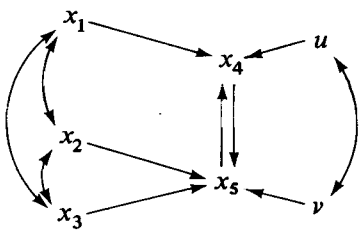
7

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 over-

identifying 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.