CHAPTER FIVE

Causality and Structural Models in Social Science and Economics

Do two men travel together
unless they have agreed?
Amos 3:3

Preface

Structural equation modeling (SEM) has dominated causal analysis in economics and the social sciences since the 1950s, yet the prevailing interpretation of SEM differs substantially from the one intended by its originators and also from the one expounded in this book. Instead of carriers of substantive causal information, structural equations are often interpreted as carriers of probabilistic information; economists view them as convenient representations of density functions, and social scientists see them as summaries of covariance matrices. The result has been that many SEM researchers have difficulty articulating the causal content of SEM, and the most distinctive capabilities of SEM are currently ill understood and underutilized.

This chapter is written with the ambitious goal of reinstating the causal interpretation of SEM. We shall demonstrate how developments in the areas of graphical models and the logic of intervention can alleviate the current difficulties and thus revitalize structural equations as the primary language of causal modeling. Toward this end, we recast several of the results of Chapters 3 and 4 in parametric form (the form most familiar to SEM researchers) and demonstrate how practical and conceptual issues of model testing and parameter identification can be illuminated through graphical methods. We then move back to nonparametric analysis, from which an operational semantics will evolve that offers a coherent interpretation of what structural equations are all about (Section 5.4). In particular, we will provide answers to the following fundamental questions: What do structural equations claim about the world? What portion of those claims is testable? Under what conditions can we estimate structural parameters through regression analysis?

In Section 5.1 we survey the history of SEM and suggest an explanation for the current erosion of its causal interpretation. The testable implications of structural models are explicated in Section 5.2. For recursive models (herein termed *Markovian*), we find that the statistical content of a structural model can be fully characterized by a set of zero partial correlations that are entailed by the model. These zero partial correlations can be read off the graph using the *d-separation* criterion, which in linear models applies to graphs with cycles and correlated errors as well (Section 5.2). The application of this criterion to model testing is discussed in Section 5.2.2, which advocates local over global testing

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

strategies. Section 5.2.3 provides simple graphical tests of model equivalence and thus clarifies the *nontestable* part of structural models.

In Section 5.3 we deal with the issue of determining the identifiability of structural parameters prior to gathering any data. In Section 5.3.1, simple graphical tests of identifiability are developed for linear Markovian and semi-Markovian models (i.e., acyclic diagrams with correlated errors). These tests result in a simple procedure for determining when a path coefficient can be equated to a regression coefficient and, more generally, when structural parameters can be estimated through regression analysis. Section 5.3.2 discusses the connection between parameter identification in linear models and causal effect identification in nonparametric models, and Section 5.3.3 offers the latter as a semantical basis for the former.

Finally, in Section 5.4 we discuss the logical foundations of SEM and resolve a number of difficulties that were kept dormant in the past. These include operational definitions for structural equations, structural parameters, error terms, and total and direct effects, as well as a formal definition of exogeneity in econometrics.

5.1 INTRODUCTION

5.1.1 Causality in Search of a Language

The word *cause* is not in the vocabulary of standard probability theory. It is an embarrassing yet inescapable fact that probability theory, the official mathematical language of many empirical sciences, does not permit us to express sentences such as "Mud does not cause rain"; all we can say is that the two events are mutually correlated, or dependent – meaning that if we find one, we can expect to encounter the other. Scientists seeking causal explanations for complex phenomena or rationales for policy decisions must therefore supplement the language of probability with a vocabulary for causality, one in which the symbolic representation for the causal relationship "Mud does not cause rain" is distinct from the symbolic representation for "Mud is independent of rain." Oddly, such distinctions have yet to be incorporated into standard scientific analysis.[1]

Two languages for causality have been proposed: path analysis or structural equation modeling (SEM) (Wright 1921; Haavelmo 1943) and the Neyman–Rubin potential-outcome model (Neyman 1923; Rubin 1974). The former has been adopted by economists and social scientists (Goldberger 1972; Duncan 1975), while a group of statisticians champion the latter (Rubin 1974; Holland 1988; Rosenbaum 2002). These two languages are mathematically equivalent (see Chapter 7, Section 7.4.4), yet neither has become standard in causal modeling – the structural equation framework because it has been greatly misused and inadequately formalized (Freedman 1987) and the potential-outcome framework because it has been only partially formalized and (more significantly) because it rests on an esoteric and seemingly metaphysical vocabulary of randomized experiments and counterfactual variables that bears no apparent relation to ordinary understanding of cause–effect processes in nonexperimental settings (see Section 3.6.3).

[1] A summary of attempts by philosophers to reduce causality to probabilities is given in Chapter 7 (Section 7.5).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

Currently, potential-outcome models are understood by few and used by even fewer. Structural equation models are used by many, but their causal interpretation is generally questioned or avoided, even by their leading practitioners. In Chapters 3 and 4 we described how structural equation models, in nonparametric form, can provide the semantic basis for theories of interventions. In Sections 1.4 and 3.6.3 we outlined how these models provide the semantical basis for a theory of counterfactuals as well. It is somewhat embarrassing that these distinctive features are hardly recognized and rarely utilized in the modern SEM literature. The current dominating philosophy treats SEM as just a convenient way to encode density functions (in economics) or covariance information (in social science). Ironically, we are witnessing one of the most bizarre circles in the history of science: causality in search of a language and, simultaneously, speakers of that language in search of its meaning.

The purpose of this chapter is to formulate the causal interpretation and outline the proper use of structural equation models, thereby reinstating confidence in SEM as the primary formal language for causal analysis in the social and behavioral sciences. First, however, we present a brief analysis of the current crisis in SEM research in light of its historical development.

5.1.2 SEM: How Its Meaning Became Obscured

Structural equation modeling was developed by geneticists (Wright 1921) and economists (Haavelmo 1943; Koopmans 1950, 1953) so that qualitative cause–effect information could be combined with statistical data to provide quantitative assessment of cause–effect relationships among variables of interest. Thus, to the often asked question, "Under what conditions can we give causal interpretation to structural coefficients?" Wright and Haavelmo would have answered, "Always!" According to the founding fathers of SEM, the conditions that make the equation $y = \beta x + \varepsilon$ *structural* is precisely the claim that the causal connection between X and Y is β and nothing about the statistical relationship between x and ε can ever change this interpretation of β. Amazingly, this basic understanding of SEM has all but disappeared from the literature, leaving modern econometricians and social scientists in a quandary over β.

Most SEM researchers today are of the opinion that extra ingredients are necessary for structural equations to qualify as carriers of causal claims. Among social scientists, James, Mulaik, and Brett (1982, p. 45), for example, stated that a condition called *self-containment* is necessary for consecrating the equation $y = \beta x + \varepsilon$ with causal status, where self-containment stands for $\text{cov}(x, \varepsilon) = 0$. According to James et al. (1982), if self-containment does not hold, then "neither the equation nor the functional relation represents a causal relation." Bollen (1989, p. 44) reiterated the necessity of self-containment (under the rubric *isolation* or *pseudo-isolation*) – contrary to the understanding that structural equations attain their causal interpretation prior to, and independently of, any statistical relationships among their constituents. Since the early 1980s, it has become exceedingly rare to find an open endorsement of the original SEM logic: that β defines the sensitivity of $E(Y)$ to experimental manipulations (or counterfactual variations) of X; that ε is defined in terms of β, not the other way around; and that the orthogonality condition $\text{cov}(x, \varepsilon) = 0$ is neither necessary nor sufficient for the causal interpretation

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

of β (see Sections 3.6.2 and 5.4.1).[2] It is therefore not surprising that many SEM textbooks have given up on causal interpretation altogether: "We often see the terms cause, effect, and causal modeling used in the research literature. We do not endorse this practice and therefore do not use these terms here" (Schumaker and Lomax 1996, p. 90).

Econometricians have just as much difficulty with the causal reading of structural parameters. Leamer (1985, p. 258) observed, "It is my surprising conclusion that economists know very well what they mean when they use the words 'exogenous,' 'structural,' and 'causal,' yet no textbook author has written adequate definitions." There has been little change since Leamer made these observations. Econometric textbooks invariably devote most of their analysis to estimating structural parameters, but they rarely discuss the role of these parameters in policy evaluation. The few books that deal with policy analysis (e.g., Goldberger 1991; Intriligator et al. 1996, p. 28) assume that policy variables satisfy the orthogonality condition by their very nature, thus rendering structural information superfluous. Hendry (1995, p. 62), for instance, explicitly tied the interpretation of β to the orthogonality condition, stating as follows:

> the status of β may be unclear until the conditions needed to estimate the postulated model are specified. For example, in the model:
>
> $$y_t = z_t\beta + u_t \quad \text{where} \quad u_t \sim \text{IN}[0, \sigma_u^2],$$
>
> until the relationship between z_t and u_t is specified the meaning of β is uncertain since $E[z_t\, u_t]$ could be either zero or nonzero on the information provided.

LeRoy (1995, p. 211) goes even further: "It is a commonplace of elementary instruction in economics that endogenous variables are not generally causally ordered, implying that the question 'What is the effect of y_1 on y_2' where y_1 and y_2 are endogenous variables is generally meaningless." According to LeRoy, causal relationships cannot be attributed to any variable whose causes have separate influence on the effect variable, a position that denies any causal reading to most of the structural parameters that economists and social scientists labor to estimate.

Cartwright (1995b, p. 49), a renowned philosopher of science, addresses these difficulties by initiating a renewed attack on the tormenting question, "*Why* can we assume that we can read off causes, including causal order, from the parameters in equations whose exogenous variables are uncorrelated?" Cartwright, like SEM's founders, recognizes that causes cannot be derived from statistical or functional relationships alone and that causal assumptions are prerequisite for validating any causal conclusion. Unlike Wright and Haavelmo, however, she launches an all-out search for the assumptions that would endow the parameter β in the regression equation $y = \beta x + \varepsilon$ with a legitimate causal meaning and endeavors to prove that the assumptions she proposes are indeed sufficient. What is revealing in Cartwright's analysis is that she does not consider the answer Haavelmo would have provided – namely, that the assumptions needed for drawing

[2] In fact, this condition is not necessary even for the *identification* of β, once β is interpreted (see the identification of α in Figures 5.7 and 5.9).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

causal conclusions from parameters are communicated to us by the scientist who declared the equation "structural"; they are already encoded in the *syntax* of the equations and can be read off the associated graph as easily as a shopping list;[3] they need not be searched for elsewhere, nor do they require new proofs of sufficiency. Again, Haavelmo's answer applies to models of any size and shape, including models with correlated exogenous variables.

