiability of a structural equation within a linear model, restricted only by the exclusion of certain variables from certain equations, is that we can form at least one nonvanishing determinant of order $G - 1$ out of those coefficients, properly arranged, with which the variables excluded from that structural equation appear in the $G - 1$ other structural equations. This is known as the *rank condition* of identifiability.

The application of these criteria to the foregoing examples is straightforward. In all cases considered, the number of structural equations is $G = 2$. Therefore, any of the equations involved can be identifiable through exclusion of variables only if at least $G - 1 = 1$ variable is excluded from it by the model. If this is so, the equation is identifiable provided at least one of the variables so excluded occurs in the other equation with nonvanishing coefficient (a determinant of order 1 equals the value of its one and only element). For instance, the conclusion already reached at the end of the discussion of our second example is now confirmed: The identifiability of the demand equation (5d) is only then safeguarded by the exclusion of the variable r from that equation if $\delta \neq 0$, that is, if that variable not only possibly but actually occurs in the supply equation.

5. THE STATISTICAL TEST OF A PRIORI UNCERTAIN IDENTIFIABILITY

The example just quoted shows that the identifiability of one structural parameter, θ, say, may depend on the value of another structural parameter, η, say. In such situations, which are of frequent occurrence, the identifiability of θ cannot be settled by a priori reasoning from the model alone. On the other hand, the identifiability of θ cannot escape all analysis because of possible nonidentifiability of η. As is argued more fully elsewhere [**13**], since the identifiability of any parameter is a

[8] Again counting lagged variables as separate variables.

property of the distribution of the observations, it is subject to some suitable statistical test, of which the degree of conclusiveness tends to certainty as the number of observations increases indefinitely. The validity of this important conclusion is not limited to linear models.

In the case of a linear model as described in Section 4, the present statement can also be demonstrated explicitly by equivalent reformulation of the rank criterion for identifiability in terms of identifiable parameters only. By the *reduced form* of a complete set of linear structural equations as described in Section 4, we mean the form obtained by solving for each of the *dependent* (i.e., nonlagged endogenous) variables, in terms of the *predetermined* (i.e., exogenous or lagged endogenous) variables, and in terms of transformed disturbances (which are linear functions of the disturbances in the original structural equations). It has been argued more fully elsewhere [**14**, Section 3.1.6], that the coefficients of the equations of the reduced form are parameters of the joint distribution of the observations, and as such are always identifiable.

It may be stated briefly without proof that the following rank criterion for identifiability of a given structural equation, in terms of coefficients of the reduced form, is equivalent to that stated in Section 4 above: Consider only those equations of the reduced form that solve for dependent variables, specified by the model as occurring in (strictly: as not excluded from) the structural equation in question. Let the number of the equations so obtained be H, where $H \leqslant G$. Now form the matrix Π^{**} of the coefficients, in these H equations, of those predetermined variables that are excluded by the model from the structural equation involved. A necessary and sufficient condition for the identifiability of that structural equation is that the rank of Π^{**} be equal to $H - 1$. A direct proof of the equivalence of the two identification criteria will be published in due course.

6. IDENTIFICATION THROUGH DISAGGREGATION AND INTRODUCTION OF SPECIFIC EXPLANATORY VARIABLES

As a further exercise in the application of these criteria, we shall consider a question which has already been the subject of a discussion between Ezekiel [**2, 3**] and Klein [**9, 10**]. The question is whether identifiability of the investment equation can be attained by the subdivision of the investment variable into separate categories of investment. In the discussion referred to, which took place before the concepts and terminology employed in this article were developed, questions of identifiability were discussed alongside with questions regarding the merit of particular economic assumptions incorporated in the model, and with questions of the statistical method of estimating parameters that have been recognized as identifiable. In the present context, we shall avoid the latter two groups of problems and concentrate on the formal analysis of identifiability, accepting a certain model as economically valid for purposes of discussion.

As a starting point we shall consider a simple model expressing the crudest elements of Keynesian theory. The variables are, in money amounts,

$$\text{(10)} \qquad \begin{cases} S & \text{savings} \\ I & \text{investment} \\ Y & \text{income} \\ Y_{-1} & \text{income lagged one year.} \end{cases}$$

The structural equations are:

$$\text{(11)} \qquad \begin{cases} \text{(11id)} & S - I = 0 \\ \text{(11S)} & S \quad - \alpha_1 Y - \alpha_2 Y_{-1} - \alpha_0 = u \\ \text{(11I)} & I - \beta_1 Y - \beta_2 Y_{-1} - \beta_0 = v. \end{cases}$$

Of these, the first is the well-known savings-investment identity arising from Keynes's definitions of these concepts.[9] The second is a behavior equation of consumers, indicating that the money amount of their savings (income not spent for consumption) is determined by present and past income, subject to a random disturbance u. The third is a behavior equation of entrepreneurs, indicating that the money amount of investment is determined by present and past income, subject to a random disturbance v.

