6

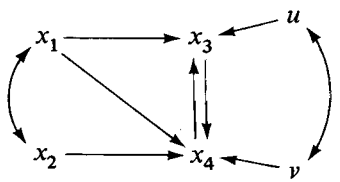
Underidentification and the Problem of Identification

Let us make one change in the model considered in Chapter 5 so that x_4 now depends on both exogenous variables. The new model is

$$x_3 = b_{31}x_1 + b_{34}x_4 + u$$

$$x_4 = b_{41}x_1 + b_{42}x_2 + b_{43}x_3 + v$$

The path diagram is



Multiplying through the x_3-equation by exogenous variables we obtain (as before)

$$\sigma_{13} = b_{31}\sigma_{11} + b_{34}\sigma_{14}$$

$$\sigma_{23} = b_{31}\sigma_{12} + b_{34}\sigma_{24}$$

Hence, we may estimate the structural coefficients of the x_3-equation by IV. (x_1 serves as its own instrument, whereas x_2 is the instrument for x_4, which cannot perform this role for itself because of its correlation with u.) The estimates are given in Set (*vi*) of Chapter 5.

Turning to the x_4-equation we note that only x_1 and x_2 are available as instrumental variables, since x_3 is correlated with v. Multiplying through by the instrumental variables we obtain:

$$\sigma_{14} = b_{41}\sigma_{11} + b_{42}\sigma_{12} + b_{43}\sigma_{13}$$

$$\sigma_{24} = b_{41}\sigma_{12} + b_{42}\sigma_{22} + b_{43}\sigma_{23}$$

We see that even if the σ's were known we could not solve uniquely for the b's, since there are three unknowns in only two equations. The x_4-equation of this model is *underidentified.* Note that the problem of identification is quite distinct from problems due to errors of sampling. We would be unable to estimate the structural coefficients in an underidentified equation even if we knew the population variances and covariances.

Another perspective on the identification problem is gained in examining the reduced form of the model. Substituting each equation into the other we obtain

$$x_3 = \frac{(b_{31} + b_{34}b_{41})x_1 + b_{34}b_{42}x_2 + u + b_{34}v}{1 - b_{34}b_{43}}$$

$$x_4 = \frac{(b_{41} + b_{43}b_{31})x_1 + b_{42}x_2 + b_{43}u + v}{1 - b_{34}b_{43}}$$

We may adopt new symbols for the reduced-form coefficients; their definitions serve to express the reduced-form coefficients in terms of the structural coefficients:

$$a_{31} = \frac{b_{31} + b_{34}b_{41}}{1 - b_{34}b_{43}}$$

$$a_{32} = \frac{b_{34}b_{42}}{1 - b_{34}b_{43}}$$

$$a_{41} = \frac{b_{41} + b_{43}b_{31}}{1 - b_{34}b_{43}}$$

$$a_{42} = \frac{b_{42}}{1 - b_{34}b_{43}}$$

The analysis of the fully identified model (see page 72) carries over to the extent that we can estimate the a's by OLS. However, even if we knew the a's, we would not be able to solve for all the b's, since there are five unknowns in the four equations defining the a's. It does turn out that we can solve the reduced-form equations for $b_{34} = a_{32}/a_{42}$ and for $b_{31} = a_{31} - b_{34}a_{41}$. This corresponds to the fact that the x_3-equation is just identified, even though the x_4-equation is underidentified.

Thus one diagnosis of underidentification arises from study of the model's reduced form: If there are not enough reduced-form coefficients to define solutions for the structural coefficients, at least one of the equations of the model is underidentified.

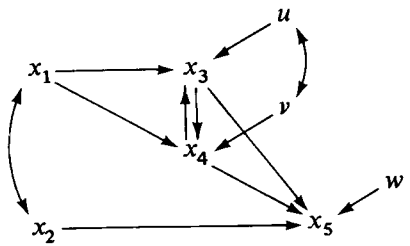
A general counting rule is perhaps easier to apply. For each equation of a model count the number (G) of explanatory variables (variables on which the dependent variable depends directly, or which have causal arrows pointing directly to it). Then count the number (H) of variables available as instrumental variables; these will include all exogenous variables in the model and any other variables that are predetermined with respect to the particular equation. (In the simple nonrecursive models considered thus far, the only predetermined variables are, in fact, the strictly exogenous ones.) A necessary condition for identification is that $H \geq G$. (This is the so-called "order condition" for identification; but we shall not explain that term here.) If $H < G$, the equation is underidentified. For the x_4-equation in our illustrative model we find $G = 3$ (counting x_3, x_2, and x_1 as explanatory variables) and $H = 2$ (counting the exogenous variables x_1 and x_2 as instrumental variables); $H < G$, so that the x_4-equation is underidentified. The counting rule is necessary, but not strictly sufficient, although it usually suffices in practice, except for the kind of pathological model noted presently.