These examples bespeak an alarming tendency among economists and social scientists to view a structural equation as an algebraic object that carries functional and statistical assumptions but is void of causal content. This statement from one leading social scientist is typical: "It would be very healthy if more researchers abandoned thinking of and using terms such as cause and effect" (Muthen 1987, p. 180). Perhaps the boldest expression of this tendency was voiced by Holland (1995, p. 54): "I am speaking, of course, about the equation: $\{y = a + bx + \varepsilon\}$. What does it mean? The only meaning I have ever determined for such an equation is that it is a shorthand way of describing the conditional distribution of $\{y\}$ given $\{x\}$."[4]

The founders of SEM had an entirely different conception of structures and models. Wright (1923, p. 240) declared that "prior knowledge of the causal relations is assumed as prerequisite" in the theory of path coefficients, and Haavelmo (1943) explicitly interpreted each structural equation as a statement about a hypothetical controlled experiment. Likewise, Marschak (1950), Koopmans (1953), and Simon (1953) stated that the purpose of postulating a structure behind the probability distribution is to cope with the hypothetical changes that can be brought about by policy. One wonders, therefore, what has happened to SEM over the past 50 years, and why the basic (and still valid) teachings of Wright, Haavelmo, Marschak, Koopmans, and Simon have been forgotten.

Some economists attribute the decline in the understanding of structural equations to Lucas's (1976) critique, according to which economic agents anticipating policy interventions would tend to act contrary to SEM's predictions, which often ignore such anticipations. However, since this critique merely shifts the model's invariants and the burden of structural modeling – from the behavioral level to a deeper level that involves agents' motivations and expectations – it does not exonerate economists from defining and representing the causal content of structural equations at some level of discourse.

I believe that the causal content of SEM has gradually escaped the consciousness of SEM practitioners mainly for the following reasons.

---

[3] These assumptions are explicated and operationalized in Section 5.4. Briefly, if G is the graph associated with a causal model that renders a certain parameter identifiable, then two assumptions are sufficient for authenticating the causal reading of that parameter: (1) every missing arrow, say between X and Y, represents the assumption that X has no effect on Y once we intervene and hold the parents of Y fixed; and (2) every missing bidirected arc $X \blacktriangleleft - - \blacktriangleright Y$ represents the assumption that all omitted factors that affect Y are uncorrected with those that affect X. Each of these assumptions is *testable* in experimental settings, where interventions are feasible (Section 5.4.1).

[4] All but forgotten, the structural interpretation of the equation (Haavelmo 1943) says nothing whatsoever about the conditional distribution of $\{y\}$ given $\{x\}$. Paraphrased in our vocabulary, it reads: "In an ideal experiment where we control X to x and any other set Z of variables (not containing X or Y) to z, Y will attain a value y given by $a + bx + \varepsilon$, where ε is a random variable that is (pointwise) independent of the settings x and z" (see Section 5.4.1). This statement implies that $E[Y \mid do(x), do(z)] = a + bx + c$ but says nothing about $E(Y \mid X = x)$.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

1. SEM practitioners have sought to gain respectability for SEM by keeping causal assumptions implicit, since statisticians, the arbiters of respectability, abhor assumptions that are not directly testable.
2. The algebraic language that has dominated SEM lacks the notational facility needed to make causal assumptions, as distinct from statistical assumptions, explicit. By failing to equip causal relations with precise mathematical notation, the founding fathers in fact committed the causal foundations of SEM to oblivion. Their disciples today are seeking foundational answers elsewhere.

Let me elaborate on the latter point. The founders of SEM understood quite well that, in structural models, the equality sign conveys the asymmetrical relation "is determined by" and hence behaves more like an assignment symbol (:=) in programming languages than like an algebraic equality. However, perhaps for reasons of mathematical purity, they refrained from introducing a symbol to represent the asymmetry. According to Epstein (1987), in the 1940s Wright gave a seminar on path diagrams to the Cowles Commission (the breeding ground for SEM), but neither side saw particular merit in the other's methods. Why? After all, a diagram is nothing but a set of nonparametric structural equations in which, to avoid confusion, the equality signs are replaced with arrows.

My explanation is that the early econometricians were extremely careful mathematicians who thought they could keep the mathematics in purely equational–statistical form and just reason about structure in their heads. Indeed, they managed to do so surprisingly well, because they were truly remarkable individuals who *could* do it in their heads. The consequences surfaced in the early 1980s, when their disciples began to mistake the equality sign for an algebraic equality. The upshot was that suddenly the "so-called disturbance terms" did not make any sense at all (Richard 1980, p. 3). We are living with the sad end to this tale. By failing to express their insights in mathematical notation, the founders of SEM brought about the current difficulties surrounding the interpretation of structural equations, as summarized by Holland's "What does it mean?"

5.1.3 Graphs as a Mathematical Language

Recent developments in graphical methods promise to bring causality back into the mainstream of scientific modeling and analysis. These developments involve an improved understanding of the relationships between graphs and probabilities, on the one hand, and graphs and causality, on the other. But the crucial change has been the emergence of graphs as a mathematical language. This mathematical language is not simply a heuristic mnemonic device for displaying algebraic relationships, as in the writings of Blalock (1962) and Duncan (1975). Rather, graphs provide a fundamental notational system for concepts and relationships that are not easily expressed in the standard mathematical languages of algebraic equations and probability calculus. Moreover, graphical methods now provide a powerful symbolic machinery for deriving the consequences of causal assumptions when such assumptions are combined with statistical data.

A concrete example that illustrates the power of the graphical language – and that will set the stage for the discussions in Sections 5.2 and 5.3 – is Simpson's paradox, discussed

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

in Section 3.3 and further analyzed in Section 6.1. This paradox concerns the reversal of an association between two variables (e.g., gender and admission to school) that occurs when we partition a population into finer groups, (e.g., departments). Simpson's reversal has been the topic of much statistical research since its discovery in 1899. This research has focused on conditions for escaping the reversal instead of addressing the practical questions posed by the reversal: "Which association is more valid, before or after partitioning?" In linear analysis, the problem surfaces through the choice of regressors – for example, determining whether a variate Z can be added to a regression equation without biasing the result. Such an addition may easily reverse the sign of the coefficients of the other regressors, a phenomenon known as "suppressor effect" (Darlington 1990).

Despite a century of analysis, questions of regressor selection or adjustment for covariates continue to be decided informally, case by case, with the decision resting on folklore and intuition rather than on hard mathematics. The standard statistical literature is remarkably silent on this issue. Aside from noting that one should not adjust for a covariate that is affected by the putative cause (X),[5] the literature provides no guidelines as to what covariates might be admissible for adjustment and what assumptions would be needed for making such a determination formally. The reason for this silence is clear: the solution to Simpson's paradox and the covariate selection problem (as we have seen in Sections 3.3.1 and 4.5.3) rests on causal assumptions, and such assumptions cannot be expressed formally in the standard language of statistics.[6]

In contrast, formulating the covariate selection problem in the language of graphs immediately yields a general solution that is both natural and formal. The investigator expresses causal knowledge (or assumptions) in the familiar qualitative terminology of path diagrams, and once the diagram is completed, a simple procedure decides whether a proposed adjustment (or regression) is appropriate relative to the quantity under evaluation. This procedure, which we called the *back-door criterion* in Definition 3.3.1, was applicable when the quantity of interest is the total effect of X on Y. If instead the direct effect is to be evaluated, then the graphical criterion of Theorem 4.5.3 is applicable. A modified criterion for identifying direct effects (i.e., a path coefficient) in linear models will be given in Theorem 5.3.1.

This example is not an isolated instance of graphical methods affording clarity and understanding. In fact, the conceptual basis for SEM achieves a new level of precision through graphs. What makes a set of equations "structural," what assumptions are expressed by the authors of such equations, what the testable implications of those assumptions are, and what policy claims a given set of structural equations advertises are some of the questions that receive simple and mathematically precise answers via graphical methods. These issues, shunned even by modern SEM writers (Heckman and Vytlacil 2007; see Section 11.5.4), will be discussed in the following sections.

[5] This advice, which rests on the causal relationship "not affected by," is (to the best of my knowledge) the *only* causal notion that has found a place in statistics textbooks. The advice is neither necessary nor sufficient, as readers can verify from the discussion in Chapter 3.

[6] Simpson's reversal, as well as the supressor effect, are paradoxical only when we attach a causal reading to the associations involved; see Section 6.1.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

5.2 GRAPHS AND MODEL TESTING

In 1919, Wright developed his "method of path coefficients," which allows researchers to compute the magnitudes of cause–effect coefficients from correlation measurements provided the path diagram represents correctly the causal processes underlying the data. Wright's method consists of writing a set of equations, one for each pair of variables (X_i, X_j), and equating the (standardized) correlation coefficient ρ_{ij} with a sum of products of path coefficients and residual correlations along the various paths connecting X_i and X_j. One can then attempt to solve these equations for the path coefficients in terms of the observed correlations. Whenever the resulting equations give a unique solution to some path coefficient p_{mn} that is independent of the (unobserved) residual correlations, that coefficient is said to be *identifiable*. If every set of correlation coefficients ρ_{ij} is compatible with some choice of path coefficients, then the model is said to be *untestable* or *unfalsifiable* (also called *saturated, just identified,* etc.), because it is capable of perfectly fitting any data whatsoever.

Whereas Wright's method is partly graphical and partly algebraic, the theory of directed graphs permits us to analyze questions of testability and identifiability in purely graphical terms, prior to data collection, and it also enables us to extend these analyses from linear to nonlinear or nonparametric models. This section deals with issues of testability in linear and nonparametric models.

5.2.1 The Testable Implications of Structural Models

When we hypothesize a model of the data-generating process, that model often imposes restrictions on the statistics of the data collected. In observational studies, these restrictions provide the only view under which the hypothesized model can be tested or falsified. In many cases, such restrictions can be expressed in the form of zero partial correlations; more significantly, the restrictions are implied by the structure of the path diagram alone, independent of the numerical values of the parameters, as revealed by the *d*-separation criterion.

Preliminary Notation

Before addressing the testable implication of structural models, let us first review some definitions from Section 1.4 and relate them to the standard notation used in the SEM literature.

The graphs we discuss in this chapter represent sets of structural equations of the form

$$x_i = f_i(pa_i, \varepsilon_i), \quad i = 1, \ldots, n, \tag{5.1}$$

where pa_i (connoting *parents*) stands for (values of) the set of variables judged to be immediate causes of X_i and where the ε_i represent errors due to omitted factors. Equation (5.1) is a nonlinear, nonparametric generalization of the standard linear equations

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + \varepsilon_i, \quad i = 1, \ldots, n, \tag{5.2}$$

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

in which pa_i correspond to those variables on the r.h.s. of (5.2) that have nonzero coefficients. A set of equations in the form of (5.1) will be called a *causal model* if each equation represents the process by which the value (not merely the probability) of variable X_i is selected. The graph G obtained by drawing an arrow from every member of pa_i to X_i will be called a *causal diagram*. In addition to full arrows, a causal diagram should contain a bidirected (i.e., double-arrowed) arc between any pair of variables whose corresponding errors are dependent.

