3

Specification Error

The title of this chapter, like much of the nomenclature and content of this book, derives from the literature on econometrics. It is quite a useful euphemism for what in a blunter language would be called "using the wrong model." There are many more wrong models than right ones, so that specification error is very common, though often not recognized and usually not easily recognizable.

The main message of the entire book would be missed if the reader did not understand by now that we can seldom be sure we have the right model, although we can sometimes be nearly certain, on the basis of empirical evidence, that we have been using the wrong one. Indeed, it would require no elaborate sophistry to show that we will never have the "right" model in any absolute sense. Hence, we shall never be able to compare one of our many wrong models with a definitively right one. Is the matter of specification error beyond rational discussion, therefore, even if we have no difficulty in "taking about" it while failing to "discuss" it?

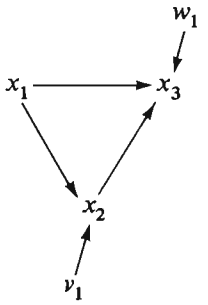
As the term will be used here, analysis of specification error relates to a rhetorical strategy in which we suggest a model as the "true" one for sake of argument, determine how our working model differs from it and what the consequences of the difference(s) are, and thereby get

some sense of how important the mistakes we will inevitably make may be.

Sometimes it is possible to secure genuine comfort by this route. For example, in using an estimation method like 2SLS, where we estimate the coefficients of just one equation in the model at a time, it turns out that having specification errors in the other equations does not matter, if only we have designated correctly the predetermined variables that are excluded from the particular equation being estimated. (This is fairly "obvious" from a review of how this method goes, and no further proof is offered here.) It could happen (if one is lucky) that the particular equation is the crucial one for the light that its coefficients shed on the theory we are working with. Hence, this partial insulation from the effects of specification error can be welcome.

A similar theorem was stated for recursive models in Chapter 3 (page 43). In estimating path (or structural) coefficients in one equation, we do not harm the results by having an erroneous causal ordering of the variables in "prior" equations. Of course, we do harm to the results for the model as a whole, so it is well not to take too much comfort from this theorem.

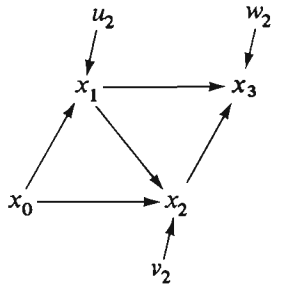
It should not be supposed, however, that a mistake in regard to the causal ordering is the only form of specification error. To illustrate the contrary, suppose we are working with three variables and have a firm basis for the causal ordering, $x_1 \rightarrow x_2 \rightarrow x_3$. Can we be sure that it is safe to take



Model 1

as our model (especially since it subsumes the case in which any one of the structural coefficients is zero)?

Imagine the true state of affairs is



Model 2

(The disturbances carry subscripts because v and w in Model 2 are not the same as v and w in Model 1.) You should be able to see at a glance that Model 1 is faulty. Even though x_1 is causally prior to x_2 and x_3, it is not legitimate (in the light of "true" Model 2) to treat x_1 by itself as the sole exogenous variable. To demonstrate the specification error more rigorously and to discover its precise consequences requires some algebra of the kind we have been doing throughout our study.

Exercise. *Write down the equations for each model and from them derive expressions for the variances and covariances of all dependent variables. Set up the formulas for* OLS *estimates of structural coefficients.*

We find that the structural coefficients and OLS estimates thereof are the same for the x_3-equation in Model 2 as in Model 1. For the x_2-equation, however, we find that Model 1 (with the usual specification on the disturbance) implies $b_{21} = \sigma_{12}/\sigma_{11}$. But, from "true" Model 2 we see (given the results of the Exercise) that

$$\frac{\sigma_{12}}{\sigma_{11}} = \frac{b_{21}\sigma_{11} + b_{20}\sigma_{01}}{\sigma_{11}}$$

$$= b_{21} + b_{20}\frac{\sigma_{01}}{\sigma_{11}}$$

Hence, if we use the "false" estimator,

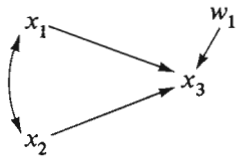
$$\hat{b}_{21}^{\mathrm{F}} = \frac{m_{12}}{m_{11}}$$

