to the causal ordering of the variables, which was required to be specified at the outset. It is, therefore, an exceedingly misleading statistic. (One does not often see the mistake in quite this crude a form. But naïvely regressing causes on effects is far from being unknown in the literature.)

Another way to raise $R^2_{3(21)}$ would be to introduce another variable, say x'_3, that is essentially an alternative measure of x_3, though giving slightly different results. The regression of x_3 on x'_3, x_2, and x_1 is then guaranteed to yield a high value of R^2.

Indeed, the best-known examples of very high correlations are those selected to convey the notion of "spurious correlation," "nonsense correlation in time series," or other kinds of artifact. This shows us that high values of R^2, in themselves, are not sufficient to evaluate a model as successful.

Before worrying too much about his R^2, therefore, the investigator does well to reconsider the entire specification of the model. If that specification cannot be faulted on other grounds, the R^2 as such is not sufficient reason to call it into question.

Exercise. *To conclude, for the time being, your study of fully recursive models, review the material on estimation and testing in Chapter 3 and restate the essential points so that they apply to the recursive model as expressed without standardization of variables* (*page* 52).

FURTHER READING

An example of a recursive sociological model presented in terms of both standardized and nonstandardized coefficients appears in Duncan (1969). Note that the more interesting conclusions were developed on the basis of the latter. On the questionable value of commonly used measures of "relative importance" or "unique contribution" of the several variables in an equation, see Ward (1969), Cain and Watts (1970), and Duncan (1970).

5

A Just-Identified Nonrecursive Model

The model considered throughout this chapter is

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

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

For convenience, $E(x_j) = 0$, $j = 1, \ldots, 4$, and $E(u) = E(v) = 0$. However, we do *not* put the variables in standard form. Variables x_1 and x_2 are *exogenous*; their variances and their covariance are not explained within the model. Variables x_3 and x_4 are *jointly dependent* or *endogenous*; the purpose of the model is to explain the behavior of these variables. Variables u and v are, respectively, the *disturbances* in the x_3-equation and the x_4-equation. Their presence accounts for the fact that x_3 and x_4 are not fully explained by their explicit determining factors. The model will be operational only if we can assume that disturbances are uncorrelated with exogenous variables: hence the specification $E(x_1 u) = E(x_1 v) = E(x_2 u) = E(x_2 v) = 0$. *This is a serious assumption.* The research worker must carefully consider what circumstances would violate it and whether his theoretical understanding of the situation under study permits him to rule out such violations.

In contrast to the case of a fully recursive model, in the nonrecursive model the specification of zero covariances between disturbances and

exogenous variables does not lead to either zero covariance between the two disturbances or zero covariance between the disturbance and each explanatory variable in an equation. To see this, let us multiply through each equation of the model by every variable in the model, and take expectations. To express the result of this operation in a convenient form, we will adopt the notation $E(x_j^2) = \sigma_{jj}$ where *sigma* with the repeated subscript refers to the (population) variance of x_j, and $E(x_h x_j) = \sigma_{hj}$ where (if $h \neq j$) *sigma* with two different subscripts refers to the (population) covariance of x_h and x_j. From the x_3-equation in the model we obtain, after multiplying through by x_1 and x_2:

$$E(x_1 x_3) = b_{31} E(x_1^2) + b_{34} E(x_1 x_4) + E(x_1 u)$$
$$E(x_2 x_3) = b_{31} E(x_1 x_2) + b_{34} E(x_2 x_4) + E(x_2 u)$$

or, since $E(x_1 u) = E(x_2 u) = 0$,

$$\sigma_{13} = b_{31}\sigma_{11} + b_{34}\sigma_{14}$$
$$\sigma_{23} = b_{31}\sigma_{12} + b_{34}\sigma_{24} \qquad \text{Set } (i)$$

However, in multiplying through the x_3-equation by endogenous variables and disturbances, not all covariances involving the disturbance drop out; we find:

$$\sigma_{33} = b_{31}\sigma_{13} + b_{34}\sigma_{34} + \sigma_{3u}$$
$$\sigma_{34} = b_{31}\sigma_{14} + b_{34}\sigma_{44} + \sigma_{4u}$$
$$\sigma_{3u} = b_{34}\sigma_{4u} + \sigma_{uu}$$
$$\sigma_{3v} = b_{34}\sigma_{4v} + \sigma_{uv} \qquad \text{Set } (ii)$$

Similarly, in multiplying through the x_4-equation by exogenous variables, we obtain:

$$\sigma_{14} = b_{42}\sigma_{12} + b_{43}\sigma_{13}$$
$$\sigma_{24} = b_{42}\sigma_{22} + b_{43}\sigma_{23} \qquad \text{Set } (iii)$$

But, in multiplying it through by endogenous variables and disturbances, we find:

$$\sigma_{34} = b_{42}\sigma_{23} + b_{43}\sigma_{33} + \sigma_{3v}$$
$$\sigma_{44} = b_{42}\sigma_{24} + b_{43}\sigma_{34} + \sigma_{4v}$$
$$\sigma_{4u} = b_{43}\sigma_{3u} + \sigma_{uu}$$
$$\sigma_{4v} = b_{43}\sigma_{3v} + \sigma_{vv} \qquad \text{Set } (iv)$$

Equations in Sets (*i*), (*ii*), (*iii*), and (*iv*) are population moment equations, analogous to but different in significant ways from those pertaining to recursive models. Note the lack of symmetry in the pattern of σ's on the right-hand side in Sets (*i*) and (*iii*), in contrast with the symmetric pattern of the normal equations on page 53. Note also—or carry out an *Exercise* to show this—that we cannot replace σ_{3u} by σ_{uu} or σ_{4v} by σ_{vv}, so that the simplification mentioned on page 53 for the recursive model is not available here. The population moment equations serve to express the relationships holding among the structural coefficients of the model (the b's) and the variances and covariances in the population under study. A number of important properties of the model are disclosed in studying these sets of equations.