It is important to emphasize that causal diagrams (as well as traditional path diagrams) should be distinguished from the wide variety of graphical models in the statistical literature whose construction and interpretation rest solely on properties of the joint distribution (Kiiveri et al. 1984; Edwards 2000; Cowell et al. 1999; Whittaker 1990; Cox and Wermuth 1996; Lauritzen 1996; Andersson et al. 1998). The missing links in those statistical models represent conditional independencies, whereas the missing links in causal diagrams represent absence of causal connections (see note 3 and Section 5.4), which may or may not imply conditional independencies in the distribution.

A causal model will be called *Markovian* if its graph contains no directed cycles and if its ε_i are mutually independent (i.e., if there are no bidirected arcs). A model is *semi-Markovian* if its graph is acyclic and if it contains dependent errors.

If the ε_i are multivariate normal (a common assumption in the SEM literature), then the X_i in (5.2) will also be multivariate normal and will be fully characterized by the correlation coefficients ρ_{ij}. A useful property of multivariate normal distributions is that the conditional variance $\sigma^2_{X|z}$, conditional covariance $\sigma_{XY|z}$, and conditional correlation coefficient $\rho_{XY|z}$ are all independent of the value z. These are known as *partial* variance, covariance, and correlation coefficient and are denoted by $\sigma_{X\cdot Z}$, $\sigma_{XY\cdot Z}$, and $\rho_{XY\cdot Z}$ (respectively), where X and Y are single variables and Z is a set of variables. Moreover, the partial correlation coefficient $\rho_{XY\cdot Z}$ is zero if and only if $(X \perp\!\!\!\perp Y \mid Z)$ holds in the distribution.

The *partial regression coefficient* is given by

$$r_{YX\cdot Z} = \rho_{YX\cdot Z}\frac{\sigma_{Y\cdot Z}}{\sigma_{X\cdot Z}};$$

it is equal to the coefficient of X in the linear regression of Y on X and Z (the order of the subscripts is essential). In other words, the coefficient of x in the regression equation

$$y = ax + b_1 z_1 + \cdots + b_k z_k$$

is given by

$$a = r_{YX\cdot Z_1 Z_2 \cdots Z_k}.$$

These coefficients can therefore be estimated by the method of least squares (Crámer 1946).

d-Separation and Partial Correlations

Markovian models (the parallel term in the SEM literature is *recursive models*;[7] Bollen 1989) satisfy the Markov property of Theorem 1.2.7; as a result, the statistical parameters

[7] The term *recursive* is ambiguous; some authors exclude correlated errors, but others do not.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

of Markovian models can be estimated by ordinary regression analysis. In particular, the *d*-separation criterion is valid in such models (here we restate Theorem 1.2.4).

Theorem 5.2.1 (Verma and Pearl 1988; Geiger et al. 1990)
If sets X and Y are d-separated by Z in a DAG G, then X is independent of Y conditional on Z in every Markovian model structured according to G. Conversely, if X and Y are not *d-separated by Z in a DAG G, then X and Y are dependent conditional on Z in* almost all *Markovian models structured according to G.*

Because conditional independence implies zero partial correlation, Theorem 5.2.1 translates into a graphical test for identifying those partial correlations that must vanish in the model.

Corollary 5.2.2
In any Markovian model structured according to a DAG G, the partial correlation $\rho_{XY \cdot Z}$ *vanishes whenever the nodes corresponding to the variables in Z d-separate node X from node Y in G, regardless of the model's parameters. Moreover, no other partial correlation would vanish for all the model's parameters.*

Unrestricted semi-Markovian models can always be emulated by Markovian models that include latent variables, with the latter accounting for all dependencies among error terms. Consequently, the *d*-separation criterion remains valid in such models if we interpret bidirected arcs as emanating from latent common parents. This may not be possible in some linear semi-Markovian models where each latent variable is restricted to influence at most two observed variables (Spirtes et al. 1996). However, it has been shown that the *d*-separation criterion remains valid in such restricted systems (Spirtes et al. 1996) and, moreover, that the validity is preserved when the network contains cycles (Spirtes et al. 1998; Koster 1999). These results are summarized in the next theorem.

Theorem 5.2.3 (*d*-Separation in General Linear Models)
For any linear model structured according to a diagram D, which may include cycles and bidirected arcs, the partial correlation $\rho_{XY \cdot Z}$ *vanishes if the nodes corresponding to the set of variables Z d-separate node X from node Y in D.* (*Each bidirected arc* $i \blacktriangleleft\!\!- - \!\!\blacktriangleright j$ *is interpreted as a latent common parent* $i \leftarrow L \rightarrow j$.)

For linear structural equation models (see (5.2)), Theorem 5.2.3 implies that those (and only those) partial correlations identified by the *d*-separation test are guaranteed to vanish independent of the model parameters α_{ik} and independent of the error variances. This suggests a simple and direct method for testing models: rather than going through the standard exercise of finding a maximum likelihood estimate for the model's parameters and scoring those estimates for fit to the data, we can directly test for each zero partial correlation implied by the free model. The advantages of using such tests were noted by Shipley (1997), who also devised implementations of these tests.

However, the question arises of whether it is feasible to test for the vast number of zero partial correlations entailed by a given model. Fortunately, these partial correlations

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

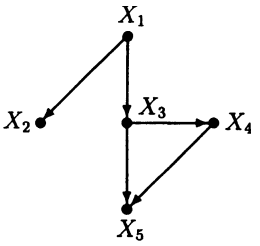

Figure 5.1 Model testable with two regressors for each missing link (equation (5.3)).

are not independent of each other; they can be derived from a relatively small number of partial correlations that constitutes a *basis* for the entire set (Pearl and Verma 1987).

Definition 5.2.4 (Basis)
Let S be a set of partial correlations. A basis *B for S is a set of zero partial correlations where* (i) *B implies* (*using the laws of probability*) *the zero of every element of S and* (ii) *no proper subset of B sustains such implication.*

An obvious choice of a basis for the zero partial correlations entailed by a DAG D is the set of equalities $B = \{\rho_{ij \cdot pa_i} = 0 \mid i > j\}$, where i ranges over all nodes in D and j ranges over all predecessors of i in any order that agrees with the arrows of D. In fact, this set of equalities reflects the "parent screening" property of Markovian models (Theorem 1.2.7), which is the source of all the probabilistic information encoded in a DAG. Testing for these equalities is therefore sufficient for testing all the statistical claims of a linear Markovian model. Moreover, when the parent sets PA_i are large, it may be possible to select a more economical basis, as shown in the next theorem.[8]

Theorem 5.2.5 (Graphical Basis)
Let (i, j) *be a pair of nonadjacent nodes in a DAG D, and let* Z_{ij} *be any set of nodes that are closer to i than j is to i and such that* Z_{ij} *d-separates i from j. The set of zero partial correlations* $B = \{\rho_{ij \cdot Z_{ij}} = 0 \mid i > j\}$, *consisting of one element per nonadjacent pair, constitutes a basis for the set of all zero partial correlations entailed by D.*

Theorem 5.2.5 states that the set of zero partial correlations corresponding to *any* separation between nonadjacent nodes in the diagram encapsulates all the statistical information conveyed by a linear Markovian model. A proof of Theorem 5.2.5 is given in Pearl and Meshkat (1999).

Examining Figure 5.1, we see that each of following two sets forms a basis for the model in the figure:

$$\begin{aligned} B_1 &= \{\rho_{32 \cdot 1} = 0,\ \ \rho_{41 \cdot 3} = 0,\ \ \rho_{42 \cdot 3} = 0,\ \ \rho_{51 \cdot 43} = 0,\ \ \rho_{52 \cdot 43} = 0\}, \\ B_2 &= \{\rho_{32 \cdot 1} = 0,\ \ \rho_{41 \cdot 3} = 0,\ \ \rho_{42 \cdot 1} = 0,\ \ \rho_{51 \cdot 3} = 0,\ \ \rho_{52 \cdot 1} = 0\}. \end{aligned} \tag{5.3}$$

The basis B_1 employs the parent set PA_i for separating i from j ($i > j$). Basis B_2, on the other hand, employs smaller separating sets and thus leads to tests that involve fewer

[8] The possibility that linear models may possess more economical bases came to my awareness during a conversation with Rod McDonald.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

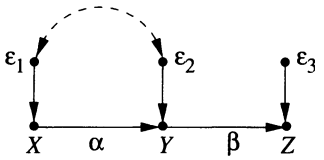

Figure 5.2 A testable model containing unidentified parameter (α).

regressors. Note that each member of a basis corresponds to a missing arrow in the DAG; therefore, the number of tests required to validate a DAG is equal to the number of missing arrows it contains. The sparser the graph, the more it constrains the covariance matrix and more tests are required to verify those constraints.

5.2.2 Testing the Testable

In linear structural equation models, the hypothesized causal relationships between variables can be expressed in the form of a directed graph annotated with coefficients, some fixed a priori (usually to zero) and some free to vary. The conventional method for testing such a model against the data involves two stages. First, the free parameters are estimated by iteratively maximizing a fitness measure such as the likelihood function. Second, the covariance matrix implied by the estimated parameters is compared to the sample covariances and a statistical test is applied to decide whether the latter could originate from the former (Bollen 1989; Chou and Bentler 1995).

There are two major weaknesses to this approach:

1. if some parameters are not identifiable, then the first phase may fail to reach stable estimates for the parameters and the investigator must simply abandon the test;
2. if the model fails to pass the data fitness test, the investigator receives very little guidance about which modeling assumptions are wrong.

For example, Figure 5.2 shows a path model in which the parameter α is not identifiable if $\text{cov}(\varepsilon_1, \varepsilon_2)$ is assumed to be unknown, which means that the maximum likelihood method may fail to find a suitable estimate for α, thus precluding the second phase of the test. Still, this model is no less testable than the one in which $\text{cov}(\varepsilon_1, \varepsilon_2) = 0$, α is identifiable, and the test can proceed. These models impose the same restrictions on the covariance matrix – namely, that the partial correlation $\rho_{XZ \cdot Y}$ should vanish (i.e., $\rho_{XZ} = \rho_{XY}\,\rho_{YZ}$) – yet the model with free $\text{cov}(\varepsilon_1, \varepsilon_2)$, by virtue of α being nonidentifiable, cannot be tested for this restriction.

