

10.3.4	Path-Switching Causation	324
10.3.5	Temporal Preemption	325
10.4	Conclusions	327
<b>11</b>	<b>Reflections, Elaborations, and Discussions with Readers</b>	<b>331</b>
11.1	Causal, Statistical, and Graphical Vocabulary	331
11.1.1	Is the Causal-Statistical Dichotomy Necessary?	331
11.1.2	<i>d</i> -Separation without Tears (Chapter 1, pp. 16–18)	335
11.2	Reversing Statistical Time (Chapter 2, p. 58–59)	337
11.3	Estimating Causal Effects	338
11.3.1	The Intuition behind the Back-Door Criterion (Chapter 3, p. 79)	338
11.3.2	Demystifying “Strong Ignorability”	341
11.3.3	Alternative Proof of the Back-Door Criterion	344
11.3.4	Data vs. Knowledge in Covariate Selection	346
11.3.5	Understanding Propensity Scores	348
11.3.6	The Intuition behind <i>do</i> -Calculus	352
11.3.7	The Validity of <i>G</i> -Estimation	352
11.4	Policy Evaluation and the <i>do</i> -Operator	354
11.4.1	Identifying Conditional Plans (Section 4.2, p. 113)	354
11.4.2	The Meaning of Indirect Effects	355
11.4.3	Can <i>do</i> ( <i>x</i> ) Represent Practical Experiments?	358
11.4.4	Is the <i>do</i> ( <i>x</i> ) Operator Universal?	359
11.4.5	Causation without Manipulation!!!	361
11.4.6	Hunting Causes with Cartwright	362
11.4.7	The Illusion of Nonmodularity	364
11.5	Causal Analysis in Linear Structural Models	366
11.5.1	General Criterion for Parameter Identification (Chapter 5, pp. 149–54)	366
11.5.2	The Causal Interpretation of Structural Coefficients	366
11.5.3	Defending the Causal Interpretation of SEM (or, SEM Survival Kit)	368
11.5.4	Where Is Economic Modeling Today? – Courting Causes with Heckman	374
11.5.5	External Variation vs. Surgery	376
11.6	Decisions and Confounding (Chapter 6)	380
11.6.1	Simpson’s Paradox and Decision Trees	380
11.6.2	Is Chronological Information Sufficient for Decision Trees?	382
11.6.3	Lindley on Causality, Decision Trees, and Bayesianism	384
11.6.4	Why Isn’t Confounding a Statistical Concept?	387
11.7	The Calculus of Counterfactuals	389
11.7.1	Counterfactuals in Linear Systems	389
11.7.2	The Meaning of Counterfactuals	391
11.7.3	<i>d</i> -Separation of Counterfactuals	393

<b>Contents</b>	xiii
11.8 Instrumental Variables and Noncompliance	395
11.8.1 Tight Bounds under Noncompliance	395
11.9 More on Probabilities of Causation	396
11.9.1 Is “Guilty with Probability One” Ever Possible?	396
11.9.2 Tightening the Bounds on Probabilities of Causation	398
<b>Epilogue The Art and Science of Cause and Effect</b>	
A public lecture delivered November 1996 as part of the UCLA Faculty Research Lectureship Program	401
<i>Bibliography</i>	429
<i>Name Index</i>	453
<i>Subject Index</i>	459

## Reflections, Elaborations, and Discussions with Readers

*As X-rays are to the surgeon,  
graphs are for causation.*

The author

In this chapter, I reflect back on the material covered in Chapters 1 to 10, discuss issues that require further elaboration, introduce new results obtained in the past eight years, and answer questions of general interest posed to me by readers of the first edition. These range from clarification of specific passages in the text, to conceptual and philosophical issues concerning the controversial status of causation, how it is taught in classrooms and how it is treated in textbooks and research articles.

The discussions follow roughly the order in which these issues are presented in the book, with section numbers indicating the corresponding chapters.

### 11.1 CAUSAL, STATISTICAL, AND GRAPHICAL VOCABULARY

#### 11.1.1 Is the Causal–Statistical Dichotomy Necessary?

##### *Question to Author (from many readers)*

Chapter 1 (Section 1.5) insists on a sharp distinction between statistical and causal concepts; the former are definable in terms of a joint distribution function (of observed variables), the latter are not. Considering that many concepts which the book classifies as “causal” (e.g., “randomization,” “confounding,” and “instrumental variables”) are commonly discussed in the statistical literature, is this distinction crisp? Is it necessary? Is it useful?