we shall be estimating *not* b_{21} but a quantity that differs from b_{21} by an amount $b_{20}\sigma_{01}/\sigma_{11}$, which will be positive or negative according to whether b_{20} and σ_{01} have the same or different signs. Indeed, since $\sigma_{01} = b_{10}\sigma_{00}$ and $\sigma_{11} = b_{10}^2\sigma_{00} + \sigma_{u_2u_2}$ (you should use results of the exercise to verify this claim), we see that the bias in $\hat{b}_{21}^{F}$ can be written

$$b_{20}b_{10}\frac{\sigma_{00}}{b_{10}^2\sigma_{00} + \sigma_{u_2u_2}}$$

so that its sign depends on whether b_{20} and b_{10} have like or different signs (the remaining terms in the expression being intrinsically positive). We also see that if either b_{20} or b_{10} is zero, the bias is nil. (This agrees with the results of common sense in comparing the two path diagrams—or, if it does not, you need further to develop your common sense about these diagrams.) If the investigator can offer plausible considerations favoring the view that one or the other of these coefficients is "small," he may take comfort in the implication that the bias in his estimator ($\hat{b}_{21}^{F}$) is small or, equivalently, that Model 1 is "nearly true." Such comfort would be prized in the realistic situation where Model 2 is "known" to be true, but measurements on x_0 are not available, whether by reason of some defect in the study design or because x_0 is not an observable variable.

Suppose, however, that x_0, though not observed, must be assumed to have substantial effects on both x_1 and x_2. What then is our recourse? We can only adjust to this unsatisfactory state of affairs by substituting for (false) Model 1 the following revised version:



Model 1′

The fact that models like Model 1′ are intrinsically less informative than those like Model 1 (if only the latter were true!) has been remarked in Chapter 3 (pages 36–43).

There is another instructive way to describe the specification error.

Let us compare the x_2-equations of the two models.

$$x_2 = b_{21}x_1 + v_1 \qquad \text{(Model 1)}$$

$$x_2 = b_{21}x_1 + b_{20}x_0 + v_2 \qquad \text{(Model 2)}$$

Now, we can render Model 1 "true" by resolving to think of b_{21} as the same coefficient, regardless of which model it appears in, and noting that this requires us to recognize a relationship between the disturbances in the two models, given by

$$v_1 = b_{20}x_0 + v_2$$

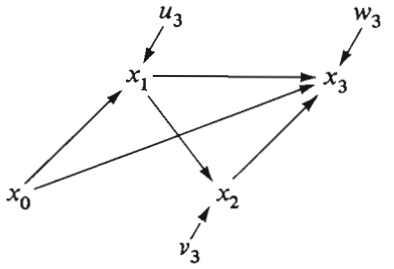
Evaluating $E(x_1 v_1)$ we find

$$E(x_1 v_1) = b_{20}E(x_0 x_1) + E(x_1 v_2)$$

$$= b_{20}\sigma_{01}$$

since in "true" Model 2 the usual specification on the disturbance would be $E(x_1 v_2) = 0$. But it is precisely this "usual" specification that we cannot legitimately adopt for Model 1, given the truth of Model 2. Another way to put the whole issue, then, is that in Model 1 (as we know from comparing it with Model 2) we must *not* assume $E(x_1 v_1) = 0$. But with that assumption not available, we cannot rely on the OLS estimate of b_{21} in the framework of Model 1. We come full circle back to the statement in Chapter 1 (page 5), "The assumption that an explanatory or causal variable is uncorrelated with the disturbance must always be weighed carefully." Perhaps the meaning of "to weigh carefully" will be a little clearer now than it was seven chapters ago.

Let us make use of this alternative way of describing specification error in attacking another illustrative problem. Suppose, now, the "true" model is



Model 3

Studying Model 3 after the fashion of Chapter 4 we find among other things:

$$\sigma_{01} = b_{10}\sigma_{00}$$
$$\sigma_{11} = b_{10}^2\sigma_{00} + \sigma_{u_3u_3}$$
$$\sigma_{02} = b_{21}b_{10}\sigma_{00}$$
$$\sigma_{12} = b_{21}b_{10}^2\sigma_{00} + b_{21}\sigma_{u_3u_3}$$
$$\sigma_{22} = b_{21}^2b_{10}^2\sigma_{00} + b_{21}^2\sigma_{u_3u_3} + \sigma_{v_3v_3}$$