Figure 5.3 illustrates the weakness associated with model diagnosis. Suppose the true data-generating model has a direct causal connection between X and W, as shown in Figure 5.3(a), while the hypothesized model (Figure 5.3(b)) has no such connection. Statistically, the two models differ in the term $\rho_{XW \cdot Z}$, which should vanish according to Figure 5.3(b) and is left free according to Figure 5.3(a). Once the nature of the discrepancy is clear, the investigator must decide whether substantive knowledge justifies alteration of the model by adding either a link or a curved arc between X and W. However, because the effect of the discrepancy will be spread over several covariance terms, global fitness tests will not be able to isolate the discrepancy easily. Even multiple fitness tests

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

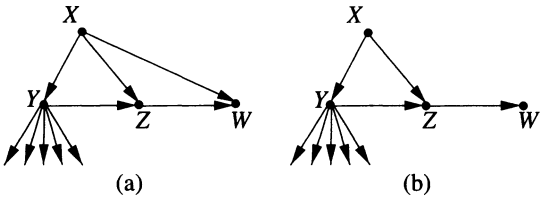

Figure 5.3 Models differing in one local test, $\rho_{XW \cdot Z} = 0$.

on various local modifications of the model (such tests are provided by LISREL) may not help much, because the results may be skewed by other discrepancies in different parts of the model, such as the subgraph rooted at Y. Thus, testing for global fitness is often of only minor use in model debugging.

An attractive alternative to global fitness testing is local fitness testing, which involves listing the restrictions implied by the model and testing them one by one. A restriction such as $\rho_{XW \cdot Z} = 0$, for example, can be tested locally without measuring Y or any of its descendants, thus keeping errors associated with those measurements from interfering with the test for $\rho_{XW \cdot Z} = 0$, which is the real source of the lack of fit. More generally, typical SEM models are often close to being "saturated," claiming but a few restrictions in the form of a few edges missing from large, otherwise unrestrictive diagrams. Local and direct tests for those restrictions are more reliable than global tests, since they involve fewer degrees of freedom and are not contaminated with irrelevant measurement errors. The missing edges approach described in Section 5.2.1 provides a systematic way of detecting and enumerating the local tests needed for testing a given model.

5.2.3 Model Equivalence

In Section 2.3 (Definition 2.3.3) we defined two structural equation models to be observationally equivalent if every probability distribution that is generated by one of the models can also be generated by the other. In standard SEM, models are assumed to be linear and data are characterized by covariance matrices. Thus, two such models are observationally indistinguishable if they are *covariance equivalent*, that is, if every covariance matrix generated by one model (through some choice of parameters) can also be generated by the other. It can be easily verified that the equivalence criterion of Theorem 1.2.8 extends to covariance equivalence.

Theorem 5.2.6

Two Markovian linear-normal models are covariance equivalent if and only if they entail the same sets of zero partial correlations. Moreover, two such models are covariance equivalent if and only if their corresponding graphs have the same sets of edges and the same sets of υ*-structures.*

The first part of Theorem 5.2.6 defines the testable implications of Markovian models. It states that, in nonmanipulative studies, Markovian structural equation models cannot be tested for any feature other than those zero partial correlations that the d-separation test reveals. It also provides a simple test for equivalence that requires, instead of checking all the d-separation conditions, merely a comparison of corresponding edges and their directionalities.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

In semi-Markovian models (DAGs with correlated errors), the *d*-separation criterion is still valid for testing independencies (see Theorem 5.2.3), but independence equivalence no longer implies observational equivalence.[9] Two models that entail the same set of zero partial correlations among the observed variables may yet impose different inequality constraints on the covariance matrix. Nevertheless, Theorems 5.2.3 and 5.2.6 still provide necessary conditions for testing equivalence.

Generating Equivalent Models

By permitting arrows to be reversed as long as no v-structures are destroyed or created, we can use Theorem 5.2.6 to generate equivalent alternatives to any Markovian model. Meek (1995) and Chickering (1995) showed that $X \rightarrow Y$ can be replaced by $X \leftarrow Y$ if and only if all parents of X are also parents of Y. They also showed that, for any two equivalent models, there is always some sequence of such edge reversals that takes one model into the other. This simple rule for edge reversal coincides with those proposed by Stelzl (1986) and Lee and Hershberger (1990).

In semi-Markovian models, the rules for generating equivalent models are more complicated. Nevertheless, Theorem 5.2.6 yields convenient graphical principles for testing the correctness of edge-replacement rules. The basic principle is that if we regard each bidirected arc $X \dashleftarrow\dashrightarrow Y$ as representing a latent common cause $X \leftarrow L \rightarrow Y$, then the "if" part of Theorem 5.2.6 remains valid; that is, any edge-replacement transformation that does not destroy or create a v-structure is allowed. Thus, for example, an edge $X \rightarrow Y$ can be replaced by a bidirected arc $X \dashleftarrow\dashrightarrow Y$ whenever X and Y have no other parents, latent or observed. Likewise, an edge $X \rightarrow Y$ can be replaced by a bidirected arc $X \dashleftarrow\dashrightarrow Y$ whenever (1) X and Y have no latent parents and (2) every parent of X or Y is a parent of both. Such replacements do not introduce new v-structures. However, since v-structures may now involve latent variables, we can tolerate the creation or destruction of some v-structures as long as this does not affect partial correlations among the observed variables. Figure 5.4(a) demonstrates that the creation of certain v-structures can be tolerated. By reversing the arrow $X \rightarrow Y$ we create two converging arrows $Z \rightarrow X \leftarrow Y$ whose tails are connected, not directly, but through a latent common cause. This is tolerated because, although the new convergence at X blocks the path (Z, X, Y), the connection between Z and Y (through the arc $Z \dashleftarrow\dashrightarrow Y$) remains unblocked and, in fact, cannot be blocked by any set of observed variables.

We can carry this principle further by generalizing the concept of v-structure. Whereas in Markovian models a v-structure is defined as two converging arrows whose tails are not connected by a link, we now define v-structure as any two converging arrowheads whose tails are "separable." By *separable* we mean that there exists a conditioning set S capable of *d*-separating the two tails. Clearly, the two tails will not be separable if they are connected by an arrow or by a bidirected arc. But a pair of nodes in a semi-Markovian model can be inseparable even when not connected by an edge (Verma and Pearl 1990). With this generalization in mind, we can state necessary conditions for edge replacement as follows.

[9] Verma and Pearl (1990) presented an example using a nonparametric model, and Richardson devised an example using linear models with correlated errors (Spirtes and Richardson 1996).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

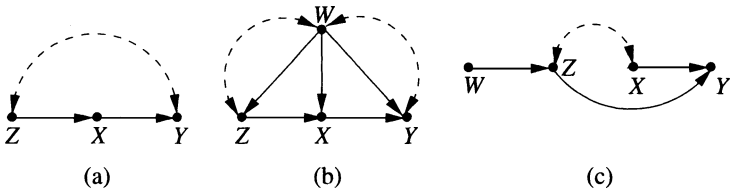

Figure 5.4 Models permitting ((a) and (b)) and forbidding (c) the reversal of $X \rightarrow Y$.

Rule 1: An arrow $X \rightarrow Y$ is interchangeable with $X \leftarrow\!\!\rightarrow Y$ only if every neighbor or parent of X is inseparable from Y. (By *neighbor* we mean a node connected (to X) through a bidirected arc.)

Rule 2: An arrow $X \rightarrow Y$ can be reversed into $X \leftarrow Y$ only if, before reversal, (i) every neighbor or parent of Y (excluding X) is inseparable from X and (ii) every neighbor or parent of X is inseparable from Y.

For example, consider the model $Z \leftarrow\!\!\rightarrow X \rightarrow Y$. The arrow $X \rightarrow Y$ cannot be replaced with a bidirected arc $X \leftarrow\!\!\rightarrow Y$ because Z (a neighbor of X) is separable from Y by the set $S = \{X\}$. Indeed, the new v-structure created at X would render Z and Y marginally independent, contrary to the original model.

As another example, consider the graph in Figure 5.4(a). Here, it is legitimate to replace $X \rightarrow Y$ with $X \leftarrow\!\!\rightarrow Y$ or with a reversed arrow $X \leftarrow Y$ because X has no neighbors and Z, the only parent of X, is inseparable from Y. The same considerations apply to Figure 5.4(b); variables Z and Y, though nonadjacent, are inseparable, because the paths going from Z to Y through W cannot be blocked.

A more complicated example, one that demonstrates that rules 1 and 2 are not sufficient to ensure the legitimacy of a transformation, is shown in Figure 5.4(c). Here, it appears that replacing $X \rightarrow Y$ with $X \leftarrow\!\!\rightarrow Y$ would be legitimate because the (latent) v-structure at X is shunted by the arrow $Z \rightarrow Y$. However, the original model shows the path from W to Y to be d-connected given Z, whereas the postreplacement model shows the same path d-separated given Z. Consequently, the partial correlation $\rho_{WY \cdot Z}$ vanishes in the postreplacement model but not in the prereplacement model. A similar disparity also occurs relative to the partial correlation $\rho_{WY \cdot ZX}$. The original model shows that the path from W to Y is blocked, given $\{Z, X\}$, but the postreplacement model shows that path to be d-connected, given $\{Z, X\}$. Consequently, the partial correlation $\rho_{WY \cdot ZX}$ vanishes in the prereplacement model but is unconstrained in the postreplacement model.[10] Evidently, it is not enough to impose rules on the parents and neighbors of X; remote ancestors (e.g., W) should be considered, too.

These rules are just a few of the implications of the d-separation criterion when applied to semi-Markovian models. A necessary and sufficient criterion for testing the d-separation equivalence of two semi-Markovian models was devised by Spirtes and Verma (1992). Spirtes and Richardson (1996) extended that criterion to include models with feedback cycles. However, we should keep in mind that, because two semi-Markovian

[10] This example was brought to my attention by Jin Tian, and a similar one by two anonymous reviewers.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

models can be zero-partial-correlation equivalent and yet not covariance equivalent, criteria based on *d*-separation can provide merely the necessary conditions for model equivalence.

The Significance of Equivalent Models

Theorem 5.2.6 is methodologically significant because it clarifies what it means to claim that structural models are "testable" (Bollen 1989, p. 78).[11] It asserts that we never test *a* model but rather a whole *class* of observationally equivalent models from which the hypothesized model cannot be distinguished by any statistical means. It asserts as well that this equivalence class can be constructed (by inspection) from the graph, which thus provides the investigator with a vivid representation of competing alternatives for consideration. Graphical representation of all models in a given equivalence class have been devised by Verma and Pearl (1990) (see Section 2.6), Spirtes et al. (1993), and Andersson et al. (1998). Richardson (1996) discusses the representation of equivalence classes of models with cycles.

Although it is true that (overidentified) structural equation models have testable implications, those implications are but a small part of what the model represents: a set of claims, assumptions, and implications. Failure to distinguish among causal assumptions, statistical implications, and policy claims has been one of the main reasons for the suspicion and confusion surrounding quantitative methods in the social sciences (Freedman 1987, p. 112; Goldberger 1992; Wermuth 1992). However, because they make the distinctions among these components vivid and crisp, graphical methods promise to make SEM more acceptable to researchers from a wide variety of disciplines.