Note, first, that if the b's in Set (*i*) are regarded as unknown and the σ's as known, it is possible to solve these two equations for the b's:

$$b_{31} = \frac{\sigma_{13}\sigma_{24} - \sigma_{14}\sigma_{23}}{\sigma_{11}\sigma_{24} - \sigma_{12}\sigma_{14}}$$
$$b_{34} = \frac{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_{11}\sigma_{24} - \sigma_{12}\sigma_{14}} \qquad \text{Set } (v)$$

In practice, of course, we would not know the population variances and covariances. However, if we replace the σ's by the corresponding sample moments, $m_{jj} = \sum (x_j^2)$ and $m_{hj} = \sum (x_h x_j)$, where the summation is over all sample observations on variables x_h and x_j, we obtain

$$\hat{b}_{31} = \frac{m_{13}m_{24} - m_{14}m_{23}}{m_{11}m_{24} - m_{12}m_{14}}$$
$$\hat{b}_{34} = \frac{m_{11}m_{23} - m_{12}m_{13}}{m_{11}m_{24} - m_{12}m_{14}} \qquad \text{Set } (vi)$$

This method of obtaining the estimates, $\hat{b}_{31}$ and $\hat{b}_{34}$, of the structural coefficients is termed instrumental variables. Here it is equivalent to the method of indirect least squares, for reasons that will become clear later. The method works here because the number of equations in Set (*i*), obtained by multiplying through the x_3-equation of the model by all exogenous variables, is just the same as the number of unknown structural coefficients. This fact is implied when we describe the model as "just identified" or "exactly identified" with respect to the x_3-equation. If Set (*i*) included more equations than structural coefficients, we would describe the x_3-equation as "overidentified"; if there were fewer moment equations in Set (*i*) than structural coefficients, the x_3-equation would be "underidentified" or "unidentified." In the case of overidentification, we replace instrumental variables (IV) or indirect least squares (ILS) by special methods of estimation. In the case of underidentification, estimation of structural coefficients is not possible.

Turning to the moment equations in Set (*iii*), we find that the x_4-equation of the model is likewise exactly identified. The solution for the b's is

$$b_{42} = \frac{\sigma_{14}\sigma_{23} - \sigma_{13}\sigma_{24}}{\sigma_{12}\sigma_{23} - \sigma_{13}\sigma_{22}}$$

$$b_{43} = \frac{\sigma_{12}\sigma_{24} - \sigma_{14}\sigma_{22}}{\sigma_{12}\sigma_{23} - \sigma_{13}\sigma_{22}} \qquad \text{Set } (vii)$$

IV estimates, $\hat{b}_{42}$ and $\hat{b}_{43}$, may be obtained as before by replacing the σ's with sample moments, m_{jj} and m_{hj}:

$$\hat{b}_{42} = \frac{m_{14}m_{23} - m_{13}m_{24}}{m_{12}m_{23} - m_{13}m_{22}}$$

$$\hat{b}_{43} = \frac{m_{12}m_{24} - m_{14}m_{22}}{m_{12}m_{23} - m_{13}m_{22}} \qquad \text{Set } (viii)$$

Further study of this model is facilitated by solving for its *reduced form*, in which each endogenous variable is represented as a function of exogenous variables and disturbances only. Each equation is substituted into the other one. That is, for x_4 in the x_3-equation we substi-

tute the right-hand side of the x_4-equation; and for x_3 in the x_4-equation we substitute the right-hand side of the x_3-equation. After collecting terms, this algebra yields the two reduced-form equations:

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

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

Let us adopt the notation,

$$a_{31} = \frac{b_{31}}{(1 - b_{34}b_{43})}$$

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

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

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

If we now multiply through each reduced-form equation by the exogenous variables, we obtain

$$\sigma_{13} = a_{31}\sigma_{11} + a_{32}\sigma_{12}$$

$$\sigma_{23} = a_{31}\sigma_{12} + a_{32}\sigma_{22} \qquad \text{Set } (ix)$$

and

$$\sigma_{14} = a_{41}\sigma_{11} + a_{42}\sigma_{12}$$

$$\sigma_{24} = a_{41}\sigma_{12} + a_{42}\sigma_{22} \qquad \text{Set } (x)$$

since terms like $b_{43}E(x_1 u)$ and $E(x_2 v)$ drop out.

It appears that the reduced-form parameters (the a's) are exact nonlinear functions of the structural coefficients (the b's), and vice versa. Indeed, the four expressions defining the a's could just as well be regarded as four equations in the unknown b's, so that if the a's were

known we could solve for the b's by the following routine:

$$b_{34} = \frac{a_{32}}{a_{42}}$$

$$b_{43} = \frac{a_{41}}{a_{31}} \qquad \text{Set } (xi)$$

whence

$$1 - b_{34}b_{43} = 1 - \frac{a_{32}a_{41}}{a_{31}a_{42}}$$

and

$$b_{31} = a_{31}\left(1 - \frac{a_{32}a_{41}}{a_{31}a_{42}}\right)$$

$$b_{42} = a_{42}\left(1 - \frac{a_{32}a_{41}}{a_{31}a_{42}}\right) \qquad \text{Set } (xii)$$