##### *Author Answer*

The distinction is crisp,<sup>1</sup> necessary, and useful, and, as I tell audiences in all my lectures: “If you get nothing out of this lecture except the importance of keeping statistical and causal concepts apart, I would consider it a success.” Here, I would dare go even further:

---

<sup>1</sup> The basic distinction has been given a variety of other nomenclatures, e.g., descriptive vs. etiological, associational vs. causal, empirical vs. theoretical, observational vs. experimental, and many others. I am not satisfied with any of these surrogates, partly because they were not as crisply defined, partly because their boundaries got blurred through the years, and partly because the concatenation “nonstatistical” triggers openness to new perspectives.

“If I am remembered for no other contribution except for insisting on the causal–statistical distinction, I would consider my scientific work worthwhile.”

The distinction is embarrassingly crisp and simple, because it is based on the fundamental distinction between statics and kinematics. Standard statistical analysis, typified by regression, estimation, and hypothesis-testing techniques, aims to assess parameters of a static distribution from samples drawn of that distribution. With the help of such parameters, one can infer associations among variables, estimate the likelihood of past and future events, as well as update the likelihood of events in light of new evidence or new measurements. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis goes one step further; its aim is to infer not only the likelihood of events under static conditions, but also the dynamics of events under *changing conditions*, for example, changes induced by treatments or external interventions, or by new policies or new experimental designs.

This distinction implies that causal and statistical concepts do not mix. There is nothing in the joint distribution of symptoms and diseases to tell us that curing the former would or would not cure the latter. More generally, there is nothing in a distribution function to tell us how that distribution would differ if external conditions were to change – say, from observational to experimental setup – because the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified. This information must be provided by extra assumptions that identify what in the distribution remains invariant when the specified modification takes place. The sum total of these extra assumptions is what we call “causal knowledge.”

These considerations imply that the slogan “correlation does not imply causation” can be translated into a useful principle: behind every causal conclusion there must lie some causal assumption that is not discernible from the distribution function.

Take the concept of randomization – why is it not statistical? Assume we are given a bivariate density function  $f(x,y)$ , and we are told that one of the variables is randomized; can we tell which one it is by just examining  $f(x,y)$ ? Of course not; therefore, following our definition, randomization is a causal, not a statistical concept. Indeed, every randomized experiment is based on external *intervention*; that is, subjects are “forced” to take one treatment or another in accordance with the experimental protocol, regardless of their natural inclination. The presence of intervention immediately qualifies the experimental setup, as well as all relationships inferred from that setup, as causal.

Note, however, that the purpose of the causal–statistical demarcation line (as stated in Section 1.4, p. 40) is not to exclude causal concepts from the province of statistical analysis but, rather, to encourage investigators to treat causal concepts distinctly, with the proper set of mathematical and inferential tools. Indeed, statisticians were the first to conceive of randomized experiments, and have used them successfully since the time of Fisher (1926). However, both the assumptions and conclusions in those studies were kept implicit, in the mind of ingenious investigators; they did not make their way into the mathematics. For example, one would be extremely hard pressed to find a statistics textbook, even at the graduate level, containing a mathematical proof that randomization indeed produces unbiased estimates of the quantities we wish estimated – i.e., efficacy of treatments or policies.

As a related example, very few statistics teachers today can write down a formula stating that “randomized experiments prove drug  $x_1$  to be twice as effective as drug  $x_2$ .”

Of course, they can write:  $P(y | x_1)/P(y | x_2) = 2$  ( $y$  being the desirable outcome), but then they must keep in mind that this ratio applies to a specific randomized condition, and should not be confused with likelihood ratios prevailing in observational studies. Scientific progress requires that such distinctions be expressed mathematically.<sup>2</sup>

The most important contribution of causal analysis in the past two decades has been the emergence of mathematical languages in which not merely the data, but the experimental design itself can be given mathematical description. Such description is essential, in fact, if one wishes the results of one experiment to serve as premises in another, or to predict outcomes of one design from data obtained under another, or merely to decide if we are in possession of sufficient knowledge to render such cross-design predictions possible.

### *Is the Distinction Necessary?*

Science thrives on distinctions, especially those that do not readily mix. The distinction between rational and irrational numbers, for example, is extremely important in number theory, for it spares us futile efforts to define the latter through some arithmetic operations on the former. The same can be said about the distinctions between prime, composite, algebraic, and transcendental numbers. Logicians, among them George Boole (1815–1864) and Augustus De Morgan (1806–1871), wasted half a century trying to prove syllogisms of first-order logic (e.g., all men are mortal) using the machinery of propositional logic; the distinction between the two was made crisp only at the end of the nineteenth century.