Since the identity (11id) is fully given a priori, no question of identifiability arises with respect to the first equation. In both the second and third equations, only one variable is excluded which appears in another equation of the model, and no other restrictions on the coefficients are stated.[10] Hence both of these equations already fail to meet the necessary order criterion of identifiability. This could be expected because the two equations connect the same savings-investment variable with the same two income variables, and therefore can not be distinguished statistically.

[9] These definitions include in investment all increases in inventory, including undesired inventories remaining in the hands of manufacturers or dealers as a result of falling demand. In principle, therefore, the "investment" equation should include a term or terms explaining such inventory changes. The absence of such terms from (11) and from later elaborations thereof may be taken as expressing the "theory" that for annual figures, say, such changes can be regarded as random. Alternatively, investment may be defined so as to exclude undesired inventory changes, and (11id) may be interpreted as an "equilibrium condition," expressing the randomness of such changes by replacing the zero in the right hand member by a disturbance w. The obvious need for refinement in this crude "theory" does not preclude its use for illustrative purposes.

[10] The normalization requirement that the variables S and I shall have coefficients $+1$ in (11S) and (11I) respectively does not restrict the relationships involved but merely serves to give a common level to coefficients which otherwise would be subject to arbitrary proportional variation.

Ezekiel attempts to obtain identifiability of the structure by a refinement of the model as a result of subdivision of aggregate investment I into the following four components:

$$(12a)\quad \begin{cases} I_1 \text{ investment in plant and equipment} \\ I_2 \text{ investment in housing} \\ I_3 \text{ temporary investment: changes in consumers' credit and in} \\ \quad \text{business inventories} \\ I_4 \text{ quasi-investment: net contributions from foreign trade and} \\ \quad \text{the government budget.} \end{cases}$$

If each of these components were to be related to the same set of explanatory variables as occurs in (11), the disaggregation would be of no help toward identification. Therefore, for each of the four types of investment decisions, Ezekiel introduces a separate explanatory equation, either explicitly or by implication in his verbal comments. In attempting to formulate these explanations in terms of a complete set of behavior equations, we shall introduce two more variables:

$$(12b)\quad \begin{cases} H \text{ semi-independent cyclical component of housing investment} \\ E \text{ exogenous component of quasi-investment.} \end{cases}$$

In addition, linear and quadratic functions of time are introduced as trend terms in some equations by Ezekiel. For purposes of the present discussion, we may as well disregard such trend terms, because they would help toward identification only if they could be excluded a priori from some of the equations while being included in others—a position advocated neither by Ezekiel nor by the present author.

With these qualifications, "Ezekiel's model" can be interpreted as follows:

$$(13)\quad \begin{cases} (13\text{id}) & S - I_1 - I_2 - I_3 - I_4 & & & = 0 \\ (13\text{S}) & S & - \alpha_1 Y - \alpha_2 Y_{-1} & & - \alpha_0 = u \\ (13\text{I}_1) & I_1 & - \beta_1 Y - \beta_2 Y_{-1} & & - \beta_0 = v_1 \\ (13\text{I}_2) & I_2 & - \gamma_1 Y - \gamma_2 Y_{-1} - H & & - \gamma_0 = v_2 \\ (13\text{I}_3) & I_3 & - \delta_1 Y + \delta_1 Y_{-1} & & - \delta_0 = v_3 \\ (13\text{I}_4) & I_4 & - \epsilon_1 Y - \epsilon_2 Y_{-1} & - E & - \epsilon_0 = v_4 \end{cases}$$

(13id) is the savings-investment identity. (13S) repeats (11S), and (13I_1) is modeled after (11I). More specific explanations are introduced for the three remaining types of investment decisions.

Housing investment decisions I_2 are explained partly on the basis of income[11] Y, partly on the basis of a "semi-independent housing cycle" H.

[11] We have added a term with Y_{-1} because the exclusion of such a term could hardly be made the basis for a claim of identifiability.