Of course, the a's are not known, since they are functions of population variances and covariances, as shown by Sets (ix) and (x). But we could obtain estimates of the reduced-form parameters by replacing the σ's in Sets (ix) and (x) by corresponding sample moments; thus:

$$\hat{a}_{31} = \frac{m_{13}m_{22} - m_{12}m_{23}}{m_{11}m_{22} - m_{12}^2}$$

$$\hat{a}_{32} = \frac{m_{11}m_{23} - m_{12}m_{13}}{m_{11}m_{22} - m_{12}^2}$$

and

$$\hat{a}_{41} = \frac{m_{14}m_{22} - m_{12}m_{24}}{m_{11}m_{22} - m_{12}^2}$$

$$\hat{a}_{42} = \frac{m_{11}m_{24} - m_{12}m_{14}}{m_{11}m_{22} - m_{12}^2}$$

We see that these are precisely the same as the estimates we would obtain for the coefficients of the ordinary least squares (OLS) regressions of (respectively) x_3 on x_2 and x_1, and x_4 on x_2 and x_1, that is, of each endogenous variable on all exogenous variables.

Suppose we now replace the a's in Sets (xi) and (xii) by the corresponding OLS estimates, the $\hat{a}$'s. We will obtain estimates of the b's that are the very IV estimates presented earlier as Sets (vi) and $(viii)$; that is,

$$\hat{b}_{34} = \frac{\hat{a}_{32}}{\hat{a}_{42}}$$

$$\hat{b}_{43} = \frac{\hat{a}_{41}}{\hat{a}_{31}}$$

$$\hat{b}_{31} = \hat{a}_{31}\left(1 - \frac{\hat{a}_{32}\hat{a}_{41}}{\hat{a}_{31}\hat{a}_{42}}\right)$$

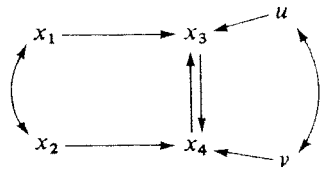
$$\hat{b}_{42} = \hat{a}_{42}\left(1 - \frac{\hat{a}_{32}\hat{a}_{41}}{\hat{a}_{31}\hat{a}_{42}}\right)$$

are the IV estimates of the b's. This fact may not be immediately obvious, but it is easily proved by algebraic substitutions.

Exercise. *Carry out this algebra.*

The fact that the $\hat{b}$'s are obtained from the $\hat{a}$'s, which in turn are OLS estimates of reduced-form regression coefficients, justifies the name indirect least squares.

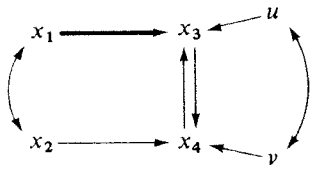
In addition to its possible use for purposes of estimation, the reduced form of the model is instructive in the way it displays the mechanisms through which the exogenous variables influence the endogenous variables. In this connection, study of the path diagram of the model is also instructive.



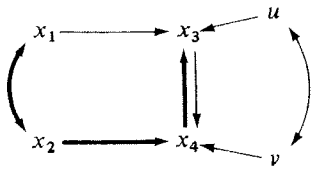
The covariance of x_1 and x_3 is obtained by multiplying through the reduced-form x_3-equation by x_1:

$$\sigma_{13} = \frac{b_{31}}{1 - b_{34}b_{43}}\sigma_{11} + \frac{b_{34}b_{42}}{1 - b_{34}b_{43}}\sigma_{12}$$

Note that there is a direct effect of x_1 on x_3:



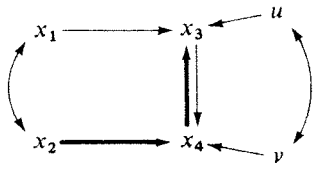
But another part of σ_{13} arises from the correlation (covariance) of x_1 with another cause (namely, x_2) of x_3, even though the latter works only indirectly:



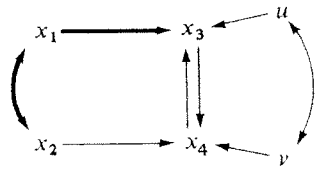
Note that we are reading the path diagram according to Sewall Wright's general principle, but are not using it as an algorithm for computing covariances. In particular, when actually calculating σ_{13} we must inflate both the direct path and the component due to a correlated cause by the factor $1/(1 - b_{34}b_{43})$. This is the "multiplier effect" in the model due to the "simultaneity" or "reciprocal causation" of the two endogenous variables. The covariance of x_2 and x_3 is

$$\sigma_{23} = \frac{b_{31}}{1 - b_{34}b_{43}}\sigma_{12} + \frac{b_{34}b_{42}}{1 - b_{34}b_{43}}\sigma_{22}$$

We see that it is produced by an indirect effect of x_2 on x_3:



and by a component due to the correlation of x_2 with x_1:



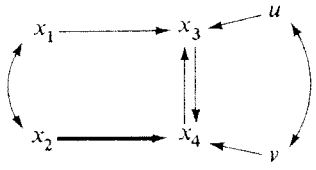
Again the multiplier effect comes into play for both components.

