5.4 Some Conceptual Underpinnings

5.4.1 What Do Structural Parameters Really Mean?

Every student of SEM has stumbled on the following paradox at some point in his or her career. If we interpret the coefficient β in the equation

$$y = \beta x + \epsilon$$

as the change in $E(Y)$ per unit change of X, then, after rewriting the equation as

$$x = (y - \epsilon)/\beta,$$

we ought to interpret $1/\beta$ as the change in $E(X)$ per unit change of Y. But this conflicts both with intuition and with the prediction of the model: the change in $E(X)$ per unit change of Y ought to be *zero* if Y does not appear as an independent variable in the original, structural equation for X.

Teachers of SEM generally evade this dilemma via one of two escape routes. One route involves denying that β has any causal reading and settling for a purely statistical interpretation, in which β measures the reduction in the variance of Y explained by X (see e.g. Muthen 1987). The other route permits causal reading of only those coefficients that meet the "isolation" restriction (Bollen 1989; James et al. 1982): the explanatory variable must be uncorrelated with the error in the equation. Because ϵ cannot be uncorrelated with both X and Y (or so the argument goes), β and $1/\beta$ cannot both have causal meaning, and the paradox dissolves.

The first route is self-consistent, but it compromises the founders' intent that SEM function as an aid to policy making and clashes with the intuition of most SEM users. The second is vulnerable to attack logically. It is well known that every pair of bivariate normal variables, X and Y, can be expressed in two equivalent ways,

$$y = \beta x + \epsilon_1 \text{ and } x = \alpha y + \epsilon_2,$$

where $\text{cov}(X, \epsilon_1) = \text{cov}(Y, \epsilon_2) = 0$ and $\alpha = r_{XY} = \beta \sigma_X^2 / \sigma_Y^2$. Thus, if the condition $\text{cov}(X, \epsilon_1) = 0$ endows β with causal meaning, then $\text{cov}(Y, \epsilon_2) = 0$ ought to endow α with causal meaning as well. But this, too, conflicts with both intuition and the intentions behind SEM; the change in $E(X)$ per unit change of Y ought to be zero, not r_{XY}, if there is no causal path from Y to X.

What then *is* the meaning of a structural coefficient? Or a structural equation? Or an error term? The interventional interpretation of causal effects, when coupled with the $do(x)$ notation, provides simple answers to these questions. The answers explicate the operational meaning of structural equations and thus should end, I hope an era of controversy and confusion regarding these entities.

Structural Equations: Operational Definition

Definition 5.4.1 (Structural Equations)
An equation $y = \beta x + \epsilon$ is said to be structural *if it is to be interpreted as follows: In an ideal experiment where we control X to x and any other set Z of variables $($not containing X or $Y)$ to z, the value y of Y is given by $\beta x + \epsilon$, where ϵ is not a function of the settings x and z.*

This definition is operational because all quantities are observable, albeit under conditions of controlled manipulation. That manipulations cannot be performed in most observational studies does not negate the operationality of the definition, much as our inability to observe bacteria with the naked eye does not negate their observability under a microscope. The challenge of SEM is to extract the maximum information concerning what we wish to observe from the little we actually can observe.

Note that the operational reading just given makes no claim about how X (or any other variable) will behave when we control Y. This asymmetry makes the equality signs in structural equations different from algebraic equality signs; the former act symmetrically in relating observations on X and Y (e.g., observing $Y = 0$ implies $\beta x = -\epsilon$), but they act asymmetrically when it comes to interventions (e.g., setting Y to zero tells us nothing about the relation between x and ϵ). The arrows in path diagrams make this dual role explicit, and this may account for the insight and inferential power gained through the use of diagrams.