A similar situation occurred in the history of causality. Philosophers have struggled for half a century trying to reduce causality to probabilities (Section 7.5) and have gotten nowhere, except for traps such as “evidential decision theory” (Section 4.1). Epidemiologists have struggled for half a century to define “confounding” in the language of associations (Chapter 6, pp. 183, 194). Some are still struggling (see Section 11.6.4). This effort could have been avoided by appealing to first principles: If confounding were a statistical concept, we would have been able to identify confounders from features of nonexperimental data, adjust for those confounders, and obtain unbiased estimates of causal effects. This would have violated our golden rule: behind any causal conclusion there must be some causal assumption, untested in observational studies. That epidemiologists did not recognize in advance the futility of such attempts is a puzzle that can have only two explanations: they either did not take seriously the causal–statistical divide, or were afraid to classify “confounding” – a simple, intuitive concept – as “nonstatistical.”

Divorcing simple concepts from the province of statistics – the most powerful formal language known to empirical scientists – can be traumatic indeed. Social scientists have been laboring for half a century to evaluate public policies using statistical analysis, anchored in regression techniques, and only recently have confessed, with great disappointment, what should have been recognized as obvious in the 1960’s: “Regression analyses typically do nothing more than produce from a data set a collection of conditional means and conditional variances” (Berk 2004, p. 237). Economists have gone through a

<sup>2</sup> The potential-outcome approach of Neyman (1923) and Rubin (1974) does offer a notational distinction, by writing  $P(Y_{x_1} = y)/P(Y_{x_2} = y) = 2$  for the former, and  $P(y | x_1)/P(y | x_2) = 2$  for the latter. However, the opaqueness of this notation and the incomplete state of its semantics (see Sections 3.6.3 and 11.3.2) have prevented it from penetrating classrooms, textbooks, and laboratories.

similar trauma with the concept of exogeneity (Section 5.4.3). Even those who recognized that a strand of exogeneity (i.e., superexogeneity) is of a causal variety came back to define it in terms of distributions (Maddala 1992; Hendry 1995) – crossing the demarcation line was irresistible. And we understand why; defining concepts in term of prior and conditional distributions – the ultimate oracles of empirical knowledge – was considered a mark of scientific prudence. We know better now.

### *Is the Distinction Useful?*

I am fairly confident that today, enlightened by failed experiments in philosophy, epidemiology, and economics, no reputable discipline would waste half a century chasing after a distribution-based definition of another causal concept, however tempted by prudence or intuition. Today, the usefulness of the demarcation line lies primarily in helping investigators trace the assumptions that are needed to support various types of scientific claims. Since every claim invoking causal concepts must rely on some judgmental premises that invoke causal vocabulary, and since causal vocabulary can only be formulated in causally distinct notation, the demarcation line provides notational tools for identifying the judgmental assumptions to which every causal claim is vulnerable.

Statistical assumptions, even untested, are testable in principle, given a sufficiently large sample and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control. This difference stands out in Bayesian analysis. Though the priors that Bayesians commonly assign to statistical parameters are untested quantities, the sensitivity to these priors tends to diminish with increasing sample size. In contrast, sensitivity to prior causal assumptions – say, that treatment does not change gender – remains high regardless of sample size.

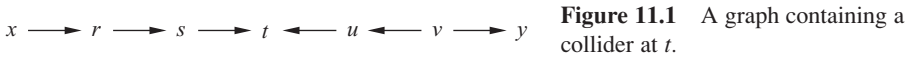
This makes it doubly important that the notation we use for expressing causal assumptions be meaningful and unambiguous so that scientists can clearly judge the plausibility or inevitability of the assumptions articulated.

### *How Does One Recognize Causal Expressions in the Statistical Literature?*

Those versed in the potential-outcome notation (Neyman 1923; Rubin 1974; Holland 1988) can recognize such expressions through the subscripts that are attached to counterfactual events and variables, e.g.,  $Y_x(u)$  or  $Z_{xy}$ . (Some authors use parenthetical expressions, e.g.,  $Y(x, u)$  or  $Z(x, y)$ .) (See Section 3.6.3 for semantics.)