Exercise. *Verify these results.*

Now, we propose to study Model 1 in light of true Model 3. But, this time, we make explicit the fact that $E(x_1 w_1) \neq 0$. That is, we take the x_3-equation from Model 1,

$$x_3 = b_{32}x_2 + b_{31}x_1 + w_1$$

and, after observing that if Model 3 is true,

$$w_1 = b_{30}x_0 + w_3$$

rewrite the x_3-equation as

$$x_3 = b_{32}x_2 + b_{31}x_1 + (b_{30}x_0 + w_3)$$

Now we multiply through by x_2 and x_1:

$$\sigma_{23} = b_{32}\sigma_{22} + b_{31}\sigma_{12} + (b_{30}\sigma_{02})$$
$$\sigma_{13} = b_{32}\sigma_{12} + b_{31}\sigma_{11} + (b_{30}\sigma_{01})$$

taking advantage of the specification in Model 3 that $E(x_1 w_3) = E(x_2 w_3) = 0$. Just to make the next step easier to follow, these two equations are rewritten

$$(\sigma_{23} - b_{30}\sigma_{02}) = b_{32}\sigma_{22} + b_{31}\sigma_{12}$$
$$(\sigma_{13} - b_{30}\sigma_{01}) = b_{32}\sigma_{12} + b_{31}\sigma_{11}$$

We solve for b_{32}:

$$b_{32} = \frac{\sigma_{11}(\sigma_{23} - b_{30}\sigma_{02}) - \sigma_{12}(\sigma_{13} - b_{30}\sigma_{01})}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}$$
$$= \frac{\sigma_{11}\sigma_{23} - \sigma_{12}\sigma_{13}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} + b_{30}\frac{\sigma_{01}\sigma_{12} - \sigma_{02}\sigma_{11}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}$$

But the second term on the right vanishes, as will be seen upon evaluating the numerator, using the expressions for the σ's already exhibited. This shows that the OLS estimator of b_{32} from Model 1 is unbiased, for if x_3 is regressed on x_2 and x_1 the estimator is

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

On the other hand, when we solve for the other structural coefficient, we obtain

$$b_{31} = \frac{\sigma_{22}\sigma_{13} - \sigma_{12}\sigma_{23}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} + b_{30}\frac{\sigma_{02}\sigma_{12} - \sigma_{01}\sigma_{22}}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}$$

The second term on the right does not vanish, so that the OLS estimator from Model 1,

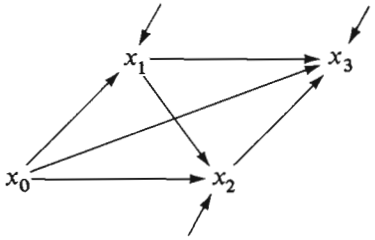
$$\hat{b}_{31}^{\mathrm{F}} = \frac{m_{22}m_{13} - m_{12}m_{23}}{m_{11}m_{22} - m_{12}^2}$$

is, in fact, biased and the bias is given by the non-zero value of that second term.

This example shows that despite the specification error, OLS applied to Model 1 would have yielded an unbiased estimate for b_{32}, though not for b_{31}. One must, perhaps, be hungry indeed to take much nourishment from such a result. Still, one can imagine a realistic situation in which x_0, though unobserved, may (with reason) be postulated to behave according to Model 3. In that event, we can still do *something* with Model 1. Certainly, it is important to realize that specification error may have diverse effects, according to where it occurs. Hence, it is not enough for a skeptical critic to say, "The model is improperly specified, hence the estimates of structural coefficients are biased." He must, on the contrary—or *you* must, taking the role of

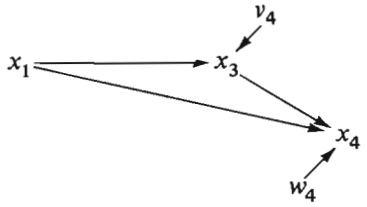
critic—propose a new, "true" model that shows just wherein the initial model errs, and then compare the two to infer as much as possible about the consequences of the specification error.

Exercise. *We have seen that if Model 2 is true we can still salvage the x_3-equation from Model 1. If Model 3 is true, show that we can salvage the x_2-equation as well as an unbiased* OLS *estimate of b_{32}. If the true model is this,*



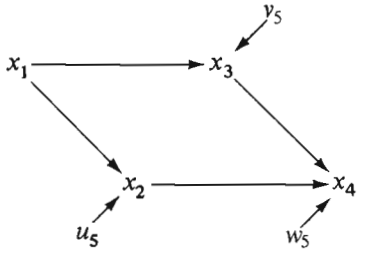
what can be salvaged from Model 1?