The strongest empirical claim of the equation $y = \beta x + \epsilon$ is made by excluding other variables from the r.h.s. of the equation, thus proclaiming X the *only* immediate cause of Y. This translates into a testable claim of *invariance*: the statistics of Y under condition $do(x)$ should remain invariant to the manipulation of any other variable in the model (see Section 1.3.2).[16] This claim can be written symbolically as

$$P(y|do(x), do(z)) = P(y|do(x)) \tag{5.23}$$

for all Z disjoint of $\{X \cup Y\}$.[17]

Note that this invariance holds relative to manipulations, not observations, of Z. The statistics of Y under condition $do(x)$ given the measurement $Z = z$, written

[16] The basic notion that structural equations remain invariant to certain changes in the system goes back to Marschak (1950) and Simon (1953), and it has received mathematical formulation at various levels of abstraction in Hurwicz (1962), Mesarovic (1969), Sims (1977), Cartwright (1989), Hoover (1990), and Woodward (1995). The simplicity, precision, and clarity of (5.23) is unsurpassed, however.

[17] This claim is, in fact, only part of the message conveyed by the equation; the other part consists of a dynamic or counterfactual claim: If we were to control X to x' instead of x, then Y would attain the value $\beta x' + \epsilon$. In other words, plotting the value of Y under various hypothetical controls of X, and under the same external conditions (ϵ), should result in a straight line with slope β. Such deterministic dynamic claims concerning system behavior under successive control conditions can only be tested under the assumption that ϵ, representing external conditions or properties of experimental units, remains unaltered as we switch from x to x'. Such counterfactual claims constitute the empirical content of every scientific law (see Section 7.2.2).

$P(y|do(x), z)$, would certainly depend on z if the measurement were taken on a consequence (i.e. descendant) of Y. Note also that the ordinary conditional probability $P(y|x)$ does not enjoy such a strong property of invariance, since $P(y|x)$ is generally sensitive to manipulations of variables other than X in the model (unless X and ϵ are independent). Equation (5.23), in contrast, remains valid regardless of the statistical relationship between ϵ and X.

Generalized to a set of several structural equations, (5.23) explicates the assumptions underlying a given causal diagram. If G is the graph associated with a set of structural equations, then the assumptions are embodied in G as follows: (1) every missing arrow—say, between X and Y—represents the assumption that X has no causal effect on Y once we intervene and hold the parents of Y fixed; and (2) every missing bidirected link between X and Y represents the assumption that the omitted factors that (directly) influence X are uncorrelated with those that (directly) influence Y. We shall define the operational meaning of the latter assumption in (5.25)–(5.27).

The Structural Parameters: Operational Definition

The interpretation of a structural equation as a statement about the behavior of Y under a hypothetical intervention yields a simple definition for the structural parameters. The meaning of β in the equation $y = \beta x + \epsilon$ is simply

$$\beta = \frac{\partial}{\partial x} E[Y|do(x),] \tag{5.24}$$

that is, the rate of change (relative to x) of the expectation of Y in an experiment where X is held at x by external control. This interpretation holds regardless of whether ϵ and X are correlated in nonexperimental studies (e.g., via another equation $x = \alpha y + \delta$).

We hardly need to add at this point that β has nothing to do with the regression coefficient r_{YX} or, equivalently, with the conditional expectation $E(Y|x)$, as suggested in many textbooks. The conditions under which β coincides with the regression coefficient are spelled out in Theorem 5.3.1.

It is important nevertheless to compare the definition of (5.24) with theories that acknowledge the invariant character of β but have difficulties explicating which changes β is invariant to. Cartwright (1989, p. 194), for example, characterizes β as an invariant of nature that she calls "capacity." She states correctly that β remains constant under change but explains that, as the statistics of X changes, "it is the ratio $[\beta = E(YX)/E(X^2)]$ which remains fixed no matter how the variances shift." This characterization is imprecise on two accounts. First, β may in general not be equal to the stated ratio nor to any other combination of statistical parameters. Second—and this is the main point of Definition 5.4.1—structural parameters are invariant to local interventions (i.e., changes in specific equations in the system) and not to general changes in the statistics of the variables. If we start with $\mathrm{cov}(X, \epsilon) = 0$ and the variance of X changes because we (or Nature) locally modify the *process* that generates X, then Cartwright is correct; the ratio $\beta = E(YX)/E(X^2)$ will remain constant. However, if the variance of X changes for any other reason—say, because we observed some evidence $Z = z$ that depends on both X and Y or because the process generating X becomes dependent on a wider set of variables—then that ratio will not remain constant.