The sufficient condition for identification (the so-called "rank condition") is that each equation of a model be distinct from every other equation in the model and from all possible linear combinations of equations in the model. [We will not try to elucidate this statement, but simply refer the sufficiently highly motivated reader to the technical econometric literature, especially Christ (1966). However, we give below an example of how the condition may be violated.] If this condition is satisfied and if $H = G$, the equation is exactly or just identified; but if $H > G$, it is overidentified.

Note how this definition of overidentification applies to the recursive model studied in Chapter 3 (pages 44–50). The x_4-equation of that model included two explanatory variables, whereas three variables in the model were predetermined with respect to x_4. There was no need there to resort to instrumental variables not in the x_4-equation (indeed, as was indicated, it would be a mistake to do so), since each of the explanatory variables was, in fact, predetermined and could serve as its own instrument.

We consider later how to proceed in the case of overidentified nonrecursive models, but note here only that both overidentified ($H > G$) and just identified ($H = G$) models are termed "identified."

Our present concern is how to recognize underidentification. We present a new example:



The equations of the model are

$$x_3 = b_{31}x_1 + b_{34}x_4 + u$$

$$x_4 = b_{41}x_1 + b_{43}x_3 + v$$

$$x_5 = b_{52}x_2 + b_{53}x_3 + b_{54}x_4 + w$$

We pause in the discussion of underidentification to observe that this model combines features of the two main kinds of models—recursive and nonrecursive. With regard to x_5 all the preceding variables are predetermined, and the specification on the disturbance of the x_5-equation is $E(x_j w) = 0, j = 1, \ldots, 4$.

Exercise. *Determine whether the x_5-equation is identified and, if it is, whether it is just identified or overidentified. If it is overidentified, determine the overidentifying restriction and suggest the appropriate methods of testing the restriction and of estimating the coefficients, assuming the overidentifying restriction is believed to obtain.*

With regard to x_3 and x_4, the model is nonrecursive, since these are jointly dependent variables. The specifications on their disturbances are $E(x_j u) = E(x_j v) = 0, j = 1, 2$, since both x_1 and x_2 are exogenous.

The model as a whole is *block-recursive.* The x_3- and x_4-equations comprise the first block; the x_5-equation by itself makes up the second block. The property of recursivity holds as between such blocks, the separability of which turns on the fact that they do not share any *endogenous* variables. (In the present example, x_3 and x_4 are endogenous with respect to the first two equations, but predetermined with respect to x_5.)

In our analysis of identification we focus on the x_3- and x_4-equations. Multiplying through by exogenous variables, we find

$$\left.\begin{aligned}\sigma_{13} &= b_{31}\sigma_{11} + b_{34}\sigma_{14}\\ \sigma_{23} &= b_{31}\sigma_{12} + b_{34}\sigma_{24}\end{aligned}\right\} \quad \text{from the } x_3\text{-equation}$$

$$\left.\begin{aligned}\sigma_{14} &= b_{41}\sigma_{11} + b_{43}\sigma_{13}\\ \sigma_{24} &= b_{41}\sigma_{12} + b_{43}\sigma_{23}\end{aligned}\right\} \quad \text{from the } x_4\text{-equation}$$

Taking the σ's as known and solving for the b's, we obtain

$$b_{31} = \frac{\sigma_{13}\sigma_{24} - \sigma_{14}\sigma_{23}}{\sigma_{11}\sigma_{24} - \sigma_{12}\sigma_{14}}$$

$$b_{34} = \frac{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_{11}\sigma_{24} - \sigma_{12}\sigma_{14}}$$

$$b_{41} = \frac{\sigma_{14}\sigma_{23} - \sigma_{13}\sigma_{24}}{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}$$

$$b_{43} = \frac{\sigma_{11}\sigma_{24} - \sigma_{12}\sigma_{14}}{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}$$

Now we observe a disconcerting feature of the solution: $b_{43} = 1/b_{34}$ whatever the values of the σ's, and similarly $b_{41} = -b_{31}/b_{34}$. So there is really only one set of coefficients that governs both of the equations. Or, more accurately, there really is only one equation, and whether we call it the x_3-equation or the x_4-equation is a matter of indifference.