In Ezekiel's treatment H is not an independently observed variable, but a smooth long cycle fitted to I. We share Klein's objection [**9,** p. 255] to this procedure, but do not think that his proposal to substitute a linear function of time for H does justice to Ezekiel's argument. The latter definitely thinks of H as produced largely by a long-cycle mechanism peculiar to the housing market, and quotes in support of this view a study by Derksen [**1**] in which this mechanism is analyzed. Derksen constructs an equation explaining residential construction in terms of the rent level, the rate of change of income, the level of building cost in the recent past, and growth in the number of families; he further explains the rent level in terms of income, the number of families, and the stock of dwelling units (all of these subject to substantial time lags). The stock of dwelling units, in its turn, represents an accumulation of past construction diminished by depreciation or demolition. Again accepting without inquiry the economic assumptions involved in these explanations, the point to be made is that H in ($13I_2$) can be thought to represent specific observable exogenous and *past* endogenous variables.

Temporary investment I_3 is related by Ezekiel to the rate of change in income. Quasi-investment I_4 is related by him partly to income[12] (especially via government revenue, imports), partly to exogenous factors underlying exports and government expenditure where used as an instrument of policy. The variable E in ($13I_4$) is therefore similar to H in that it can be thought to represent observable exogenous or past endogenous variables.

It cannot be said that this interpretation of the variables H and E establishes the completeness of the set of equations (13) in the sense defined above. The variable H has been found to depend on the past values of certain indubitably endogenous variables (building cost, rent level) of which the present values do not occur in the equation system (13), and which therefore remain unexplained by (13). The reader is asked to accept what could be proved explicitly: that incompleteness of this kind does not invalidate the criteria of identifiability indicated.[13]

Let us then apply our criteria of identifiability to the behavior equations in (13). In each of these, the number of excluded variables is at least 5, i.e., at least the necessary number for identifiability in a model of 6 equations. In order to apply the rank criterion for the identifiability of the savings equation (13S), say, we must consider the matrix

[12] We have again added a term with Y_{-1} on grounds similar to those stated with respect to ($13I_2$).

[13] Provided, as indicated in footnote 7, there is no linear functional relationship between the exogenous and lagged endogenous variables occurring in (13).

$$
(14) \qquad \begin{array}{cccccc}
(I_1) & (I_2) & (I_3) & (I_4) & (H) & (E) \\
\end{array}
$$
$$
\begin{bmatrix}
-1 & -1 & -1 & -1 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & -1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -1
\end{bmatrix}
$$

There are several ways in which a nonvanishing determinant of order 5 can be selected from this matrix. One particular way is to take the columns labeled I_1, I_2, I_3, H, E. It follows that if the present model is valid, the savings equation is indeed identifiable.

It is easily seen that the same conclusion applies to the equations explaining investment decisions of the types I_1 and I_3. Let us now inspect the rank criterion matrix for the identifiability of (13I$_2$):

$$
(15) \qquad \begin{array}{ccccc}
(S) & (I_1) & (I_3) & (I_4) & (E)
\end{array}
$$
$$
\begin{bmatrix}
1 & -1 & -1 & -1 & 0 \\
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & -1
\end{bmatrix}
$$

Again the determinant value of this square matrix of order 5 is different from zero. Hence the housing equation is identifiable. A similar analysis leads to the same conclusion regarding the equation (13I$_4$) for quasi-investment.

It may be emphasized again that identifiability was attained not through the mere subdivision of total investment, but as a result of the introduction of specific explanatory variables applicable to some but not all components of investment.[14] Whenever such specific variables are available in sufficient number and variety of occurence, on good grounds of economic theory as defined above, the door has been opened in princi-

[14] In fact, more specific detail was introduced than the minimum necessary to produce identifiability. Starting again from (11), identifiability can already be obtained if it is possible to break off from investment I some observable exogenous component, like public works expenditure P (supposing that to be exogenous for the sake of argument). Writing $Q = I - P$ for the remainder of investment, (11) is then modified to read

$$
(11a) \qquad \begin{cases}
S - Q - P & = 0 \\
S \qquad\qquad - \alpha_1 Y - \alpha_2 Y_{-1} - \alpha_0 & = u \\
\qquad Q \qquad - \beta_1 Y - \beta_2 Y_{-1} - \beta_0 & = v,
\end{cases}
$$

of which each equation meets our criteria of identifiability. The intent of this remark is largely formal, because (11a) is not as defensible a "theory" as (13).

ple to statistical inference regarding behavior parameters—inference conditional upon the assumptions derived from "theory."