The Mystical Error Term: Operational Definition

The interpretations given in Definition 5.4.1 and (5.24) provide an operational definition for that mystical error term

$$\epsilon = y - E[Y|do(x)], \tag{5.25}$$

which, despite being unobserved in nonmanipulative studies, is far from being metaphysical or definitional as suggested by some researchers (e.g. Richard 1980; Holland 1988, p. 460; Hendry 1995, p. 62). Unlike errors in regression equations, ϵ measures the deviation of Y from its controlled expectation $E[Y|do(x)]$ and not from its conditional expectation $E[Y|x]$. The statistics of ϵ can therefore be measured from observations on Y once X is controlled. Alternatively, because β remains the same regardless of whether X is manipulated or observed, the statistics of $\epsilon = y - \beta x$ can be measured in observational studies if we know β.

Likewise, correlations among errors can be estimated empirically. For any two nonadjacent variables X and Y, (5.25) yields

$$E[\epsilon_Y \epsilon_X] = E[YX|do(pa_Y, pa_X)] - E[Y|do(pa_Y)]E[X|do(pa_X)]. \tag{5.26}$$

Once we have determined the structural coefficients, the controlled expectations $E[Y|do(pa_Y)]$, $E[X|do(pa_X)]$, and $E[YX|do(pa_Y, pa_X)]$ become known linear functions of the observed variables pa_Y and pa_X; hence, the expectations on the r.h.s. of (5.26) can be estimated in observational studies. Alternatively, if the coefficients are not determined, then the expression can be assessed directly in interventional studies by holding pa_X and pa_Y fixed (assuming X and Y are not in parent-child relationship) and estimating the covariance of X and Y from data obtained under such conditions.

Finally, we are often interested not in assessing the numerical value of $E[\epsilon_Y \epsilon_X]$ but rather in determining whether ϵ_Y and ϵ_X can be assumed to be uncorrelated. For this determination, it suffices to test whether the equality

$$E[Y|x, do(s_{XY})] = E[Y|do(x), do(s_{XY})] \tag{5.27}$$

holds true, where s_{XY} stands for (any setting of) all variables in the model excluding X and Y. This test can be applied to any two variables in the model *except* when Y is a parent of X, in which case the symmetrical equation (with X and Y interchanged) is applicable.

The Mystical Error Term: Conceptual Interpretation

The authors of SEM textbooks usually interpret error terms as representing the influence of omitted factors. Many SEM researchers are reluctant to accept this interpretation, however, partly because unspecified omitted factors open the door to metaphysical speculations and partly because arguments based on such factors were improperly used as a generic, substance-free license to omit bidirected arcs from path diagrams (McDonald 1997). Such concerns are answered by the operational interpretation of error terms, (5.25), since it prescribes how errors are measured, not how they originate.

It is important to note, though, that this operational definition is no substitute for the omitted-factors conception when it comes to deciding whether pairs of error terms can be assumed to be uncorrelated. Because such decisions are needed at a stage when the model's parameters are still "free," they cannot be made on the basis of numerical assessments of correlations but must rest instead on qualitative structural knowledge about how mechanisms are tied together and how variables affect each other. Such judgmental decisions are hardly aided by the operational criterion of (5.26), which instructs the investigator to assess whether two deviations—taken on two different variables under complex experimental conditions—would be correlated or uncorrelated. Such assessments are cognitively unfeasible.

In contrast, the omitted-factors conception instructs the investigator to judge whether there could be factors that simultaneously influence several observed variables. Such judgments are cognitively manageable because they are qualitative and rest on purely structural knowledge—the only knowledge available during this phase of modeling.