The same kinds of components can be found for the covariances of the two exogenous variables with x_4, shown below as they are obtained from the reduced-form x_4-equation:

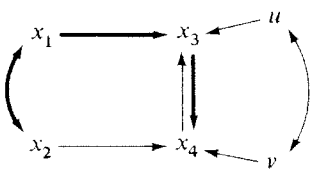
$$\sigma_{24} = \frac{b_{42}}{1 - b_{34}b_{43}}\sigma_{22} + \frac{b_{43}b_{31}}{1 - b_{34}b_{43}}\sigma_{12}$$

$$\sigma_{14} = \frac{b_{42}}{1 - b_{34}b_{43}}\sigma_{12} + \frac{b_{43}b_{31}}{1 - b_{34}b_{43}}\sigma_{11}$$

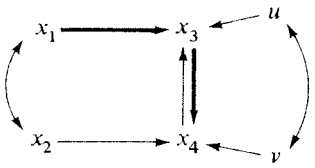
Thus, σ_{24} arises from the direct effect:



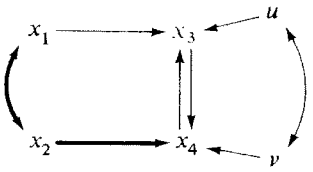
and the contribution due to a correlated cause, operating indirectly:



while σ_{14} arises from an indirect effect



and the contribution of a correlated cause:



Again, all these components are seen to include the multiplier effect of the factor $1/(1 - b_{34}b_{43})$.

It is important to understand precisely why special methods, instead of the conventional statistical procedure of OLS, are required for estimating structural coefficients in nonrecursive models. In the x_3-equation (to take it as a typical example), the two explanatory variables on the right-hand side are x_1 and x_4. If one were to estimate the coefficients of the x_3-equation by OLS, therefore, he would obtain as the estimate of b_{31}

$$\frac{m_{13}m_{44} - m_{14}m_{34}}{m_{11}m_{44} - m_{14}^2}$$

and as the estimate of b_{34}

$$\frac{m_{11}m_{34} - m_{13}m_{14}}{m_{11}m_{44} - m_{14}^2}$$

As we have seen, in Sets (*i*) and (*ii*), the covariances of x_3 with these explanatory variables are

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

$$\sigma_{34} = b_{31}\sigma_{14} + b_{34}\sigma_{44} + \sigma_{4u}$$

We may solve for the b's in terms of the σ's:

$$b_{31} = \frac{\sigma_{13}\sigma_{44} - \sigma_{14}\sigma_{34} + \sigma_{14}\sigma_{4u}}{\sigma_{11}\sigma_{44} - \sigma_{14}^2}$$

$$b_{34} = \frac{\sigma_{11}\sigma_{34} - \sigma_{13}\sigma_{14} - \sigma_{11}\sigma_{4u}}{\sigma_{11}\sigma_{44} - \sigma_{14}^2}$$

The implication is clear if we note that, apart from $\sigma_{14}\sigma_{4u}$ in the numerator, the expression for b_{31} is the population counterpart of the OLS estimator of b_{31}; and apart from $-\sigma_{11}\sigma_{4u}$ in the numerator, the expression for b_{34} is the population counterpart of the OLS estimator of b_{34}. Thus, even if our sample were infinitely large, so that we could form OLS estimators from population variances and covariances (instead of sample moments), the OLS estimates would be biased. Indeed, the OLS procedure would not estimate b_{31} but rather

$$b_{31} - \frac{\sigma_{14}\sigma_{4u}}{\sigma_{11}\sigma_{44} - \sigma_{14}^2}$$

and it would not estimate b_{34} but rather

$$b_{34} + \frac{\sigma_{11}\sigma_{4u}}{\sigma_{11}\sigma_{44} - \sigma_{14}^2}$$

Similarly we can show that OLS applied to the x_4-equation would estimate not b_{42} but rather

$$b_{42} - \frac{\sigma_{23}\sigma_{3v}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2}$$

and, instead of b_{43}, it would estimate

$$b_{43} + \frac{\sigma_{22}\sigma_{3v}}{\sigma_{22}\sigma_{33} - \sigma_{23}^2}$$

The basic reason for the failure of OLS, then, is that not all the explanatory variables in the equation are uncorrelated with the disturbance. And this is inescapably so given the jointly dependent (simultaneous, reciprocally influencing) relationships of the endogenous variables of a nonrecursive model. For if $x_3 \rightarrow x_4$, then $x_u \rightarrow x_3$ implies that $x_u \rightarrow x_3 \rightarrow x_4$ will contribute a nonzero component to σ_{4u}. Similarly,

$x_v \rightarrow x_4 \rightarrow x_3$ will contribute a nonzero component to σ_{3v}. It could happen that one or the other of these is cancelled out by σ_{uv}, for in Sets (*ii*) and (*iv*) we find

$$\sigma_{3v} = b_{34}\sigma_{4v} + \sigma_{uv}$$

and

$$\sigma_{4u} = b_{43}\sigma_{3u} + \sigma_{uv}$$

Hence, if $\sigma_{uv} = -b_{34}\sigma_{4v}$, then $\sigma_{3v} = 0$; or if $\sigma_{uv} = -b_{43}\sigma_{3u}$, then $\sigma_{4u} = 0$. But for either of these to hold would be merely a coincidence, and for both to hold would be a rare coincidence indeed.

Exercise. *Working from Sets (i), (ii), (iii), and (iv) in the spirit of Chapter 4, pages 53 - 55, show that the variances and covariances of the observable variables in the model studied in this chapter may be expressed as functions of (1) the variances and covariances of the exogenous variables, (2) a (nonlinear) combination of structural coefficients, and (3) variances and covariances of the disturbances. Verify Table 5.1.*

Table 5.1 Sources of Observable Variances and Covariances

	Variance or covariance	*Is a function of*									
		σ_{11}	σ_{12}	σ_{22}	b_{31}	b_{42}	b_{34}	b_{43}	σ_{uu}	σ_{uv}	σ_{vv}
*	σ_{11}	×	···	···	···	···	···	···	···	···	···
*	σ_{12}	···	×	···	···	···	···	···	···	···	···
*	σ_{22}	···	···	×	···	···	···	···	···	···	···
	σ_{13}	×	×	···	×	×	×	×	···	···	···
	σ_{14}	×	×	···	×	×	×	×	···	···	···
	σ_{23}	···	×	×	×	×	×	×	···	···	···
	σ_{24}	···	×	×	×	×	×	×	···	···	···
	σ_{33}	×	×	×	×	×	×	×	×	×	×
	σ_{34}	×	×	×	×	×	×	×	×	×	×
	σ_{44}	×	×	×	×	×	×	×	×	×	×

* Exogenous

Discuss implications of the possibility that some, if not all, of the parameters across the top of the table are invariant across populations.