How wide the door has been opened, i.e., how much accuracy of estimation can be attained from given data, is of course a matter depending on many circumstances, and to be explored separately by the appropriate procedures of statistical inference.[15] In the present case, the extent to which the exclusion of H and/or E from certain equations contributes to the reliability of estimates of their parameters depends very much on whether or not there are pronounced differences in the time-paths of the three *predetermined variables* Y_{-1}, H, E, i.e., the variables determined either exogenously or in earlier time units. These time-paths represent in a way the basic patterns of movement in the economic model considered, such that the time-paths of all other variables are linear combinations of these three paths, modified by disturbances. If the three basic paths are sufficiently distinct, conditions are favorable for estimation of identifiable parameters. If there is considerable similarity between any two of them, or even if there is only a considerable multiple correlation between the three, conditions are adverse.

7. IMPLICATIONS OF THE CHOICE OF THE MODEL

It has already been stressed repeatedly that any statistical inference regarding identifiable parameters of economic behavior is conditional upon the validity of the model. This throws great weight on a correct choice of the model. We shall not attempt to make more than a few tentative remarks about the considerations governing this choice.[16]

It is an important question to what extent certain aspects of a model of the kind considered above are themselves subject to statistical test. For instance, in the model (13) we have specified linearity of each equation, independence of disturbances in successive time units, time lags which are an integral multiple of the chosen unit of time, as well as exclusions of specific variables from specific equations. It is often possible to subject one particular aspect or set of specifications of the model to a statistical test which is conditional upon the validity of the remaining

[15] We are not concerned here with an evaluation of the particular estimation procedures applied by Ezekiel.

[16] In an earlier article [11] I have attempted, in a somewhat different terminology, to discuss that problem. That article needs rewriting in the light of subsequent developments in econometrics. It unnecessarily clings to the view that each structural equation represents a causal process in which one single dependent variable is determined by the action upon it of all other variables in the equation. Moreover, use of the concept of identifiability will contribute to sharper formulation and treatment of the problem of the choice of a model. However, the most serious defect of the article, in my view, cannot yet be corrected. It arises from the fact that we do not yet have a satisfactory statistical theory of choice among several alternative hypotheses.

specifications. This is, for instance, the case with respect to the exclusion of any variable from any equation whenever the equation involved is identifiable even without that exclusion. However, at least *four* difficulties arise which point to the need for further fundamental research on the principles of statistical inference.

In the *first* place, on a given basis of maintained hypotheses (not subjected to test) there may be several alternative hypotheses to be tested. For instance, if there are two variables whose exclusion, either jointly or individually, from a given equation is not essential to its identifiability, it is possible to test separately (a) the exclusion of the first variable, or (b) of the second variable, or (c) of both variables simultaneously, as against (d) the exclusion of neither variable. However, instead of three separate tests, of (a) against (d), (b) against (d), and (c) against (d), we need a procedure permitting selection of one of the four alternatives (a), (b), (c), (d). An extension of current theory with regard to the testing of hypotheses, which is concerned mainly with choices between two alternatives, is therefore needed.

Secondly, if certain specifications of a model can be tested given all other specifications, it is usually possible in many different ways to choose the set of "other" specifications which is not subjected to test. It may not be possible to choose the minimum set of untested specifications in any way so that strong a priori confidence in the untested specifications exists. Even in such a case, it may nevertheless happen that for any choice of the set of untested specifications, the additional specifications that are confirmed by test also inspire some degree of a priori confidence. In such a case, the model as a whole is more firmly established than any selected minimum set of untested specifications. However, current theory of statistical inference provides no means of giving quantitative expression to such partial and indirect confirmation of anticipation by observation.

Thirdly, if the choice of the model is influenced by the same data from which the structural parameters are estimated, the estimated sampling variances of these estimated parameters do not have that direct relation to the reliability of the estimated parameters which they would have if the estimation were based on a model of which the validity is given a priori with certainty.

Finally, the research worker who constructs a model does not really believe that reality is exactly described by a "true" structure contained in the model. Linearity, discrete time lags, are obviously only approximations. At best, the model builder hopes to construct a model that contains a structure which approximates reality to a degree sufficient for the practical purposes of the investigation. The tests of current statistical theory are formulated as an (uncertain) choice, from two or more sets of structures (single or composite hypotheses), of that one which contains the "true" structure. Instead we need to choose the simplest possible set—in some sense—which contains a structure sufficiently approximative—in some sense—to economic reality.