Another source of error correlation that should be considered by investigators is *selection bias*. If two uncorrelated unobserved factors have a common effect that is omitted from the analysis but influences the selection of samples for the study, then the corresponding error terms will be correlated in the sampled population; hence, the expectation in (5.26) will not vanish when taken over the sampled population (see discussion of Berkson's paradox in Section 1.2.3).

We should emphasize, however, that the arcs *missing* from the diagram, not those *in* the diagram, demand the most attention and careful substantive justification. Adding an extra bidirected arc can at worst compromise the identifiability of parameters, but deleting an existing bidirected arc may produce erroneous conclusions as well as a false sense of model testability. Thus, bidirected arcs should be assumed to exist, by default, between any two nodes in the diagram. They should be deleted only by well-motivated justifications, such as the unlikely existence of a common cause for the two variables and the unlikely existence of selection bias. Although we can never be cognizant of all the factors that may affect our variables, substantive knowledge sometimes permits us to state that the influence of a possible common factor is not likely to be significant.

Thus, as often happens in the sciences, the way we measure physical entities does not offer the best way of thinking about them. The omitted-factor conception of errors, because it rests on structural knowledge, is a more useful guide than the operational definition when building, evaluating, and thinking about causal models.

5.4.2 Interpretation of Effect Decomposition

Structural equation modeling prides itself, and rightly so, for providing principled methodology for distinguishing direct from indirect effects. We have seen in Section 4.5 that such distinction is important in many applications, ranging from process control to legal disputes, and that SEM indeed provides a coherent methodology of defining, identifying, and estimating direct and indirect effects. However, the reluctance of most SEM researchers to admit the causal reading of structural parameters—coupled with their preoccupation with algebraic manipulations—has resulted in inadequate definitions of direct and indirect effects, as pointed out by Freedman (1987) and Sobel (1990). In this section we hope to correct this confusion by adhering to the operational meaning of the structural coefficients.

We start with the general notion of a causal effect $P(y|do(x))$, as in Definition 3.2.1. We then specialize it to define direct effect, as in Section 4.5, and finally express the definitions in terms of structural coefficients.

Definition 5.4.2 (Total Effect)
The total effect *of X on Y is given by $P(y|do(x))$, namely, the distribution of Y while X is held constant at x and all other variables are permitted to run their natural course.*

Definition 5.4.3 (Direct Effect)
The direct effect *of X on Y is given by $P(y|do(x),\ do(s_{XY}))$, where S_{XY} is the set of all observed variables in the system except X and Y.*

In linear analysis, Definitions 5.4.2 and 5.4.3 yield, after differentiation with respect to x, the familiar path coefficients in terms of which direct and indirect effects are usually defined. Yet they differ from conventional definitions in several important aspects. First, direct effects are defined in terms of hypothetical experiments in which intermediate variables are held constant by *physical intervention*, not by statistical adjustment (which is often disguised under the misleading phrase "control for"). Figure 5.10 depicts a simple example where adjusting for the intermediate variables $(Z \text{ and } W)$ would not give the correct value of zero for the direct effect of X on Y, whereas $\frac{\partial}{\partial_x} E(Y|do(x,y,w))$ does yield the correct value: $\frac{\partial}{\partial_x}(\beta w + \gamma z) = 0$. Section 4.5.3 (Table 4.1) provides another such example, one that involves dichotomous variables.

Second, there is no need to limit control to only intermediate variables; *all* variables in the system may be held constant (except for X and Y). Hypothetically, the scientist controls for all possible conditions S_{XY}, and measurements may commence without knowing the structure of the diagram. Finally, our definitions differ from convention by interpreting total and direct effects independently of each other, as outcomes of two different experiments. Textbook definitions (e.g. Bollen 1989, p. 376; Mueller 1996, p. 141; Kline 1998, p. 175) usually equate the total effect with a power series of path coefficient matrices. This algebraic definition coincides with the operational definition (Definition 5.4.2) in recursive (semi-Markovian) systems, but it yields erroneous expressions in models with feedback. For instance, given the pair of equations $\{y = \beta x + \epsilon,\ x = \alpha y + \delta\}$, the total effect of X on Y is simply β, not $\beta(1-\alpha\beta)^{-1}$ as stated in Bollen (1989, p. 379). The latter has no operational significance worthy of the phrase "effect of X."[18]