FURTHER READING

Chapters 5, 6, and 7 are essentially the elementary portions of the standard econometric presentation of simultaneous-equation models. Several advanced econometrics texts are listed among the references at the end of this book. More accessible presentations are available in Wonnacott and Wonnacott (1970, Part I), and Wallis (1973). The latter does, however, make use of matrices.

6

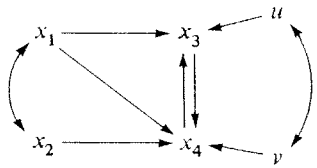
Underidentification and the Problem of Identification

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.

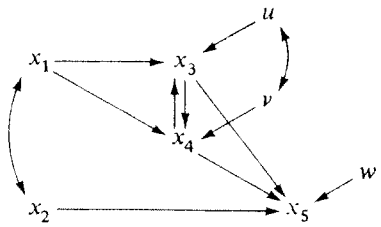
A general counting rule is perhaps easier to apply. For each equation of a model count the number (G) of explanatory variables (variables on which the dependent variable depends directly, or which have causal arrows pointing directly to it). Then count the number (H) of variables available as instrumental 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.

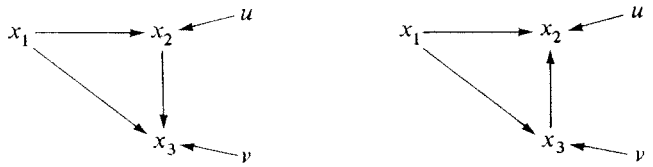
Perhaps that should have been clear at the outset, for now we see that it is possible to rearrange the x_3-equation to read

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

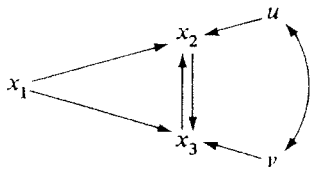
which is indistinguishable in form from the original x_4-equation. There is simply no way to tell whether we are estimating b_{41} or $-b_{31}/b_{34}$, whether we are estimating b_{43} or $1/b_{34}$.

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)*

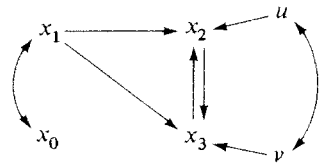


He resolves to let the question be decided by the data and specifies the nonrecursive model



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

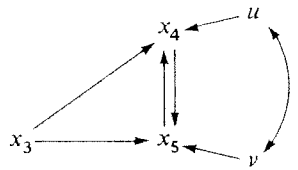
the counting rule appears to be identified:



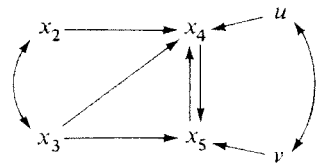
But we know already the end of this story. The hapless investigator works so hurriedly in making his computations (which go smoothly enough, offering no hint that anything is wrong) that he fails to notice the curious fact that $\hat{b}_{32} = 1/\hat{b}_{23}$, precisely. He does note, however, that $\hat{b}_{23}$ 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.

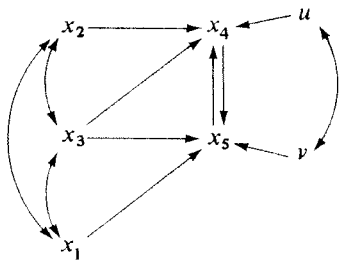
Exercise. *Imagine that the reasoning of the hypothetical investigator had been different. She began with a model in which both equations were underidentified:*



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

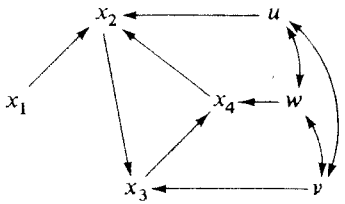


Show that this results in the x_5-equation being just identified, while the x_4-equation is still underidentified. If the investigator next introduced still another exogenous variable, which this time appears only in the x_5-equation, she would have the model



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.

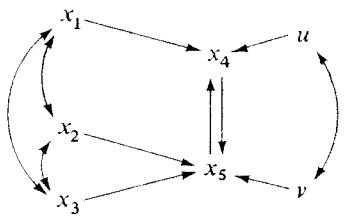
FURTHER READING

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

exogenous variables in the model as a whole). The x_4-equation is overidentified (there are only two explanatory variables in this equation).