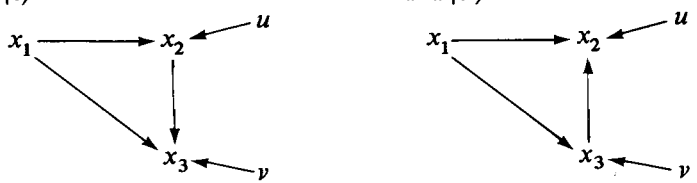
Perhaps that should have been clear at the outset, for now we see that it is possible to rearrange the x_3-equation to read

$$x_4 = -\frac{b_{31}}{b_{34}}x_1 + \frac{1}{b_{34}}x_3 - \frac{1}{b_{34}}u$$

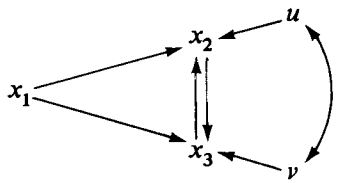
which is indistinguishable in form from the original x_4-equation. There is simply no way to tell whether we are estimating b_{41} or $-b_{31}/b_{34}$, whether we are estimating b_{43} or $1/b_{34}$.

Let us imagine a scenario—one with a basis in experience and not wholly fictitious. An investigator is working on a three-variable problem. He feels confident that x_1 precedes x_2 and x_3 in a causal ordering, but is uncertain which way the causal arrow runs between x_2 and x_3. That is, he is trying to choose between the models,

(i) *and (ii)*

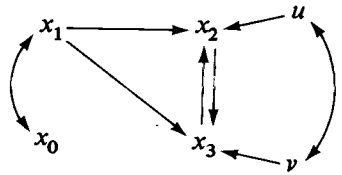


He resolves to let the question be decided by the data and specifies the nonrecursive model



Thus, he reasons, if b_{23} is large and b_{32} is small, I will conclude that the predominance of the causation is in the direction $x_3 \rightarrow x_2$, so that model (*ii*) is preferred; if the opposite is true, I will decide for model (*i*). At this point, he sees that by the counting rule, both the x_2-equation and the x_3-equation are underidentified. Hence, he introduces an instrumental variable, x_0, and considers the following model, which by

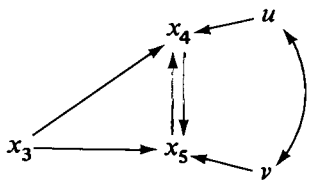
the counting rule appears to be identified:



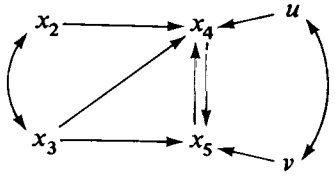
But we know already the end of this story. The hapless investigator works so hurriedly in making his computations (which go smoothly enough, offering no hint that anything is wrong) that he fails to notice the curious fact that $\hat{b}_{32} = 1/\hat{b}_{23}$, precisely. He does note, however, that $\hat{b}_{23}$ is large while $\hat{b}_{32}$ is small, and (without reporting this preliminary investigation to anyone in particular) in his further research treats the $x_2 \rightarrow x_3$ path as negligible.

Moral: Underidentification, not "causal inference," is achieved by "letting the data decide which way the causal arrow runs." The data cannot decide this matter, except in the context of a very strong theory, as is illustrated in the next exercise.

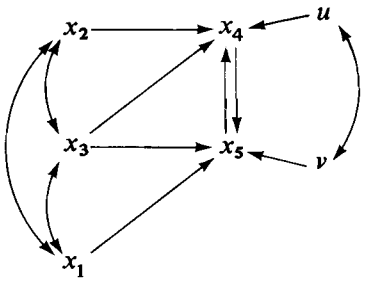
Exercise. *Imagine that the reasoning of the hypothetical investigator had been different. She began with a model in which both equations were underidentified:*



It then occurred to her to introduce an exogenous variable that appeared in the x_4-equation but not the x_5-equation:

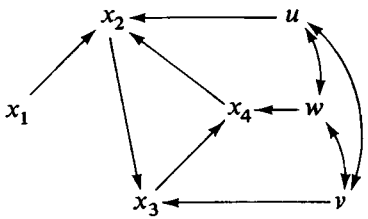


Show that this results in the x_5-equation being just identified, while the x_4-equation is still underidentified. If the investigator next introduced still another exogenous variable, which this time appears only in the x_5-equation, she would have the model



What is the status of each equation in this last model with respect to identification? What do you conclude about the kind of theory that is needed as a basis for specifying nonrecursive models in such a way that they are identified? What theories in sociology are known to you that persuasively provide such a basis?