By and large, the SEM literature has ignored the explicit analysis of equivalent models. Breckler (1990), for example, found that only one of 72 articles in the areas of social and personality psychology even acknowledged the existence of an equivalent model. The general attitude has been that the combination of data fitness and model over-identification is sufficient to confirm the hypothesized model. Recently, however, the existence of multiple equivalent models seems to have jangled the nerves of some SEM researchers. MacCallum et al. (1993, p. 198) concluded that "the phenomenon of equivalent models represents a serious problem for empirical researchers using CSM" and "a threat to the validity of interpretation of CSM results" (CSM denotes "covariance structure modeling"; this does not differ from SEM, but the term is used by some social scientists to disguise euphemistically the causal content of their models). Breckler (1990, p. 262) reckoned that "if one model is supported, so too are all of its equivalent models" and hence ventured that "the term *causal modeling* is a misnomer."

Such extremes are not justifiable. The existence of equivalent models is logically inevitable if we accept the fact that causal relations cannot be inferred from statistical data alone; as Wright (1921) stated, "prior knowledge of the causal relations is assumed as prerequisite" in SEM. But this does not make SEM useless as a tool for causal modeling.

[11] In response to an allegation that "path analysis does not derive the causal theory from the data, or test any major part of it against the data" (Freedman 1987, p. 112), Bollen (1989, p. 78) stated, "we can test and reject structural models.… Thus the assertion that these models cannot be falsified has little basis."

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

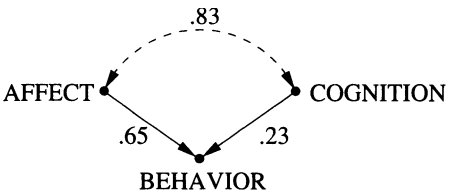

Figure 5.5 Untestable model displaying quantitative causal information derived.

The move from the qualitative causal premises represented by the structure of a path diagram (see note 3) to the quantitative causal conclusions advertised by the coefficients in the diagram is neither useless nor trivial. Consider, for example, the model depicted in Figure 5.5, which Bagozzi and Burnkrant (1979) used to illustrate problems associated with equivalent models. Although this model is saturated (i.e., just identified) and although it has (at least) 27 semi-Markovian equivalent models, finding that the influence of AFFECT on BEHAVIOR is almost three times stronger (on a standardized scale) than the influence of COGNITION on BEHAVIOR is still very illuminating – it tells us about the relative effectiveness of different behavior modification policies if some are known to influence AFFECT and others COGNITION. The significance of this quantitative analysis on policy analysis may be more dramatic when a path coefficient turns negative while the corresponding correlation coefficient measures positive. Such quantitative results may have profound impact on policy decisions, and learning that these results are logically implied by the data and the qualitative premises embedded in the diagram should make the basis for policy decisions more transparent to defend or criticize (see Section 11.5.3).

In summary, social scientists need not abandon SEM altogether; they need only abandon the notion that SEM is a method of *testing* causal models. Structural equation modeling is a method of testing a tiny fraction of the premises that make up a causal model and, in cases where that fraction is found to be compatible with the data, the method elucidates the necessary quantitative consequences of both the premises and the data. It follows, then, that users of SEM should concentrate on examining the implicit theoretical premises that enter into a model. As we will see in Section 5.4, graphical methods make these premises vivid and precise.

5.3 GRAPHS AND IDENTIFIABILITY

5.3.1 Parameter Identification in Linear Models

Consider a directed edge $X \rightarrow Y$ embedded in a path diagram G, and let α stand for the path coefficient associated with that edge. It is well known that the regression coefficient $r_{YX} = \rho_{XY}\sigma_Y/\sigma_X$ can be decomposed into the sum

$$r_{YX} = \alpha + I_{YX},$$

where I_{YX} is not a function of α, since it is computed (e.g., using Wright's rules) from other paths connecting X and Y excluding the edge $X \rightarrow Y$. (Such paths traverse both unidirected and bidirected arcs.) Thus, if we remove the edge $X \rightarrow Y$ from the path diagram and find that the resulting subgraph entails zero correlation between X and Y, then

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

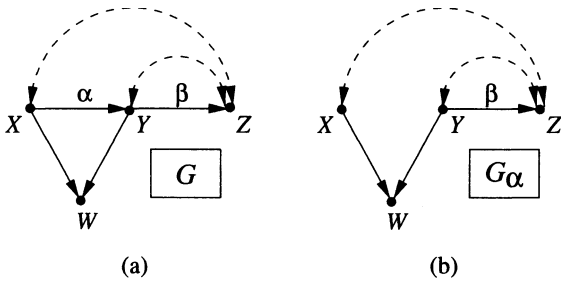

Figure 5.6 Test of whether structural parameter α can be equated with regression coefficient r_{YX}.

we know that $I_{YX} = 0$ and $\alpha = r_{YX}$; hence, α is identified. Such entailment can be established graphically by testing whether X is d-separated from Y (by the empty set $Z = \{\emptyset\}$) in the subgraph. Figure 5.6 illustrates this simple test for identification: all paths between X and Y in the subgraph G_α are blocked by converging arrows, and α can immediately be equated with r_{YX}.

We can extend this basic idea to cases where I_{YX} is not zero but can be made zero by adjusting for a set of variables $Z = \{Z_1, Z_2, \ldots, Z_k\}$ that lie on various d-connected paths between X and Y. Consider the partial regression coefficient $r_{YX \cdot Z} = \rho_{YX \cdot Z}\, \sigma_{Y \cdot Z} / \sigma_{X \cdot Z}$, which represents the residual correlation between Y and X after Z is "partialled out." If Z contains no descendant of Y, then again we can write[12]

$$r_{YX \cdot Z} = \alpha + I_{YX \cdot Z},$$

where $I_{YX \cdot Z}$ represents the partial correlation between X and Y resulting from setting α to zero, that is, the partial correlation in a model whose graph G_α lacks the edge $X \rightarrow Y$ but is otherwise identical to G. If Z d-separates X from Y in G_α, then $I_{YX \cdot Z}$ would indeed be zero in such a model, and so we can conclude that, in our original model, α is identified and is equal to $r_{YX \cdot Z}$. Moreover, since $r_{YX \cdot Z}$ is given by the coefficient of x in the regression of Y on X and Z, α can be estimated using the regression

$$y = \alpha x + \beta_1 z_1 + \cdots + \beta_k z_k + \varepsilon.$$

This result provides a simple graphical answer to the questions, alluded to in Section 5.1.3, of (i) what constitutes an adequate set of regressors and (ii) when a regression coefficient provides a consistent estimate of a path coefficient. The answers are summarized in the following theorem.[13]

Theorem 5.3.1 (Single-Door Criterion for Direct Effects)
Let G be any path diagram in which α is the path coefficient associated with link $X \rightarrow Y$, and let G_α denote the diagram that results when $X \rightarrow Y$ is deleted from G. The coefficient α is identifiable if there exists a set of variables Z such that (i) *Z contains no*

[12] This can be seen when the relation between Y and its parents, $Y = \alpha x + \Sigma_i \beta_i w_i + \varepsilon$, is substituted into the expression for $r_{YX \cdot Z}$, which yields α plus an expression $I_{YX \cdot Z}$ involving partial correlations among the variables $\{X, W_1, \ldots, W_k, Z, \varepsilon\}$. Because Y is assumed not to be an ancestor of any of these variables, their joint density is unaffected by the equation for Y; hence, $I_{YX \cdot Z}$ is independent of α.

[13] This result is presented in Pearl (1998a) and Spirtes et al. (1998).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

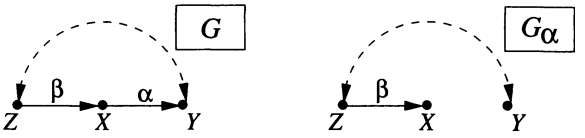

Figure 5.7 The identification of α with $r_{YX \cdot Z}$ (Theorem 5.3.1) is confirmed by G_α.

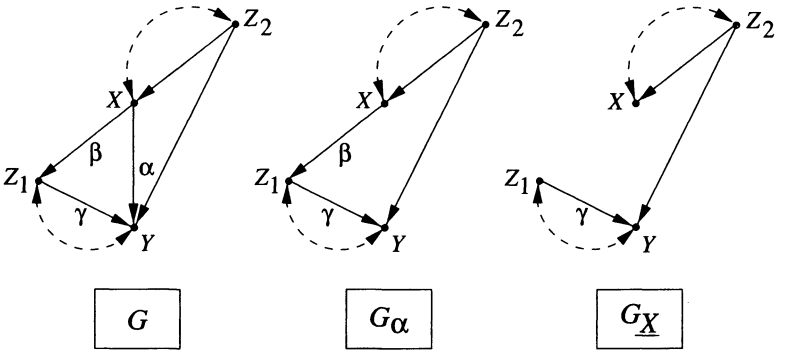

Figure 5.8 Graphical identification of the total effect of X on Y, yielding $\alpha + \beta\gamma = r_{YX \cdot Z_2}$.

descendant of Y and (ii) *Z d-separates X from Y in* G_α. *If Z satisfies these two conditions, then* α *is equal to the regression coefficient* $r_{YX \cdot Z}$. *Conversely, if Z does not satisfy these conditions, then* $r_{YX \cdot Z}$ *is not a consistent estimand of* α (*except in rare instances of measure zero*).

The use of Theorem 5.3.1 can be illustrated as follows. Consider the graphs G and G_α in Figure 5.7. The only path connecting X and Y in G_α is the one traversing Z, and since that path is d-separated (blocked) by Z, α is identifiable and is given by $\alpha = r_{YX \cdot Z}$. The coefficient β is identifiable, of course, since Z is d-separated from X in G_β (by the empty set $\emptyset$) and thus $\beta = r_{XZ}$. Note that this "single-door" test differs slightly from the back-door criterion for total effects (Definition 3.3.1); the set Z here must block *all* indirect paths from X to Y, not only back-door paths. Condition (i) is identical to both cases, because if X is a parent of Y then every descendant of Y must also be a descendant of X.

We now extend the identification of structural parameters through the identification of total effects (rather than direct effects). Consider the graph G in Figure 5.8. If we form the graph G_α by removing the link $X \rightarrow Y$, we observe that there is no set Z of nodes that d-separates all paths from X to Y. If Z contains Z_1, then the path $X \rightarrow Z_1 \dashleftarrow\dashrightarrow Y$ will be unblocked through the converging arrows at Z_1. If Z does not contain Z_1, the path $X \rightarrow Z_1 \rightarrow Y$ is unblocked. Thus we conclude that α cannot be identified using our previous method. However, suppose we are interested in the total effect of X on Y, which is given by $\alpha + \beta\gamma$. For this sum to be identified by r_{YX}, there should be no contribution to r_{YX} from paths other than those leading from X to Y. However, we see that two such paths, called *confounding* or *back-door* paths, exist in the graph – namely, $X \leftarrow Z_2 \rightarrow Y$ and $X \dashleftarrow\dashrightarrow Z_2 \rightarrow Y$. Fortunately, these paths are blocked by Z_2 and so we may conclude that adjusting for Z_2 would render $\alpha + \beta\gamma$ identifiable; thus we have

$$\alpha + \beta\gamma = r_{YX \cdot Z_2}.$$

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

This line of reasoning is captured by the back-door criterion of Definition 3.3.1, which we restate here for completeness.