Multiplying through by exogenous variables, we obtain

$$\left.\begin{aligned}\sigma_{14} &= b_{41}\sigma_{11} + b_{45}\sigma_{15}\\ \sigma_{24} &= b_{41}\sigma_{12} + b_{45}\sigma_{25}\\ \sigma_{34} &= b_{41}\sigma_{13} + b_{45}\sigma_{35}\end{aligned}\right\} \text{ from the } x_4\text{-equation}$$

$$\left.\begin{aligned}\sigma_{15} &= b_{52}\sigma_{12} + b_{53}\sigma_{13} + b_{54}\sigma_{14}\\ \sigma_{25} &= b_{52}\sigma_{22} + b_{53}\sigma_{23} + b_{54}\sigma_{24}\\ \sigma_{35} &= b_{52}\sigma_{23} + b_{53}\sigma_{33} + b_{54}\sigma_{34}\end{aligned}\right\} \text{ from the } x_5\text{-equation}$$

We see that the IV method is available for estimating coefficients in the x_5-equation. The estimates are obtained by solving the following set of normal equations for the $\hat{b}$'s:

$$\begin{aligned}m_{15} &= \hat{b}_{52}m_{12} + \hat{b}_{53}m_{13} + \hat{b}_{54}m_{14}\\ m_{25} &= \hat{b}_{52}m_{22} + \hat{b}_{53}m_{23} + \hat{b}_{54}m_{24}\\ m_{35} &= \hat{b}_{52}m_{23} + \hat{b}_{53}m_{33} + \hat{b}_{54}m_{34}\end{aligned}$$

The situation is not so straightforward for the overidentified x_4-equation. The overidentifying restriction implies that

$$b_{41} = \underset{(i)}{\frac{\sigma_{14}\sigma_{25} - \sigma_{24}\sigma_{15}}{\sigma_{11}\sigma_{25} - \sigma_{12}\sigma_{15}}} = \underset{(ii)}{\frac{\sigma_{14}\sigma_{35} - \sigma_{34}\sigma_{15}}{\sigma_{11}\sigma_{35} - \sigma_{13}\sigma_{15}}} = \underset{(iii)}{\frac{\sigma_{24}\sigma_{35} - \sigma_{34}\sigma_{25}}{\sigma_{12}\sigma_{35} - \sigma_{13}\sigma_{25}}}$$

and

$$b_{45} = \frac{\sigma_{11}\sigma_{24} - \sigma_{14}\sigma_{12}}{\sigma_{11}\sigma_{25} - \sigma_{12}\sigma_{15}} = \frac{\sigma_{11}\sigma_{34} - \sigma_{13}\sigma_{14}}{\sigma_{11}\sigma_{35} - \sigma_{13}\sigma_{15}} = \frac{\sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24}}{\sigma_{12}\sigma_{35} - \sigma_{13}\sigma_{25}}$$

(Although there are several equalities here, they are redundant. There is actually only one overidentifying restriction.) We might estimate these b's by replacing the σ's with sample moments in any one of these solutions. Note that neither solution (*i*), (*ii*), nor (*iii*) leads to an OLS estimate, in contrast to the result for the overidentified equation in a recursive model (page 46). (We already know, in any event, that OLS does not yield unbiased estimates in nonrecursive models, however large the sample may be.) If we replace the σ's by sample moments in the foregoing solutions, we will, in general, obtain three different pairs

of values for the estimated b's. Because of sampling error the equalities among the solutions will be only approximate, not exact, even if the model—or, in particular, its overidentifying restriction—is true. The essence of the overidentified case, then, is that there are "too many" distinct estimates of the structural coefficients. It is not obvious how to choose the best one from among them, or how to reconcile them. It might seem plausible to average the estimates. In a sense, this is what is done by the method that will be described later on. But the appropriate average is not a simple, unweighted mean of the three estimates.

Since a direct application of the IV method does not work for an overidentified equation (there are "too many" instrumental variables and no firm basis for choosing among them), we look for help in another direction, by studying the reduced form of the model. We find

$$\begin{aligned}x_4 &= a_{41}x_1 + a_{42}x_2 + a_{43}x_3 + u'\\ x_5 &= a_{51}x_1 + a_{52}x_2 + a_{53}x_3 + v'\end{aligned}$$

where the reduced-form coefficients and disturbances are the following functions of the structural-form coefficients and disturbances:

$$\begin{aligned}a_{41} &= \frac{b_{41}}{1 - b_{45}b_{54}}\\ a_{42} &= \frac{b_{45}b_{52}}{1 - b_{45}b_{54}}\\ a_{43} &= \frac{b_{45}b_{53}}{1 - b_{45}b_{54}}\\ a_{51} &= \frac{b_{54}b_{41}}{1 - b_{45}b_{54}}\\ a_{52} &= \frac{b_{52}}{1 - b_{45}b_{54}}\\ a_{53} &= \frac{b_{53}}{1 - b_{45}b_{54}}\\ u' &= \frac{u + b_{45}v}{1 - b_{45}b_{54}}\\ v' &= \frac{b_{54}u + v}{1 - b_{45}b_{54}}\end{aligned}$$

Exercise. *Using techniques similar to those in Chapters 4 and 5, verify these expressions. Find the variances of u' and v' and their covariance in terms of structural coefficients and variances and covariances of the structural-form disturbances. Remember that, in the nonrecursive case, $\sigma_{uv} = 0$ does not follow from the usual specification on the structural disturbances.*

It is apparent that our work in deriving the reduced-form coefficients has not solved our problem immediately. We find that $a_{42}/a_{52} = b_{45}$ but also $a_{43}/a_{53} = b_{45}$. If the model is true, both equalities must hold, so that

$$\frac{a_{42}}{a_{52}} = \frac{a_{43}}{a_{53}} \quad \text{or} \quad a_{42}a_{53} = a_{43}a_{52}$$