Exercise. *Discuss the following model from the standpoint of identification:*



If one or more equations are underidentified, describe modifications of the model that would render all its equations identified.

The Aim of the Game

When the identification problem is presented in a purely formal way—as we have done here, for compactness—one's suspicions are certainly aroused that achieving identification is only a game. If your

first sketch of a model turns out to be underidentified, just put in another variable in the "right" place and see if that shortcoming is remedied. But, of course, however simple "putting in another variable" may be in mathematical terms, it is a difficult undertaking in substantive terms. Our training in what passes for sociological theory tends to inculcate the healthy instinct to presume that "everything is connected to everything else." But a model in which this is true and in which all the connections are direct is an underidentified model—sometimes called, for rhetorical purposes, a "hopelessly underidentified" model.

Moreover, it is not enough just to "put in another variable," even if that variable is in the "right place." The additional variable(s), such as x_1 and x_2 in the exercise on page 87, must really belong in the model. Such variables must "make a difference" in the endogenous variable of the equation whose identifiability is in question, even though that difference is produced solely via indirect paths. From the standpoint of statistical estimation it must be the case that the variance in the endogenous variable produced (indirectly) by the exogenous variable(s) omitted from its equation is nontrivial. Looked at in this way (although the issue is too difficult to explore with our elementary methods), degree of identifiability may vary from weak to strong, whereas our formal analysis seemingly suggests that identification is an all-or-none proposition. For this reason, Klein (1962) advises us, "Identification cannot be cheaply achieved in any particular investigation by simply adding some weak or marginal variable to one of the relationships of a system. One must add something substantial and significant which had been previously neglected [p. 18]."

The identification problem with nonrecursive models is much the same as the problem of causal ordering with recursive models. You have to be able to argue convincingly that certain logically possible direct connections between variables are, in reality, nonexistent. Your theory must provide you with a secure basis for "sectoring" the world in such a way that the causal mechanisms of Equation 1 are really different from those operating in Equation 2 while still a different set of mechanisms comes into play in Equation 3, and so on. If the endogenous variables in all these equations are really just slightly different measures of the same thing—say, an individual's attitudes on three different but closely related issues—it is going to require a very

subtle and elaborate theory indeed to produce distinct sets of determinants of those attitudes. If, by contrast, the first equation describes the behavior of labor, the second the behavior of management, and the third the behavior of government (or, respectively, the behaviors of the father, the mother, and the child), we may more easily argue that at least some of the causes involved in each equation do not appear in all the other equations.

Sociological studies involving serious efforts to construct nonrecursive models are still so few that no conclusion can be drawn as to the productivity of this approach. One can only offer conjectures, as already stated, concerning the kinds of problem that may prove amenable to study by such models. It does seem likely, however, that some modifications in our habits of theory construction—and not only in our practice of statistical analysis—will have to occur before many convincing examples of nonrecursive models are forthcoming. An investment in the study of the formal properties of such models amounts to making a wager as to the direction of development in the subject matter discipline in which one will work.

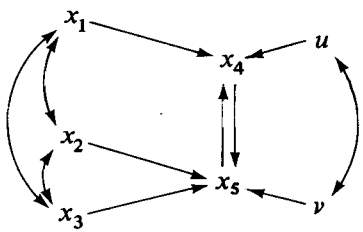
FURTHER READING

A comprehensive treatment of the identification problem is Fisher (1966); although it is heavily mathematical, a number of instructive points are formulated verbally. A discussion of the identifiability of a particular sociological model is found in Henry and Hummon (1971) with reply by Woelfel and Haller (1971). For an appreciation of Sewall Wright's long-neglected contribution to the identification problem see Goldberger (1972a). The classic paper employing modern nomenclature in expounding the identification problem was published by Koopmans in 1949; it is reprinted in Blalock (1971, Chap. 9). Criteria for "good" instrumental variables are suggested by Fisher (in Blalock, 1971, pages 260ff.).

7

Overidentification in a Nonrecursive Model

Let us enlarge the model considered in Chapter 5. We assume there are three exogenous variables, and their direct effects on the two jointly dependent variables are as shown in the path diagram:



The model, therefore, is:

$$x_4 = b_{41}x_1 + b_{45}x_5 + u$$

$$x_5 = b_{52}x_2 + b_{53}x_3 + b_{54}x_4 + v$$

with the usual specification on the disturbances. Application of the counting rule (page 83) suggests that the x_5-equation is just identified (there are three explanatory variables in that equation and three