To illustrate still another situation, let the initial, possibly erroneous model be



Model 4

Further consideration of relevant theory shows that the "true" model is



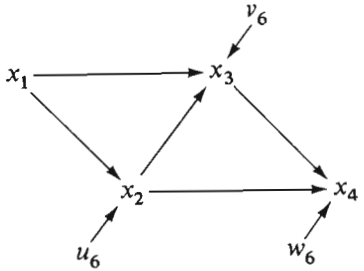
Model 5

This is an easy one.

Exercise. *Show that the* OLS *estimator of b_{43} in Model 4 estimates b_{43} in Model 5 without bias and that the* OLS *estimator of b_{41} in Model 4 estimates $b_{42}b_{21}$ without bias.*

The principle is that insertion of an "intervening variable" into one path of an initial model does not invalidate that model, but merely elaborates it.

But suppose the "true" model is



Model 6

The elements of this situation have already been covered in the episode of Model 1 versus Model 2, although the additional variable turns up at a different place in the causal ordering.

Exercise. *If Model 6 is true, with the usual specification on its disturbances, show that the specification on the disturbances in Model 4 may include $E(x_1 w_4) = E(x_1 v_4) = 0$ but not $E(x_3 w_4) = 0$. Set forth the implications for* OLS *estimation of structural coefficients in Model 4.*

Here the principal is that insertion of an "intervening variable" into an initial model, where the intervening variable then operates as a common cause of two or more later variables, reveals specification errors in the initial model. Every such case must be evaluated on its own terms. In the illustration at hand (Model 6 versus Model 4) we may make reference to the earlier experience (Model 3 versus Model 1) to observe that specification error may have one-sided consequences. Let us write the "semireduced form" equations for Model 6, eliminating x_2 from the other two equations:

$$x_3 = (b_{31} + b_{32}b_{21})x_1 + b_{32}u_6 + v_6$$

$$x_4 = b_{43}x_3 + b_{42}b_{21}x_1 + b_{42}u_6 + w_6$$

Or, more compactly,

$$x_3 = a_{31}x_1 + v'_6$$
$$x_4 = a_{43}x_3 + a_{41}x_1 + w'_6$$

where

$$a_{31} = b_{31} + b_{32}b_{21}$$
$$a_{43} = b_{43}$$
$$a_{41} = b_{42}b_{21}$$
$$v'_6 = b_{32}u_6 + v_6$$
$$w'_6 = b_{42}u_6 + w_6$$

We see that a_{31} may be estimated by OLS regression of x_3 on x_1, since $E(x_1v'_6) = 0$.

Exercise. *Show that* $E(x_1v'_6) = 0$.

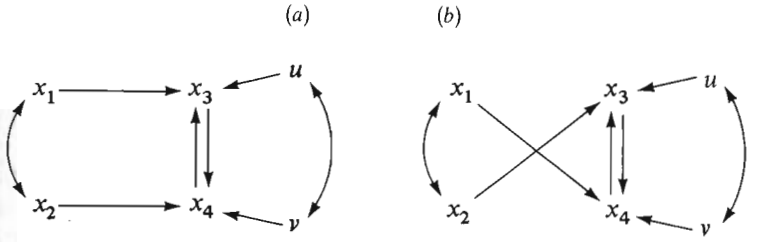
But OLS is not suitable for estimating a_{43} and a_{41}, since not all the explanatory variables are uncorrelated with w'_6, the disturbance in the semireduced form of the x_4-equation. We note that the semireduced form of Model 6 is just Model 4, with the appropriate reservation concerning nonzero covariances of causal variables and disturbances.

Much theorizing in sociology takes the form of suggesting "intervening variables" to interpret causal linkages that have been recognized earlier. For the research worker using structural equation models, it becomes vital to know whether the hypotheses about these intervening variables tend to leave his work intact (though pointing toward a useful elaboration of it) or whether they tend to suggest that his estimates of causal influences are seriously biased. Since our examples of specification error have been nonnumerical, it is easy to miss the point that specification errors may engender biases that are real but nonetheless quantitatively trivial. In addition to a qualitative analysis of the nature of the bias (if any), the investigator will, therefore, often put hypothetical (but conceivable) values on certain structural coefficients in a "true" model in order to conjecture, with some plausibility, how wide of the mark he may be if he continues to work

with the erroneous one. Such "sensitivity analysis" may sometimes provide comfort where otherwise the examination of specification error would yield pessimistic conclusions as to the acceptability of a model.