8. FOR WHAT PURPOSES IS IDENTIFICATION NECESSARY?

The question should finally be considered why it is at all desirable to postulate a structure behind the probability distribution of the variables and thus to become involved in the sometimes difficult problems of identifiability. If we regard as the main objective of scientific inquiry to make prediction possible and its reliability ascertainable, why do we need more than a knowledge of the probability distribution of the variables to permit prediction of one variable on the basis of known (or hypothetical) simultaneous or earlier values of other variables?

The answer to this question is implicit in Haavelmo's discussion of the degree of permanence of economic laws [**6,** see p. 30] and has been formulated explicitly by Hurwicz [**8**]. Knowledge of the probability distribution is in fact sufficient whenever there is no change in the structural parameters between the period of observation from which such knowledge is derived and the period to which the prediction applies. However, in many practical situations it is required to predict the values of one or more economic variables, either under changes in structure that come about independently of the economist's advice, or under hypothetical changes in structural parameters that can be brought about through policy based in part on the prediction made. In the first case knowledge may, and in the second case it is likely to, be available as to the effect of such structural change on the parameters. An example of the first case is a well-established change in consumers' preferences. An example of the second case is a change in the average level or in the progression of income tax rates.

In such cases, the "new" distribution of the variables on the basis of which predictions are to be constructed can only be derived from the "old" distribution prevailing before the structural change, if the known structural change can be applied to identifiable structural parameters, i.e., parameters of which knowledge is implied in a knowledge of the "old" distribution combined with the a priori considerations that have entered into the model.

Cowles Commission for Research in Economics

References

[1] J. B. D. Derksen, "Long Cycles in Residential Building: An Explanation." Econometrica, Vol. 8, April 1940, pp. 97–116.

[2] M. Ezekiel, "Saving, Consumption and Investment," *American Economic Review*, Vol. 32, March 1942, pp. 22–49; June 1942, pp. 272–307.

[3] M. Ezekiel, "The Statistical Determination of the Investment Schedule," Econometrica, Vol. 12, January 1944, pp. 89–90.

[4] R. Frisch, *Pitfalls in the Statistical Construction of Demand and Supply Curves*, Veröffentlichungen der Frankfurter Gesellschaft für Konjunkturforschung, Neue Folge, Heft 5, Leipzig, 1933.

[5] R. Frisch, "Statistical versus Theoretical Relations in Economic Macrodynamics," Mimeographed document prepared for a League of Nations conference concerning Tinbergen's work, 1938.

[6] T. Haavelmo, "The Probability Approach in Econometrics," Econometrica, Vol. 12, Supplement, 1944, also Cowles Commission Paper, New Series, No. 4.

[7] L. Hurwicz, "Generalization of the Concept of Identification," in *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10, New York, John Wiley and Sons. (forthcoming).

[8] L. Hurwicz, "Prediction and Least-Squares," in *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10, New York, John Wiley and Sons. (forthcoming).

[9] L. Klein, "Pitfalls in the Statistical Determination of the Investment Schedule," Econometrica, Vol. 11, July–October 1943, pp. 246–258.

[10] L. Klein, "The Statistical Determination of the Investment Schedule: A Reply," Econometrica, Vol. 12, January 1944, pp. 91–92.

[11] T. C. Koopmans, "The Logic of Econometric Business Cycle Research," *Journal of Political Economy*, Vol. 49, 1941, pp. 157–181.

[12] T. C. Koopmans, "Measurement Without Theory," *The Review of Economic Statistics*, Vol. 29, No. 3, August 1947, pp. 161–172, also Cowles Commission Paper, New Series, No. 25.

[13] T. C. Koopmans and O. Reiersol, "Identification as a Problem in Inference," to be published.

[14] T. C. Koopmans, H. Rubin and R. B. Leipnik, "Measuring the Equation Systems of Dynamic Economics," in *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10, New York, John Wiley and Sons. (forthcoming).

[15] J. Marschak, "Economic Interdependence and Statistical Analysis," in *Studies in Mathematical Economics and Econometrics*, in memory of Henry Schultz, Chicago, The University of Chicago Press, 1942, pp. 135–150.

[16] A. C. Pigou, "A Method of Determining the Numerical Values of Elasticities of Demand," *Economic Journal*, Vol. 20, 1910, pp. 636–640, reprinted as Appendix II in *Economics of Welfare*.

[17] Henry Schultz, *Theory and Measurement of Demand*, Chicago, The University of Chicago Press, 1938.

[18] A. Wald, "Note on the Identification of Economic Relations," in *Statistical Inference in Dynamic Economic Models*, Cowles Commission Monograph 10, New York, John Wiley and Sons. (forthcoming).