(This is another way of expressing the overidentifying restriction.) But, in practice, we do not know the a's and can only hope to secure estimates of them. Suppose we adopted as our estimates of the a's the OLS regression coefficients of x_4 on x_1, x_2, and x_3, and x_5 on x_1, x_2, and x_3, taking advantage of the fact (which the reader should verify) that covariances of u' and v' with the three exogenous variables are all zero. There is nothing about the OLS method which guarantees that estimates of coefficients in two different equations, estimated independently, will satisfy exactly the proportionality just cited. The best we could hope for is that

$$\frac{\hat{a}_{42}}{\hat{a}_{52}} \cong \frac{\hat{a}_{43}}{\hat{a}_{53}}$$

(where $\cong$ means "approximately equal to"). But this would leave us in the position of having to choose between the two distinct estimates

$$\hat{b}_{45}^{(1)} = \frac{\hat{a}_{42}}{\hat{a}_{52}}$$

and

$$\hat{b}_{45}^{(2)} = \frac{\hat{a}_{43}}{\hat{a}_{53}}$$

or otherwise reconciling the two. But there is no obvious way to do this, unless the reader, whose patience has by now worn thin, considers

it "obvious" that it is good enough to "split the difference." We would still have to harass the distraught fellow, however, by insisting that there is more than one way to "split the difference"—for example, by reconciling different estimates obtained via the reduced form or by reconciling those obtained on the IV approach.

We can, however, put our OLS estimates of reduced form coefficients to good use for the purpose at hand. They will serve as "first stage regression" coefficients for the method known as two-stage least squares (2SLS).

We are working on the overidentified x_4-equation, and it is the presence of x_5 in that equation that occasions much of our difficulty. To finesse that source of difficulty, we proceed as follows. Define $\hat{x}_5$ as

$$\hat{x}_5 = \hat{a}_{51}x_1 + \hat{a}_{52}x_2 + \hat{a}_{53}x_3$$

where the $\hat{a}$'s are the OLS estimates of coefficients in the reduced-form x_5 equation. We have then

$$x_5 = \hat{x}_5 + \hat{v}'$$

where $\hat{v}'$ is an estimate, calculated as the sample residual from the estimated regression equation, of the reduced-form disturbance v'. We substitute this expression for x_5 into the x_4-equation of the model, obtaining

$$x_4 = b_{41}x_1 + b_{45}\hat{x}_5 + b_{45}\hat{v}' + u$$

Let us multiply through by the two explanatory variables and take expectations:

$$E(x_1x_4) = b_{41}E(x_1^2) + b_{45}E(x_1\hat{x}_5) + b_{45}E(x_1\hat{v}') + E(x_1u)$$

$$E(\hat{x}_5x_4) = b_{41}E(\hat{x}_5x_1) + b_{45}E(\hat{x}_5^2) + b_{45}E(\hat{x}_5\hat{v}') + E(\hat{x}_5u)$$

We must look closely at the four terms involving disturbances. First, $E(x_1u) = 0$ by the original specification of the model. Second, $E(x_1\hat{v}')$ must be zero, since $\hat{v}'$ is the sample residual from a regression in which x_1 is one of the independent variables. It is a property of OLS that each regressor has a covariance of zero, identically, with the sample residual. Third, for much the same reason $E(\hat{x}_5\hat{v}') = 0$; for $\hat{x}_5$ is the value of the dependent variable calculated from a regression equation, and $\hat{v}'$ is the residual from that same regression. The OLS method ensures that the covariance of the two is identically zero.

The situation is messier with respect to the last term, $E(\hat{x}_5 u)$. We recall that, by definition,

$$\hat{x}_5 = \hat{a}_{51} x_1 + \hat{a}_{52} x_2 + \hat{a}_{53} x_3$$

Suppose, for the moment, that we knew the actual values of the a's in the population and did not have to use the $\hat{a}$'s. Then we could compute a slightly different quantity,

$$x_5^* = a_{51} x_1 + a_{52} x_2 + a_{53} x_3$$

If we then considered $E(x_5^* u)$ we would find that its value is

$$a_{51} E(x_1 u) + a_{52} E(x_2 u) + a_{53} E(x_3 u) = 0$$

The a's can be written to the left of the expectation sign since they are constants. (The same is not true of the $\hat{a}$'s; they are random variables that vary from one sample to another.) And, of course, each of the expectations, $E(x_h u) = 0$, $h = 1, 2, 3$, since x_1, x_2, and x_3 are exogenous variables.

All this is very nice, but $\hat{x}_5$ is not the same as x_5^*: it is only an estimate of x_5^*. Here, we must appeal to some statistical theory that lies beyond the scope of this exposition. While we may only write $\hat{x}_5 \cong x_5^*$, the error in the approximation will diminish, on the average, as we take larger and larger samples. In the limit, as the sample gets indefinitely large, the probability that $\hat{x}_5$ differs from x_5^* by more than any prespecified amount tends to zero. Replacing x_5^* by $\hat{x}_5$ in the expectation $E(x_5^* u)$, therefore, we may write

$$E(\hat{x}_5 u) \cong 0$$

and the error in the approximation gets smaller, on the average, the larger the sample. Hence, $E(\hat{x}_5 u)$ is "asymptotically equal" to zero. The approximation involved in taking it to be identically zero is of the same kind that we use whenever we employ "large-sample statistics" or "asymptotic estimators."

We have presented a long argument to the effect that, when working with a "large" sample, we are justified in dropping the last two terms in the expressions for $E(x_1 x_4)$ and $E(\hat{x}_5 x_4)$ previously given. We find, therefore, that

$$E(x_1 x_4) = b_{41} E(x_1^2) + b_{45} E(x_1 \hat{x}_5)$$

$$E(\hat{x}_5 x_4) \cong b_{41} E(\hat{x}_5 x_1) + b_{45} E(\hat{x}_5^2)$$

We see that if the " $\cong$ " is replaced by " $=$," and the several expectations by the corresponding sample moments, we will produce OLS estimates, $\hat{b}_{41}$ and $\hat{b}_{45}$, by simply regressing x_4 on x_1 and $\hat{x}_5$, where $\hat{x}_5$ is calculated from the result of the first-stage regression and the two $\hat{b}$'s are estimated in this second-stage regression. What we do, in effect, is replace x_5 in the x_4-equation by the estimate of x_5 given by its regression on *all* the exogenous variables in the model, and then use OLS on this revised equation.