We end this section of effect decomposition with a few remarks that should be of interest to researchers dealing with dichotomous variables. The relations among such

[18] This error was noted by Sobel (1990) but, perhaps because constancy of path coefficients was presented as a new and extraneous assumption, Sobel's correction has not brought about a shift in practice or philosophy.

variables are usually nonlinear, so the results of Section 4.5 should be applicable. In particular, the direct effect of X on Y should will depend on the levels at which we hold the other parents of Y. If we wish to average over these values, we obtain the expression given in Section 4.5.4.

In standard linear analysis, an indirect effect may be defined as the difference between the total effect and the direct effects (Bollen 1989). In nonlinear analysis, differences lose their significance, and one must isolate the contribution of mediating paths in some other way. Expressions of the form $P(y|do(x), do(z))$ cannot be used to isolate such contributions because there is no physical means of selectively disabling a direct causal link from X to Y by holding some variables constant. This suggests that the notion of indirect effect has no intrinsic operational meaning apart from providing a comparison between the direct and the total effects. In other words, a policy maker who asks for that part of the total effect transmitted by a particular intermediate variable or by a group Z of such variables is really asking for a comparison of the effects of two policies, one where Z is held constant versus the other where it is not. The expressions corresponding to these policies are $P(y|do(x),\ do(z))$ and $P(y|do(x))$, and this pair of distributions should be taken as the most general representation of indirect effects. Similar conclusions have been expressed by Robins (1986) and Robins and Greenland (1992).

5.4.3 Exogeneity, Superexogeneity, and Other Frills

Economics textbooks invariably warn readers that the distinction between exogenous and endogenous variables is, on the one hand, “most important for model building” (Darnell 1994, p. 127) and, on the other hand, “a subtle and sometimes controversial complication” (Greene 1997, p. 712). Economics students would naturally expect the concepts and tools developed in this chapter to shed some light on the subject, and rightly so. We next offer a simple definition of exogeneity that captures the important nuances appearing in the literature and that is both palatable and precise,

It is fashionable today to distinguish three types of exogeneity: weak, strong, and super (Engle et al. 1983); the former two are statistical and the latter causal. However, the importance of exogeneity—and the reason for its controversial status—lies in its implications for policy interventions. Some economists believe, therefore, that only the causal aspect (i.e. superexogeneity) deserves the exogenous title and that the statistical versions are unwarranted intruders that tend to confuse issues of identification and interpretability with those of estimation efficiency (Ed Leamer, personal communication).[19] I will serve both camps by starting with a simple definition of causal exogeneity and then offering a more general definition, from which both the causal and the statistical aspects would follow as special cases. Thus, what we call “exogeneity” corresponds to what Engle et al. called “superexogeneity,” a notion that captures economists’ interest in the structural invariance of certain relationships under policy intervention.

Suppose that we consider intervening on a set of variables X and that we wish

[19] Similiar opinions have also been communicated by John Aldrich and James Heckman. See also Aldrich (1993).

to characterize the statistical behavior of a set Y of outcome variables under the intervention $do(X = x)$. Denote the postintervention distribution of Y by the usual expression $P(y|do(x))$. If we are interested in a set λ of parameters of that distribution, then our task is to estimate $\lambda[P(y|do(x)]$ from the available data. However, the data available is typically generated under a different set of conditions: X was not held constant but instead was allowed to vary with whatever economical pressures and expectations prompted decision makers to set X in the past. Denoting the process that generated data in the past by M and the probability distribution associated with M by $P_M(v)$, we ask whether $\lambda[P_M(y|do(x)]$ can be estimated consistently from samples drawn from $P_M(v)$, given our background knowledge T (connoting "theory") about M. This is essentially the problem of identification that we have analyzed in this and previous chapters, with one important difference; we now ask whether $\lambda[P(y|do(x)]$ can be identified from the conditional distribution $P(y|x)$ alone, instead of from the entire joint distribution $P(v)$. When identification holds under this restricted condition, X is said to be *exogenous* relative to (Y, λ, T).