Theorem 5.3.2 (Back-Door Criterion)
For any two variables X and Y in a causal diagram G, the total effect of X on Y is identifiable if there exists a set of measurements Z such that

1. *no member of Z is a descendant of X; and*
2. *Z d-separates X from Y in the subgraph $G_{\underline{X}}$ formed by deleting from G all arrows emanating from X.*

Moreover, if the two conditions are satisfied, then the total effect of X on Y is given by $r_{YX \cdot Z}$.

The two conditions of Theorem 5.3.2, as we have seen in Section 3.3.1, are also valid in nonlinear non-Gaussian models as well as in models with discrete variables. The test ensures that, after adjustment for Z, the variables X and Y are not associated through confounding paths, which means that the regression coefficient $r_{YX \cdot Z}$ is equal to the total effect. In fact, we can view Theorems 5.3.1 and 5.3.2 as special cases of a more general scheme: In order to identify any *partial effect*, as defined by a select bundle of causal paths from X to Y, we ought to find a set Z of measured variables that block all nonselected paths between X and Y. The partial effect will then equal the regression coefficient $r_{YX \cdot Z}$.

Figure 5.8 demonstrates that some total effects can be determined directly from the graphs without having to identify their individual components. Standard SEM methods (Bollen 1989; Chou and Bentler 1995) that focus on the identification and estimation of individual parameters may miss the identification and estimation of effects such as the one in Figure 5.8, which can be estimated reliably even though some of the constituents remain unidentified.

Some total effects cannot be determined directly as a unit but instead require the determination of each component separately. In Figure 5.7, for example, the effect of Z on Y ($= \alpha\beta$) does not meet the back-door criterion, yet this effect can be determined from its constituents α and β, which meet the back-door criterion individually and evaluate to

$$\beta = r_{XZ}, \qquad \alpha = r_{YX \cdot Z}.$$

There is yet a third kind of causal parameter: one that cannot be determined either directly or through its constituents but rather requires the evaluation of a broader causal effect of which it is a part. The structure shown in Figure 5.9 represents an example of this case. The parameter α cannot be identified either directly or from its constituents (it has none), yet it can be determined from $\alpha\beta$ and β, which represent the effect of Z on Y and of Z on X, respectively. These two effects can be identified directly, since there are no back-door paths from Z to either Y or X; therefore, $\alpha\beta = r_{YZ}$ and $\beta = r_{XZ}$. It follows that

$$\alpha = r_{YZ}/r_{XZ},$$

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

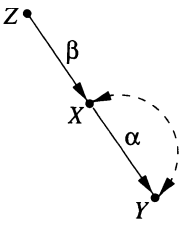

Figure 5.9 Graphical identification of α using instrumental variable Z.

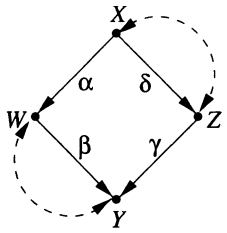

Figure 5.10 Graphical identification of α, β, and γ.

which is familiar to us as the *instrumental variable* formula (Bowden and Turkington 1984; see also Section 3.5, equation (3.46)).

The example shown in Figure 5.10 combines all three methods considered thus far. The total effect of X on Y is given by $\alpha\beta + \gamma\delta$, which is not identifiable because it does not meet the back-door criterion and is not part of another identifiable structure. However, suppose we wish to estimate β. By conditioning on Z, we block all paths going through Z and obtain $\alpha\beta = r_{YX \cdot Z}$, which is the effect of X on Y mediated by W. Because there are no back-door paths from X to W, α itself evaluates directly to $\alpha = r_{WX}$. We therefore obtain

$$\beta = r_{YX \cdot Z} / r_{WX}.$$

On the other hand, γ can be evaluated directly by conditioning on X (thus blocking all back-door paths from Z to Y through X), which gives

$$\gamma = r_{YZ \cdot X}.$$

The methods that we have been using suggest the following systematic procedure for recognizing identifiable coefficients in a graph.

1. Start by searching for identifiable causal effects among pairs of variables in the graph, using the back-door criterion and Theorem 5.3.1. These can be either direct effects, total effects, or partial effects (i.e., effects mediated by specific sets of variables).
2. For any such identified effect, collect the path coefficients involved and put them in a bucket.
3. Begin labeling the coefficients in the buckets according to the following procedure:
   (a) if a bucket is a singleton, label its coefficient I (denoting *identifiable*);
   (b) if a bucket is not a singleton but contains only a single unlabeled element, label that element I.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

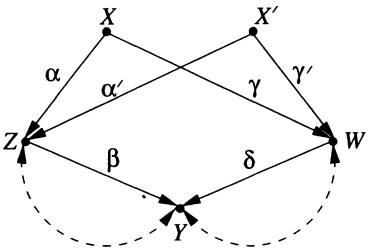

Figure 5.11 Identifying β and δ using two instrumental variables.

4. Repeat this process until no new labeling is possible.
5. List all labeled coefficients; these are identifiable.

The process just described is not complete, because our insistence on labeling coefficients one at a time may cause us to miss certain opportunities. This is shown in Figure 5.11. Starting with the pairs (X, Z), (X, W), (X', Z), and (X', W), we discover that α, γ, α', and γ' are identifiable. Going to (X, Y), we find that $\alpha\beta + \delta\gamma$ is identifiable; likewise, from (X', Y) we see that $\alpha'\beta + \gamma'\delta$ is identifiable. This does not yet enable us to label β or δ, but we can solve two equations for the unknowns β and δ as long as the determinant $\begin{vmatrix} \alpha & \gamma \\ \alpha' & \gamma' \end{vmatrix}$ is nonzero. Since we are not interested in identifiability at a point but rather in identifiability "almost everywhere" (Koopmans et al. 1950; Simon 1953), we need not compute this determinant. We merely inspect the symbolic form of the determinant's rows to make sure that the equations are nonredundant; each imposes a new constraint on the unlabeled coefficients for at least one value of the labeled coefficients.

With a facility to detect redundancies, we can increase the power of our procedure by adding the following rule:

3*. If there are k nonredundant buckets that contain at most k unlabeled coefficients, label these coefficients and continue.

Another way to increase the power of our procedure is to list not only identifiable effects but also expressions involving correlations due to bidirected arcs, in accordance with Wright's rules. Finally, one can endeavor to list effects of several variables jointly as is done in Section 4.4. However, such enrichments tend to make the procedure more complex and might compromise our main objective of providing investigators with a way to immediately recognize the identified coefficients in a given model and immediately understand those features in the model that influence the identifiability of the target quantity. We now relate these results to the identification in nonparametric models, such as those treated in Section 3.3.

5.3.2 Comparison to Nonparametric Identification

The identification results of the previous section are significantly more powerful than those obtained in Chapters 3 and 4 for nonparametric models. Nonparametric models should nevertheless be studied by parametric modelers for both practical and conceptual

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

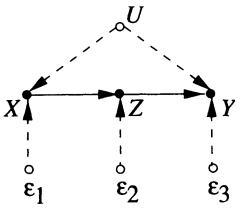

Figure 5.12 Path diagram corresponding to equations (5.4)–(5.6), where $\{X, Z, Y\}$ are observed and $\{U, \varepsilon_1, \varepsilon_2, \varepsilon_3\}$ are unobserved.

reasons. On the practical side, investigators often find it hard to defend the assumptions of linearity and normality (or other functional-distributional assumptions), especially when categorical variables are involved. Because nonparametric results are valid for nonlinear functions and for any distribution of errors, having such results allows us to gauge how sensitive standard techniques are to assumptions of linearity and normality. On the conceptual side, nonparametric models illuminate the distinctions between structural and algebraic equations. The search for nonparametric quantities analogous to path coefficients forces explication of what path coefficients really mean, why one should labor at their identification, and why structural models are not merely a convenient way of encoding covariance information.

In this section we cast the problem of nonparametric causal effect identification (Chapter 3) in the context of parameter identification in linear models.

Parametric versus Nonparametric Models: An Example

Consider the set of structural equations

$$x = f_1(u, \varepsilon_1), \tag{5.4}$$

$$z = f_2(x, \varepsilon_2), \tag{5.5}$$

$$y = f_3(z, u, \varepsilon_3), \tag{5.6}$$

where X, Z, Y are observed variables, f_1, f_2, f_3 are unknown arbitrary functions, and U, ε_1, ε_2, ε_3 are unobservables that we can regard either as latent variables or as disturbances. For the sake of this discussion, we will assume that U, ε_1, ε_2, ε_3 are mutually independent and arbitrarily distributed. Graphically, these influences can be represented by the path diagram of Figure 5.12.

The problem is as follows. We have drawn a long stream of independent samples of the process defined by (5.4) – (5.6) and have recorded the values of the observed variables X, Z, and Y; we now wish to estimate the unspecified quantities of the model to the greatest extent possible.

To clarify the scope of the problem, we consider its linear version, which is given by

$$x = u + \varepsilon_1, \tag{5.7}$$

$$z = \alpha x + \varepsilon_2, \tag{5.8}$$

$$y = \beta z + \gamma u + \varepsilon_3, \tag{5.9}$$

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

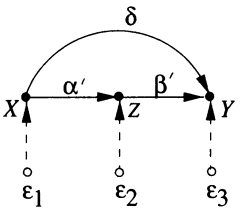

Figure 5.13 Diagram representing model M' of (5.12)–(5.14).

where U, ε_1, ε_2, ε_3 are uncorrelated, zero-mean disturbances.[14] It is not hard to show that parameters $\alpha, \beta,$ and γ can be determined uniquely from the correlations among the observed quantities X, Z, and Y. This identification was demonstrated already in the example of Figure 5.7, where the back-door criterion yielded

$$\beta = r_{YZ \cdot X}, \quad \alpha = r_{ZX}, \tag{5.10}$$

and hence

$$\gamma = r_{YX} - \alpha\beta. \tag{5.11}$$

Thus, returning to the nonparametric version of the model, it is tempting to generalize that, for the model to be identifiable, the functions $\{f_1, f_2, f_3\}$ must be determined uniquely from the data. However, the prospect of this happening is unlikely, because the mapping between functions and distributions is known to be many-to-one. In other words, given any nonparametric model M, if there exists one set of functions $\{f_1, f_2, f_3\}$ compatible with a given distribution $P(x, y, z)$, then there are infinitely many such functions (see Figure 1.6). Thus, it seems that nothing useful can be inferred from loosely specified models such as the one given by (5.4)–(5.6).