We might equally well estimate the x_5-equation by 2SLS. In that event, we would calculate the first-stage OLS regression of x_4 on x_1, x_2, and x_3; compute the calculated value ($\hat{x}_4$) of x_4 from that regression; replace x_4 in the x_5-equation by $\hat{x}_4$; and estimate the coefficients in the x_5-equation by OLS regression of x_5 on x_2, x_3, and $\hat{x}_4$. It turns out that the 2SLS estimates obtained for the just-identified x_5-equation are the same as the IV estimates. The fact that 2SLS and IV give the same result in the case of any equation that is just identified may be seen as a heuristic justification for the approximation used in deriving the 2SLS method.

Exercise. *What if the x_5-equation were underidentified; specifically, what if it were specified as*

$$x_5 = b_{51} x_1 + b_{52} x_2 + b_{53} x_3 + b_{54} x_4 + v$$

could we then estimate the coefficients by 2SLS, *where the second-stage regression is x_5 on x_1, x_2, x_3, and $\hat{x}_4$?*

Answer: No, because $\hat{x}_4$ is a weighted sum (with $\hat{a}$'s as weights) of x_1, x_2, and x_3, while all three of these variables appear elsewhere in the second-stage regression equation. Our efforts to calculate OLS estimates would fail because of "singularity." If this form of pathology is not known to you, ask your teacher to explain it.

Conclusion: Neither 2SLS nor any other method of estimation is a cure for underidentification, because the identification problem does not arise from sampling errors but from difficulties of a logical kind.

We have tried only to sketch the logic of 2SLS and not to describe efficient computational procedures. We do not, moreover, deal with tests of hypotheses about the individual structural coefficients. The

standard errors required for such tests are produced by any good computer program for calculating the estimated 2SLS coefficients themselves. The intent of our discussion was to show the reader that some special method of estimation is both required and feasible whenever a model contains one or more overidentified equations.

There are several other statistically efficient methods besides 2SLS for estimating overidentified equations or, in the case of some methods, for estimating all equations in the model at once. All of these methods are more complex, conceptually and computationally, than 2SLS. Therefore, the well-motivated reader must be referred to the advanced textbooks of econometrics, of which there are several excellent ones (Johnston, Goldberger, Christ, Malinvaud, Theil, Kmenta, among others). For all the interest in these methods, most empirical work uses 2SLS, which is quite flexible in applications and which appears to have a certain robustness in the face of the practical difficulties that always arise in a serious piece of empirical work.

From estimation, we turn to some cursory remarks on the problem of testing overidentifying restriction(s). Formal procedures of statistical inference have been proposed in connection with this problem. But these procedures are little used in practice, and some questions remain about their purely statistical properties.

It is easy to see one reason why the outcome of any such test may not be highly instructive. First, suppose the model passes the test; one is not required to reject the null hypothesis which asserts the overidentifying restriction(s) to be true. But this outcome, of course, does not guarantee that the model is true; it only provides some reassurance to the investigator who thinks he has other, adequate reasons for believing it to be true.

Second, suppose the null hypothesis must be rejected. One then concludes, with only a small probability of being mistaken, that "something" about the overidentifying restriction(s) is wrong. In the example used throughout this discussion, the overidentifying restriction takes the simplest possible form; it asserts the equality of two ratios of reduced-form coefficients: $a_{42}/a_{52} = a_{43}/a_{53}$. In more highly overidentified models there will be several such conditions. But the rejection of the null hypothesis only tells one that "something" is wrong, not what in particular is likely to be wrong.

This is true even in our simple example. If we must reject the overidentifying restriction, how may we remedy the situation? If we draw in a causal arrow, $x_2 \rightarrow x_4$, and thereby revise the x_4-equation to read,

$$x_4 = b_{41}x_1 + b_{42}x_2 + b_{45}x_5 + u$$

this equation, as well as the x_5-equation, will be just identified. There will be no overidentifying restriction(s) and thus no way of rejecting the model for failure to fit the overidentifying restriction(s). But exactly the same thing will be true if, instead, we put in the arrow, $x_3 \rightarrow x_4$, so that the x_4-equation will read

$$x_4 = b_{41}x_1 + b_{43}x_3 + b_{45}x_5 + u$$

The result of the original test of the overidentifying restriction is of no help in deciding which (if not some other) route to take. A formal analysis can only reveal formal conditions that a good model must satisfy (or satisfy approximately). Whether it is really any good must be determined on substantive grounds, with the guidance of the best theory available.

Exercise. *What happens if we revise the x_4-equation to include both x_2 and x_3?*

Exercise. *Enumerate all possible two-equation nonrecursive models comprising two endogenous and three exogenous variables, each equation of every model being just identified. Let every model have the same set of endogenous variables (say, x_4 and x_5) and the same set of exogenous variables (x_1, x_2, and x_3). What possibility, if any, do you see for letting the choice of one from among this list of models be decided by the results of a statistical analysis?*

FURTHER READING

An overidentified sociological model is estimated by 2SLS in Duncan, Haller, and Portes (in Blalock, 1971, Chapter 13), but the presentation is unnecessarily complicated by the use of standardized variables.