We may state this formally as follows.

Definition 5.4.4 (Exogeneity)
Let X and Y be two sets of variables, and let λ be any set of parameters of the postintervention probability $P(y|do(x))$. We say that X is exogenous relative to (Y, λ, T) if λ is identifiable from the conditional distribution $P(y|x)$, that is, if

$$P_{M_1}(y|x) = P_{M_2}(y|x) \implies \lambda[P_{M_1}(y|do(x))] = \lambda[P_{M_2}(y|do(x))] \qquad (5.28)$$

for any two models, M_1 and M_2, satisfying theory T.

In the special case where λ constitutes a complete specification of the postintervention probabilities, (5.28) reduces to the implication

$$P_{M_1}(y|x) = P_{M_2}(y|x) \implies P_{M_1}(y|do(x)) = P_{M_2}(y|do(x)). \qquad (5.29)$$

If we further assume that, for every $P(y|x)$, our theory T does not a priori exclude some model M_2 satisfying $P_{M_2}(y|do(x)) = P_{M_2}(y|x)$,[20] then (5.29) reduces to the equality

$$P(y|do(x)) = P(y|x), \qquad (5.30)$$

a condition we recognize as "no confounding" (see Sections 3.3 and 6.2). Equation (5.30) follows (from (5.29)) because (5.29) must hold for all M_1 in T. Note that, since the theory T is not mentioned explicitly, (5.30) can be applied to any individual model M and can be taken as yet another definition of exogeneity—albeit a stronger one than (5.28).

The motivation for insisting that λ be identifiable from the conditional distribution $P(y|x)$ alone, even though the marginal distribution $P(x)$ is available, lies in its ramification for the process of estimation. As stated in (5.30), discovering that X is

[20] For example, if T stands for all models possessing the same graph structure, then such M_2 is not a priori excluded.

exogenous permits us to predict the effect of interventions (in X) directly from passive observations, without even adjusting for confounding factors. Our analyses in Sections 3.3 and 5.3 further provide a graphical test of exogeneity: X is exogenous for Y if there is no unblocked back-door path from X to Y (Theorem 5.3.2). This test supplements the declarative definition of (5.30) with a procedural definition and thus completes the formalization of exogeneity. That the invariance properties usually attributable to superexogeneity are discernible from the topology of the causal diagram should come as no surprise, considering that each causal diagram represents a structural model and that each structural model already embodies the invariance assumptions necessary for policy predictions (see Definition 5.4.1).

Leamer (1985) defined X to be exogenous if $P(y|x)$ remains invariant to changes in the "process that generates" X. This definition coincides[21] with (5.30) because $P(y|do(x))$ is governed by a structural model in which the equations determining X are wiped out; thus, $P(y|x)$ must be insensitive to the nature of those equations. In contrast, Engle et al. (1983) defined exogeneity (i.e., their superexogeneity) in terms of changes in the "marginal density" of X; as usual, the transition from process language to statistical terminology leads to ambiguities. According to Engle et al. (1983, p. 284), exogeneity requires that all the parameters of the conditional distribution $P(y|x)$ be "invariant for any change in the distribution of the conditioning variables"[22] (i.e. $P(x)$). This requirement of constancy under *any* change in $P(x)$ is too strong—changing conditions or new observations can easily alter both $P(x)$ and $P(y|x)$ even when X is perfectly exogenous. (To illustrate, consider a change that turns a randomized experiment, where X is indisputably exogenous, into a nonrandomized experiment; we should not insist on $P(y|x)$ remaining invariant under such change.) The class of changes considered must be restricted to local modification of the mechanisms (or equations) that determine X, as stated by Leamer, and this restriction must be incorporated into any definition of exogeneity. In order to make this restriction precise, however, the vocabulary of SEMs must be invoked as in the definition of $P(y|do(x))$; the vocabulary of marginal and conditional densities is far too coarse to properly define the changes against which $P(y|x)$ ought to remain invariant.