Alternatively, this book also uses expressions of the form  $P(Y = y \mid do(X = x))$  or  $P(Y_x = y)$  to denote the probability (or frequency) that event  $(Y = y)$  would occur if treatment condition  $X = x$  were enforced uniformly over the population. (Clearly,  $P(Y = y \mid do(X = x))$  is equivalent to  $P(Y_x = y)$ .) Still a third formal notation is provided by graphical models, where the arrows represent either causal influences, as in Definition 1.3.1, or functional (i.e., counterfactual) relationships, as in Figure 1.6(c).

These notational devices are extremely useful for detecting and tracing the causal premises with which every causal inference study must commence. Any causal premise that is cast in standard probability expressions, void of graphs, counterfactual subscripts, or  $do(*)$  operators, can safely be discarded as inadequate. Consequently, any article describing an empirical investigation that does not commence with expressions involving graphs, counterfactual subscripts, or  $do(*)$  can safely be proclaimed as inadequately written.



**Figure 11.1** A graph containing a collider at  $t$ .

While this harsh verdict may condemn valuable articles in the empirical literature to the province of inadequacy, it can save investigators endless hours of confusion and argumentation in deciding whether causal claims from one study are relevant to another. More importantly, the verdict should encourage investigators to visibly explicate causal premises, so that they can be communicated unambiguously to other investigators and invite professional scrutiny, deliberation, and refinement.

### 11.1.2 $d$ -Separation without Tears (Chapter 1, pp. 16–18)

At the request of many who have had difficulties switching from algebraic to graphical thinking, I am including a gentle introduction to  $d$ -separation, supplementing the formal definition given in Chapter 1, pp. 16–18.

#### *Introduction*

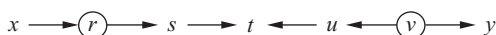
$d$ -separation is a criterion for deciding, from a given causal graph, whether a set  $X$  of variables is independent of another set  $Y$ , given a third set  $Z$ . The idea is to associate “dependence” with “connectedness” (i.e., the existence of a connecting path) and “independence” with “unconnectedness” or “separation.” The only twist on this simple idea is to define what we mean by “connecting path,” given that we are dealing with a system of directed arrows in which some vertices (those residing in  $Z$ ) correspond to measured variables, whose values are known precisely. To account for the orientations of the arrows we use the terms “ $d$ -separated” and “ $d$ -connected” ( $d$  connotes “directional”). We start by considering separation between two singleton variables,  $x$  and  $y$ ; the extension to sets of variables is straightforward (i.e., two sets are separated if and only if each element in one set is separated from every element in the other).

#### *Unconditional Separation*

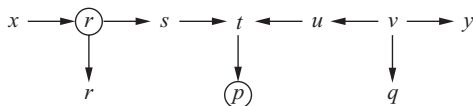
**Rule 1:**  $x$  and  $y$  are  $d$ -connected if there is an unblocked path between them.

By a “path” we mean any consecutive sequence of edges, disregarding their directionalities. By “unblocked path” we mean a path that can be traced without traversing a pair of arrows that collide “head-to-head.” In other words, arrows that meet head-to-head do not constitute a connection for the purpose of passing information; such a meeting will be called a “collider.”

**Example 11.1.1** The graph in Figure 11.1 contains one collider, at  $t$ . The path  $x - r - s - t$  is unblocked, hence  $x$  and  $t$  are  $d$ -connected. So also is the path  $t - u - v - y$ , hence  $t$  and  $y$  are  $d$ -connected, as well as the pairs  $u$  and  $y$ ,  $t$  and  $v$ ,  $t$  and  $u$ ,  $x$  and  $s$ , etc. However,  $x$  and  $y$  are not  $d$ -connected; there is no way of tracing a path from  $x$  to  $y$  without traversing the collider at  $t$ . Therefore, we conclude that  $x$  and  $y$  are  $d$ -separated, as well as  $x$  and  $v$ ,  $s$  and  $u$ ,  $r$  and  $u$ , etc. (In linear models, the ramification is that the covariance terms corresponding to these pairs of variables will be zero, for every choice of model parameters.)



**Figure 11.2** The set  $Z = \{r, v\}$   $d$ -separates  $x$  from  $t$  and  $t$  from  $y$ .



**Figure 11.3**  $s$  and  $y$  are  $d$ -connected given  $p$ , a descendant of the collider  $t$ .

**Blocking by Conditioning**

**Motivation:** When we measure a set  $Z$  of variables, and take their values as given, the conditional distribution of the remaining variables changes character; some dependent variables become independent, and some independent variables become dependent. To represent this dynamic in the graph, we need the notion of “conditional  $d$ -connectedness” or, more concretely, “ $d$ -connectedness, conditioned on a set  $Z$  of measurements.”