Identification is not an end in itself, however, even in linear models. Rather, it serves to answer practical questions of prediction and control. At issue is not whether the data permit us to identify the form of the equations but, instead, whether the data permit us to provide unambiguous answers to questions of the kind traditionally answered by parametric models.

When the model given by (5.4)–(5.6) is used strictly for prediction (i.e., to determine the probabilities of some variables given a set of observations on other variables), the question of identification loses much (if not all) of its importance; all predictions can be estimated directly from either the covariance matrices or the sample estimates of those covariances. If dimensionality reduction is needed (e.g., to improve estimation accuracy) then the covariance matrix can be encoded in a variety of simultaneous equation models, all of the same dimensionality. For example, the correlations among X, Y, and Z in the linear model M of (5.7)–(5.9) might well be represented by the model M' (Figure 5.13):

$$x = \varepsilon_1, \tag{5.12}$$

$$z = \alpha' x + \varepsilon_2, \tag{5.13}$$

$$y = \beta' z + \delta x + \varepsilon_3. \tag{5.14}$$

[14] An equivalent version of this model is obtained by eliminating U from the equations and allowing ε_1 and ε_3 to be correlated, as in Figure 5.7.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

This model is as compact as (5.7)–(5.9) and is covariance equivalent to M with respect to the observed variables X, Y, Z. Upon setting $\alpha' = \alpha, \beta' = \beta$, and $\delta = \gamma$, model M' will yield the same probabilistic predictions as those of the model of (5.7)–(5.9). Still, when viewed as data-generating mechanisms, the two models are not equivalent. Each tells a different story about the processes generating X, Y, and Z, so naturally their predictions differ concerning the changes that would result from subjecting these processes to external interventions.

5.3.3 Causal Effects: The Interventional Interpretation of Structural Equation Models

The differences between models M and M' illustrate precisely where the structural reading of simultaneous equation models comes into play, and why even causally shy researchers consider structural parameters more "meaningful" than covariances and other statistical parameters. Model M', defined by (5.12)–(5.14), regards X as a direct participant in the process that determines the value of Y, whereas model M, defined by (5.7)–(5.9), views X as an indirect factor whose effect on Y is mediated by Z. This difference is not manifested in the data itself but rather in the way the data would change in response to outside interventions. For example, suppose we wish to predict the expectation of Y after we intervene and fix the value of X to some constant x; this is denoted $E(Y \mid do(X = x))$. After $X = x$ is substituted into (5.13) and (5.14), model M' yields

$$E[Y \mid do(X = x)] = E[\beta'\alpha' x + \beta'\varepsilon_2 + \delta x + \varepsilon_3] \tag{5.15}$$

$$= (\beta'\alpha' + \delta)x; \tag{5.16}$$

model M yields

$$E[Y \mid do(X = x)] = E[\beta\alpha x + \beta\varepsilon_2 + \gamma u + \varepsilon_3] \tag{5.17}$$

$$= \beta\alpha x. \tag{5.18}$$

Upon setting $\alpha' = \alpha, \beta' = \beta$, and $\delta = \gamma$ (as required for covariance equivalence; see (5.10) and (5.11)), we see clearly that the two models assign different magnitudes to the (total) causal effect of X on Y: model M predicts that a unit change in x will change $E(Y)$ by the amount $\beta\alpha$, whereas model M' puts this amount at $\beta\alpha + \gamma$.

At this point, it is tempting to ask whether we should substitute $x - \varepsilon_1$ for u in (5.9) prior to taking expectations in (5.17). If we permit the substitution of (5.8) into (5.9), as we did in deriving (5.17), why not permit the substitution of (5.7) into (5.9) as well? After all (the argument runs), there is no harm in upholding a mathematical equality, $u = x - \varepsilon_1$, that the modeler deems valid. This argument is fallacious, however.[15] Structural equations are not meant to be treated as immutable mathematical equalities. Rather, they are meant to define a state of equilibrium – one that is *violated* when the equilibrium is perturbed by outside interventions. In fact, the power of structural equation models is

[15] Such arguments have led to Newcomb's paradox in the so-called evidential decision theory (see Section 4.1.1).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

that they encode not only the initial equilibrium state but also the information necessary for determining which equations must be violated in order to account for a new state of equilibrium. For example, if the intervention consists merely of holding X constant at x, then the equation $x = u + \varepsilon_1$, which represents the preintervention process determining X, should be overruled and replaced with the equation $X = x$. The solution to the new set of equations then represents the new equilibrium. Thus, the essential characteristic of structural equations that sets them apart from ordinary mathematical equations is that the former stand not for one but for many sets of equations, each corresponding to a subset of equations taken from the original model. Every such subset represents some hypothetical physical reality that would prevail under a given intervention.

If we take the stand that the value of structural equations lies not in summarizing distribution functions but in encoding causal information for predicting the effects of policies (Haavelmo 1943; Marschak 1950; Simon 1953), it is natural to view such predictions as the proper generalization of structural coefficients. For example, the proper generalization of the coefficient β in the linear model M would be the answer to the control query, "What would be the change in the expected value of Y if we were to intervene and change the value of Z from z to $z + 1$?", which is different, of course, from the observational query, "What would be the difference in the expected value of Y if we were to *find* Z at level $z + 1$ instead of level z?" Observational queries, as we discussed in Chapter 1, can be answered directly from the joint distribution $P(x, y, z)$, while control queries require causal information as well. Structural equations encode this causal information in their syntax by treating the variable on the left-hand side of the equality sign as the effect and treating those on the right as causes. In Chapter 3 we distinguished between the two types of queries through the symbol $do(\cdot)$. For example, we wrote

$$E(Y \mid do(x)) \triangleq E\,[Y \mid do(X = x)] \tag{5.19}$$

for the controlled expectation and

$$E(Y \mid x) \triangleq E(Y \mid X = x) \tag{5.20}$$

for the standard conditional or observational expectation. That $E(Y \mid do(x))$ does not equal $E(Y \mid x)$ can easily be seen in the model of (5.7)–(5.9), where $E(Y \mid do(x)) = \alpha\beta x$ but $E(Y \mid x) = r_{YX}x = (\alpha\beta + y)\,x$. Indeed, the passive observation $X = x$ should not violate any of the equations, and this is the justification for substituting both (5.7) and (5.8) into (5.9) before taking the expectation.

In linear models, the answers to questions of direct control are encoded in the path (or structural) coefficients, which can be used to derive the total effect of any variable on another. For example, the value of $E(Y \mid do(x))$ in the model defined by (5.7)–(5.9) is $\alpha\beta x$, that is, x times the product of the path coefficients along the path $X \rightarrow Z \rightarrow Y$. Computation of $E(Y \mid do(x))$ would be more complicated in the nonparametric case, even if we knew the functions f_1, f_2, and f_3. Nevertheless, this computation is well defined; it requires the solution (for the expectation of Y) of a modified set of equations in which f_1 is "wiped out" and X is replaced by the constant x:

$$z = f_2(x, \varepsilon_2), \tag{5.21}$$

$$y = f_3(z, u, \varepsilon_3). \tag{5.22}$$

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

Thus, computation of $E(Y \mid do(x))$ requires evaluation of

$$E(Y \mid do(x)) = E\{f_3\ [f_2(x, \varepsilon_2), u, \varepsilon_3]\},$$

where the expectation is taken over U, ε_2, and ε_3. Remarkably, graphical methods perform this computation without knowledge of f_2, f_3, and $P(\epsilon_2, \epsilon_3, u)$ (Section 3.3.2).

This is indeed the essence of identifiability in nonparametric models. The ability to answer interventional queries *uniquely*, from the data and the graph, is precisely how Definition 3.2.3 interprets the identification of the causal effect $P(y \mid do(x))$. As we have seen in Chapters 3 and 4, that ability can be discerned graphically, almost by inspection, from the diagrams that accompany the equations.

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

$$y = \beta x + \varepsilon$$

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

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

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,

$$y = \beta x + \varepsilon_1 \quad \text{and} \quad x = \alpha y + \varepsilon_2,$$

where $\text{cov}(X, \varepsilon_1) = \text{cov}(Y, \varepsilon_2) = 0$ and $\alpha = r_{XY} = \beta\sigma_X^2/\sigma_Y^2$. Thus, if the condition $\text{cov}(X, \varepsilon_1) = 0$ endows β with causal meaning, then $\text{cov}(Y, \varepsilon_2) = 0$ ought to endow α with causal meaning as well. But this too conflicts with both intuition and the intentions

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

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 $y = \beta x + \varepsilon$ *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* $\beta x + \varepsilon$, *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 $\beta x = -\varepsilon$), 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 $y = \beta x + \varepsilon$ 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

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

for all Z disjoint of $\{X \cup Y\}$.[17] In contrast, regression equations make no empirical claims whatsoever.

[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

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

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 $P(y \mid do(x), z)$, 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 $P(y \mid x)$ does not enjoy such a strong property of invariance, since $P(y \mid x)$ 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 uncorrected 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 $y = \beta x + \varepsilon$ is simply

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

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 $E(Y \mid x)$, 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

the value $\beta x' + \varepsilon$. 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 be tested only 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).

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

specific equations in the system) and not to general changes in the statistics of the variables. If we start with $\text{cov}(X, \varepsilon) = 0$ 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

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

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 $E[Y \mid do(x)]$ and not from its conditional expectation $E[Y \mid x]$. 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 $\varepsilon = y - \beta x$ 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

$$E[\varepsilon_Y \varepsilon_X] = E[YX \mid do(pa_Y, pa_X)] - E[Y \mid do(pa_Y)]E[X \mid do(pa_X)]. \tag{5.26}$$

Once we have determined the structural coefficients, the controlled expectations $E[Y \mid do(pa_Y)]$, $E[X \mid do(pa_X)]$, and $E[YX \mid do(pa_Y, pa_X)]$ 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 $E[\varepsilon_Y \varepsilon_X]$ but rather in determining whether ε_Y and ε_X can be assumed to be uncorrected. For this determination, it suffices to test whether the equality

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

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,

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

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 uncorrected. 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 uncorrected. 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 uncorrected 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, on 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

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

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 $P(y \mid do(x))$ 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* $P(y \mid do(x))$*, 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* $P(y \mid do(x), do(s_{XY}))$*, 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 intenvention*, 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 and W) would not give the correct value of zero for the direct effect of X on Y, whereas $\frac{\partial}{\partial x} E(Y \mid do(x, z, w))$ 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 only to 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 $\{y = \beta x + \varepsilon, x = \alpha y + \delta\}$, 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.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

variables are usually nonlinear, so the results of Section 4.5 should be applicable. In particular, the direct effect of X on Y 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 (4.11), which invokes nested counterfactuals, and may be reduced to (4.12).