We are now ready to define a more general notion of exogeneity, one that includes "weak" and "super" exogeneities under the same umbrella.[23] Toward that end, we remove from Definition 5.4.4 the restriction that λ must represent features of the postintervention distribution. Instead, we allow λ to represent *any* feature of the underlying model M, including structural features such as path coefficients, causal effects, and counterfactuals, and including statistical features (which could, of course, be ascertained from the joint distribution alone). With this generalization, we also obtain a simpler definition of exogeneity.

Definition 5.4.5 (General Exogeneity)
Let X and Y be two sets of variables, and let λ be any set of parameters defined on a

[21] Provided that changes are confined to modification of functions without changing the set of arguments (i.e. parents) in each function.

[22] This requirement is repeated verbatim in Darnell (1994, p. 131) and Maddala (1992, p. 192).

[23] We leave out discussion of "strong" exogeneity, which is a slightly more involved version of weak exogeneity applicable to time-series analysis.

structural model M in a theory T. We say that X is exogenous relative to (Y, λ, T) if λ is identifiable from the conditional distribution $P(y|x)$, that is, if

$$P_{M_1}(y|x) = P_{M_2}(y|x) \implies \lambda(M_1) = \lambda(M_2) \tag{5.31}$$

for any two models, M_1 and M_2, satisfying theory T.

When λ consists of structural parameters, such as path coefficients or causal effects, (5.31) expresses invariance to a variety of interventions, not merely $do(X = x)$. Although the interventions themselves are not mentioned explicitly in (5.31), the equality $\lambda(M_1) = \lambda(M_2)$ reflects such interventions through the structural character of λ. In particular, if λ stands for the values of the causal effect function $P(y|do(x))$ at selected points of x and y, then (5.31) reduces to the implication

$$P_{M_1}(y|x) = P_{M_2}(y|x) \implies P_{M_1}(y|do(x)) = P_{M_2}(y|do(x)), \tag{5.32}$$

which is identical to (5.29). Hence the causal properties of exogeneity follow.

When λ consists of strictly statistical parameters—such as means, modes, regression coefficients, or other distributional features—the structural features of M do not enter into consideration; we have $\lambda(M) = \lambda(P_M)$ and so (5.31) reduces to

$$P_1(y|x) = P_2(y|x) \Longrightarrow \lambda(P_1) = \lambda(P_2) \tag{5.33}$$

for any two probability distributions $P_1(x, y)$ and $P_2(x, y)$ that are consistent with T. We have thus obtained a statistical notion of exogeneity that permits us to ignore the marginal $P(x)$ in the estimation of λ and that we may call "weak exogeneity".[24]

Finally, if λ consists of causal effects among variables in Y (excluding X), we obtain a generalized definition of *instrumental variables*. For example, if our interest lies in the causal effect $\lambda = P(w|do(z))$, where W and Z are two sets of variables in Y, then the exogeneity of X relative to this parameter ensures the identification of $P(w|do(z))$ from the conditional probability $P(z, w|x)$. This is indeed the role of an instrumental variable—to assist in the identification of causal effects not involving the instrument. (See Figure 5.9, with $Z,\ X,\ Y$ representing $X,\ Z,\ W$, respectively.)