We have given only a few rudimentary illustrations of specification error but, it is hoped, enough to suggest that the serious investigator will muster all his ingenuity to anticipate threats to the validity of his results from this source. We have focussed on the "omitted variable" as a threat to validity, because this is one of the most common arguments encountered in discussions of models—perhaps because it is easy to suggest the name of a variable that has been overlooked, though not always so easy to justify a "true" model that includes it. Specification error arises in other ways, however.

Suppose one is in a position to compare estimates obtained for these two models:



At least one of these two models ought to give results that are questionable in the light of theoretical expectations. If not, it would appear that the domain under study actually is not very highly structured in causal terms, or the measurements all are grossly contaminated by error, or our theory is virtually noncommittal on the significant issues. In any event, the clue to specification error in (*a*) or (*b*) is primarily the substantive implausibility of the estimates.

Exercise. *Let the coefficients in diagram* (a) *be* b_{31}, b_{34}, b_{42}, b_{43} *and the coefficients in diagram* (*b*) *be* c_{32}, c_{34}, c_{41}, c_{43}. *Show that* $c_{43} = 1/b_{34}$ *and* $c_{34} = 1/b_{43}$. *Similarly, express* c_{32} *and* c_{41} *in terms of the b's. How would you answer a mathematician who claims that Models* (*a*) *and* (*b*) *are interchangeable, so that if one is "true" the other must be "true" too?*

In addition to wrongly omitted variables, the model may involve erroneously included variables. In theory, one such mistake should not be fatal if the model is otherwise correctly specified. The erroneously included variable should have a nearly zero coefficient. However, a mistake of this kind does impair the efficiency with which the model's coefficients are estimated. Hence, the investigator should not try to get by on the strategy of "including everything" on an initial run of his model. This strategy, pursued relentlessly, leads to underidentification, as we have seen in Chapter 6 (page 87).

Other issues properly subsumed under specification error include (*1*) tests of overidentifying restrictions (Chapter 3, pages 46–50; Chapter 7, pages 98–99); (*2*) the validity of the specification of linear and additive functional form, a matter that receives considerable attention in most statistical presentations of linear models; and (*3*) the acceptability of the homoskedasticity assumption in regard to disturbances, likewise treated in some statistics texts. The neglect of these last two topics in our sketch of this subject should not be mistaken for a judgement that they are unimportant. On the contrary, they are so important that they must be squarely faced throughout a project in constructing a model, but especially when considering the initial specification of the model. Statistical tests of linearity and homoskedasticity may be of use, but detailed inspection of the data aided by graphic plots is perhaps even more useful.

Exercise. *On page 17 we suggested that one could pose a countermodel to Model* II′ *as a means of discussing possible specification error in that model. On page 18, we made a similar suggestion regarding Model* III. *If you have not already done so, show how this might be done in each instance.*

FURTHER READING

The text by Rao and Miller (1971) is unusual in regard to the amount of attention given to matters relevant to this topic and also in that it is accessible to the reader not familiar with matrix algebra.

9

Measurement Error, Unobserved Variables

From a formal point of view, the topic of error in (measurement of) variables is much the same thing as that of unobserved variables. All observation is fallible, no matter how refined the measuring instrument and no matter how careful the procedure of applying it. In a strict sense, therefore, we never measure exactly the true variables discussed in our theories. In this same strict sense, all (true) variables are "unobserved."

It may happen that errors of measurement are negligible, relative to the magnitudes of the disturbances in our equations or the standard errors of sampling in our estimates of coefficients. Of course, an investigator will not blithely assume this is so, but will make every effort to assess his measurement errors and their impact on his results. If the verdict is reassuring (perhaps because the other threats to valid inference are so uncomfortably large!) he may proceed, for the moment, to treat his variables as error-free, as we have been doing implicitly throughout this book.

A second possibility is that measurement error is appreciable but "well-behaved" and, perhaps, estimable. That is, a relatively simple and manageable specification of the behavior of the errors is acceptable. Either the errors can be shown not to impair seriously the results