The manipulative account that we have invoked in defining the empirical content of structural equations (Definition 5.4.1) is adequate in linear systems, where most causal quantities of interest can be inferred from experimental studies at the population level (see Section 11.7.1). In nonlinear and nonparametric models, we occasionally need to go down to the individual unit level and invoke the (more fundamental) counterfactual reading of structural equations, as articulated in equation (3.51) and footnote 17, page 160. The analysis of indirect effects is a case in point; its definition (4.14) rests on nested counterfactuals and cannot be expressed in terms of population averages. Such analysis is necessary to give indirect effects operational meaning, independent of total and direct effects (see Section 11.4.2). With the help of counterfactual language, however, we can give indirect effects a simple operational definition: The indirect effect of X on Y is the increase we would see in Y while holding X constant and increasing the mediating variables Z to whatever value Z would have attained under a unit increase of X (see Section 4.5.5 for a formal definition). In linear systems, this definition coincides, indeed, with the difference between the total and direct effects. See Chapter 11 for further discussion of the role of indirect effects in social science and policy analysis (Pearl 2005a).

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 the structural invariance of certain relationships under policy intervention.

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

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

Suppose that we consider intervening on a set of variables X and that we wish 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 $P(y \mid do(x))$. If we are interested in a set λ of parameters of that distribution, then our task is to estimate $\lambda\,[P(y \mid do(x)]$ 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 economic 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 $\lambda\,[P_M(y \mid do(x)]$ 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 $\lambda[P(y \mid do(x)]$ can be identified from the *conditional* distribution $P(y \mid x)$ 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 $P(y \mid do(x))$. We say that X is exogenous *relative to (Y, λ, T) if λ is identifiable from the conditional distribution $P(y \mid x)$, that is, if*

$$P_{M_1}(y \mid x) = P_{M_2}(y \mid x) \Longrightarrow \lambda[P_{M_1}(y \mid do(x))] = \lambda[P_{M_2}(y \mid do(x))] \tag{5.28}$$

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

$$P_{M_1}(y \mid x) = P_{M_2}(y \mid x) \Longrightarrow P_{M_1}(y \mid do(x)) = P_{M_2}(y \mid do(x)). \tag{5.29}$$

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

$$P(y \mid do(x)) = P(y \mid x), \tag{5.30}$$

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 $P(y \mid x)$ 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 exogenous

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

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

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 $P(y \mid x)$ remains invariant to changes in the "process that generates" X. This definition coincides[21] with (5.30) because $P(y \mid do(x))$ is governed by a structural model in which the equations determining X are wiped out; thus, $P(y \mid x)$ 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 $P(y \mid x)$ 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 $P(y \mid x)$ 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 $P(y \mid x)$ remaining invariant under such a 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 $P(y \mid do(x))$; the vocabulary of marginal and conditional densities is far too coarse to properly define the changes against which $P(y \mid x)$ 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.

[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.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

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 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* $P(y \mid x)$*, that is, if*

$$P_{M_1}(y \mid x) = P_{M_2}(y \mid x) \Longrightarrow \lambda(M_1) = \lambda(M_2) \tag{5.31}$$

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 $P(y \mid do(x))$ at selected points of x and y, then (5.31) reduces to the implication

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

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

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

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 $\lambda = P(w \mid do(z))$, where W and Z are two sets of variables in Y, then the exogeneity of X relative to this parameter ensures the identification of $P(w \mid do(z))$ from the conditional probability $P(z, w \mid x)$. 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,

[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.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

if X and Y are dichotomous and the parameter of interest is $\lambda = P(y_0 \mid x_0)$, then the marginal probability $P(x)$ certainly "involves" parameters such as

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

and λ "involves" their ratio:

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

Therefore, writing $P(x_0) = \lambda_2/\lambda$ shows 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 $P(y_0 \mid x_0)$; hence it is determined uniquely by $P(y_0 \mid x_0)$ 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)
A variable X is exogenous (relative to $\lambda = P(y \mid do(x))$) 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

[25] Engle et al. (1983, p. 281) and Hendry (1995, pp. 162–3) overcome this ambiguity by using selective "reparameterization" – a necessary step which textbooks tend to ignore.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

"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]

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).

5.5 CONCLUSION

Today the enterprise known as structural equation modeling is increasingly under fire. The founding fathers have retired, their teachings are forgotten, and practitioners, teachers, and researchers currently find the methodology they inherited difficult to either defend or supplant. Modern SEM textbooks are preoccupied with parameter estimation and rarely explicate the role that those parameters play in causal explanations or in policy analysis; examples dealing with the effects of interventions are conspicuously absent, for instance. Research in SEM now focuses almost exclusively on model fitting, while issues pertaining to the meaning and usage of SEM's models are subjects of confusion and controversy. Some of these confusions are reflected in the many questions that I have received from readers (Section 11.5), to whom I dedicated an "SEM Survival Kit" (Section 11.5.3) – a set of arguments for defending the causal reading of SEM and its scientific rationale.

I am thoroughly convinced that the contemporary crisis in SEM originates in the lack of a mathematical language for handling the causal information embedded in structural equations. Graphical models have provided such a language. They have thus helped us answer many of the unsettled questions that drive the current crisis:

1. Under what conditions can we give causal interpretation to structural coefficients?
2. What are the causal assumptions underlying a given structural equation model?
3. What are the statistical implications of any given structural equation model?
4. What is the operational meaning of a given structural coefficient?
5. What are the policy-making claims of any given structural equation model?
6. When is an equation not structural?

[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. 1996, p. 446). I disagree; "random assignment" is a misleading metaphor, while "omitted factors" shines in clarity.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

This chapter has described the conceptual developments that now resolve such foundational questions. (Sections 11.5.2 and 11.5.3 provide further elaboration.) In addition, we have presented several tools to be used in answering questions of practical importance:

1. When are two structural equation models observationally indistinguishable?
2. When do regression coefficients represent path coefficients?
3. When would the addition of a regressor introduce bias?
4. How can we tell, prior to collecting any data, which path coefficients can be identified?
5. When can we dispose of the linearity–normality assumption and still extract causal information from the data?

I remain hopeful that researchers will recognize the benefits of these concepts and tools and use them to revitalize causal analysis in the social and behavioral sciences.

5.6 Postscript for the Second Edition

5.6.1 An Econometric Awakening?

After decades of neglect of causal analysis in economics, a surge of interest seems to be in progress. In a recent series of papers, Jim Heckman (2000, 2003, 2005, 2007 (with Vytlacil)) has made great efforts to resurrect and reassert the Cowles Commission interpretation of structural equation models, and to convince economists that recent advances in causal analysis are rooted in the ideas of Haavelmo (1943), Marschak (1950), Roy (1951), and Hurwicz (1962). Unfortunately, Heckman still does not offer econometricians clear answers to the questions posed in this chapter (pp. 133, 170, 171, 215–217). In particular, unduly concerned with implementational issues, Heckman rejects Haavelmo's "equation wipe-out" as a basis for defining counterfactuals and fails to provide econometricians with an alternative definition, namely, a procedure, like that of equation (3.51), for computing the counterfactual $Y(x, u)$ in a well-posed economic model, with X and Y two arbitrary variables in the model. (See Sections 11.5.4–5.) Such a definition is essential for endowing the "potential outcome" approach with a formal semantics, based on SEM, and thus unifying the two econometric camps currently working in isolation.

Another sign of positive awakening comes from the social sciences, through the publication of Morgan and Winship's book *Counterfactual and Causal Inference* (2007), in which the causal reading of SEM is clearly reinstated.[27]

5.6.2 Identification in Linear Models

In a series of papers, Brito and Pearl (2002a,b, 2006) have established graphical criteria that significantly expand the class of identifiable semi-Markovian linear models beyond those discussed in this chapter. They first proved that identification is ensured in all

[27] Though the SEM basis of counterfactuals is unfortunately not articulated.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007

graphs that do not contain bow-arcs, that is, no error correlation is allowed between a cause and its *direct* effect, while no restrictions are imposed on errors associated with indirect causes (Brito and Pearl 2002b). Subsequently, generalizing the concept of instrumental variables beyond the classical patterns of Figures 5.9 and 5.11, they establish a general identification condition that is testable in polynomial time and subsumes all conditions known in the literature. See McDonald (2002a) for an algebraic approach, and Brito (2010) for a gentle introduction and a survey of results.

5.6.3 Robustness of Causal Claims

Causal claims in SEM are established through a combination of data and the set of causal assumptions embodied in the model. For example, the claim that the causal effect $E(Y \mid do(x))$ in Figure 5.9 is given by $\alpha x = r_{YZ}/r_{XZ}\, x$ is based on the assumptions: $cov(e_Z, e_Y) = 0$ and $E(Y \mid do(x, z)) = E(Y \mid do(x))$; both are shown in the graph. A claim is *robust* when it is insensitive to violations of some of the assumptions in the model. For example, the claim above is insensitive to the assumption $cov(e_Z, e_X) = 0$, which is shown in the model.

When several distinct sets of assumptions give rise to k distinct estimands for a parameter α, that parameter is called k-identified; the higher the k, the more robust are claims based on α, because equality among these estimands imposes $k - 1$ constraints on the covariance matrix which, if satisfied in the data, indicate an agreement among k distinct sets of assumptions, thus supporting their validity. A typical example emerges when several (independent) instrumental variables are available $Z_1, Z_2, \ldots, Z_k$ for a single link $X \rightarrow Y$, which yield the equalities $\alpha = r_{YZ_1}/r_{XZ_1} = r_{YZ_2}/r_{XZ_Z} = \cdots = r_{YZ_k}/r_{XZ_k}$.

Pearl (2004) gives a formal definition for this notion of robustness, and established graphical conditions for quantifying the degree of robustness of a given causal claim. k-identification generalizes the notion of *degree of freedom* in standard SEM analysis; the latter characterizes the entire model, while the former applies to individual parameters and, more generally, to individual causal claims.

Acknowledgments

This chapter owes its inspiration to the generations of statisticians who have asked, with humor and disbelief, how SEM's methodology could make sense to any rational being – and to the social scientists who (perhaps unwittingly) have saved the SEM tradition from drowning in statistical interpretations. The comments of Herman Ader, Peter Bentler, Kenneth Bollen, Jacques Hagenaars, Rod McDonald, Les Hayduk, and Stan Mulaik have helped me gain a greater understanding of SEM practice and vocabulary. John Aldrich, Nancy Cartwright, Arthur Goldberger, James Heckman, Kevin Hoover, Ed Leamer, and Herbert Simon helped me penetrate the mazes of structural equations and exogeneity in econometrics. Jin Tian was instrumental in revising Sections 5.2.3 and 5.3.1.

Downloaded from https://www.cambridge.org/core. Access paid by the UC Berkeley Library, on 29 Oct 2020 at 00:22:06, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511803161.007