A word of caution regarding the language used in most textbooks: exogeneity is frequently defined by asking whether parameters "enter" into the expressions of the conditional or the marginal density. For example, Maddala (1992, p. 392) defined weak exogeneity as the requirement that the marginal distribution $P(x)$ "does not involve" λ. Such definitions are not unambiguous, because the question of whether a parameter "enters" a density or whether a density "involves" a parameter are syntax-dependent; different algebraic representations may make certain parameters explicit or obscure. For example, if X and Y are dichotomous, then the marginal probability $P(x)$ certainly "involves" parameters such as

$$\lambda_1 = P(x_0, y_0) + P(x_0, y_1) \text{ and } \lambda_2 = P(x_0, y_0),$$

[24] Engle et al. (1983) further imposed a requirement called "variation-free," which is satisfied by default when dealing with genuinely structural models M in which mechanisms do not constrain one another.

as well as their ratio:

$$\lambda = \lambda_2/\lambda_1.$$

Therefore, writing $P(x_0) = \lambda_2/\lambda$ whos that both λ and λ_2 are involved in the marginal probability $P(x_0)$, and one may be tempted to conclude that X is not exogenous relative to λ. Yet X *is* in fact exogenous relative to λ, because the ratio $\lambda = \lambda_2/\lambda_1$ is none other than $P(y_0|x_0)$; hence it is determined uniquely by $P(y_0|x_0)$ as required by (5.33).[25]

The advantage of the definition given in (5.31) is that it depends not on the syntactic representation of the density function but rather on its semantical content alone. Parameters are treated as quantities *computed from* a model, and not as mathematical symbols that *describe* a model. Consequently, the definition applies to both statistical and structural parameters and, in fact, to any quantity λ that can be computed from a structural model M, regardless of whether it serves (or may serve) in the description of the marginal or conditional densities.

The Mystical Error Term Revisited

Historically, the definition of exogeneity that has evoked most controversy is the one expressed in terms of correlation between variables and errors. It reads as follows.

Definition 5.4.6 (Error-Based Exogeneity)
As variable X is exogenous $($relative to $\lambda = P(y|do(x)))$ if X is independent of all errors that influence Y, except those mediated by X.

This definition, which Hendry and Morgan (1995) trace to Orcutt (1952), became standard in the econometric literature between 1950 and 1970 (e.g. Christ 1966, p. 156; Dhrymes 1970, p. 169) and still serves to guide the thoughts of most econometricians (as in the selection of instrumental variables; Bowden and Turkington 1984). However, it came under criticism in the early 1980s when the distinction between structural errors (equation (5.25)) and regression errors became obscured (Richard 1980). (Regression errors, by definition, are orthogonal to the regressors.) The Cowles Commission logic of structural equations (see Section 5.1) has not reached full mathematical maturity and—by denying notational distinction between structural and regressional parameters—has left all notions based on error terms suspect of ambiguity. The prospect of establishing an entirely new foundation of exogeneity—seemingly free of theoretical terms such as "errors" and "structure" (Engle et al. 1983)—has further dissuaded economists from tidying up the Cowles Commission logic, and criticism of the error-based definition of exogeneity has become increasingly fashionable. For example, Hendry and Morgan (1995) wrote that "the concept of exogeneity rapidly evolved into a loose notion as a property of an observable variable being uncorrelated with an unobserved error," and Imbens (1997) readily agreed that this notion "is inadequate."[26]

[25] Engle et al. (1983, p. 281) and Hendry (1995, pp. 162–3) attempted to overcome this ambiguity by using "reparameterization"—an unnecessary complication.

[26] Imbens prefers definitions in terms of experimental metaphors such as "random assignment assumption," fearing, perhaps, that "[t]ypically the researcher does not have a firm idea what these disturbances really represent" (Angrist et al. p. 446).

These critics are hardly justified if we consider the precision and clarity with which structural errors can be defined when using the proper notation (e.g. (5.25)). When applied to structural errors, the standard error-based criterion of exogeneity coincides formally with that of (5.30), as can be verified using the back-door test of Theorem 5.3.2 (with $Z = \emptyset$). Consequently, the standard definition conveys the same information as that embodied in more complicated and less communicable definitions of exogeneity. I am therefore convinced that the standard definition will eventually regain the acceptance and respectability that it has always deserved.

Relationships between graphical and counterfactual definitions of exogeneity and instrumental variables will be discussed in Chapter 7 (Section 7.4.5).