**Rule 2:**  $x$  and  $y$  are  $d$ -connected, conditioned on a set  $Z$  of nodes, if there is a collider-free path between  $x$  and  $y$  that traverses no member of  $Z$ . If no such path exists, we say that  $x$  and  $y$  are  $d$ -separated by  $Z$ . We also say then that every path between  $x$  and  $y$  is “blocked” by  $Z$ .

**Example 11.1.2** Let  $Z$  be the set  $\{r, v\}$  (marked by circles in Figure 11.2). Rule 2 tells us that  $x$  and  $y$  are  $d$ -separated by  $Z$ , and so also are  $x$  and  $s$ ,  $u$  and  $y$ ,  $s$  and  $u$ , etc. The path  $x - r - s$  is blocked by  $Z$ , and so also are the paths  $u - v - y$  and  $s - t - u$ . The only pairs of unmeasured nodes that remain  $d$ -connected in this example, conditioned on  $Z$ , are  $s$  and  $t$  and  $u$  and  $t$ . Note that, although  $t$  is not in  $Z$ , the path  $s - t - u$  is nevertheless blocked by  $Z$ , since  $t$  is a collider, and is blocked by Rule 1.

**Conditioning on Colliders**

**Motivation:** When we measure a common effect of two independent causes, the causes become dependent, because finding the truth of one makes the other less likely (or “explained away,” p. 17), and refuting one implies the truth of the other. This phenomenon (known as Berkson paradox, or “explaining away”) requires a slightly special treatment when we condition on colliders (representing common effects) or on any descendant of a collider (representing evidence for a common effect).

**Rule 3:** If a collider is a member of the conditioning set  $Z$ , or has a descendant in  $Z$ , then it no longer blocks any path that traces this collider.

**Example 11.1.3** Let  $Z$  be the set  $\{r, p\}$  (again, marked with circles in Figure 11.3). Rule 3 tells us that  $s$  and  $y$  are  $d$ -connected by  $Z$ , because the collider at  $t$  has a descendant ( $p$ ) in  $Z$ , which unblocks the path  $s - t - u - v - y$ . However,  $x$  and  $u$  are still  $d$ -separated by  $Z$ , because although the linkage at  $t$  is unblocked, the one at  $r$  is blocked by Rule 2 (since  $r$  is in  $Z$ ).

This completes the definition of  $d$ -separation, and readers are invited to try it on some more intricate graphs, such as those shown in Chapter 1, Figure 1.3.

**Typical application:** Consider Example 11.1.3. Suppose we form the regression of  $y$  on  $p$ ,  $r$ , and  $x$ ,

$$y = c_1p + c_2r + c_3x + \epsilon,$$

and wish to predict which coefficient in this regression is zero. From the discussion above we can conclude immediately that  $c_3$  is zero, because  $y$  and  $x$  are  $d$ -separated given  $p$  and  $r$ , hence  $y$  is independent of  $x$  given  $p$  and  $r$ , or,  $x$  cannot offer any information about  $y$  once we know  $p$  and  $r$ . (Formally, the partial correlation between  $y$  and  $x$ , conditioned on  $p$  and  $r$ , must vanish.)  $c_1$  and  $c_2$ , on the other hand, will in general not be zero, as can be seen from the graph:  $Z = \{r, x\}$  does not  $d$ -separate  $y$  from  $p$ , and  $Z = \{p, x\}$  does not  $d$ -separate  $y$  from  $r$ .

**Remark on correlated errors:** Correlated exogenous variables (or error terms) need no special treatment. These are represented by bi-directed arcs (double-arrowed), and their arrowheads are treated as any other arrowheads for the purpose of path tracing. For example, if we add to the graph in Figure 11.3 a bi-directed arc between  $x$  and  $t$ , then  $y$  and  $x$  will no longer be  $d$ -separated (by  $Z = \{r, p\}$ ), because the path  $x - t - u - v - y$  is  $d$ -connected – the collider at  $t$  is unblocked by virtue of having a descendant,  $p$ , in  $Z$ .

## 11.2 REVERSING STATISTICAL TIME (CHAPTER 2, pp. 58–59)

### *Question to Author:*

Keith Markus requested a general method of achieving time reversal by changing coordinate systems or, in the specific example of equation (2.3), a general method of solving for the parameters  $a$ ,  $b$ ,  $c$ , and  $d$  to make the statistical time run opposite to the physical time (p. 59).

### *Author's Reply:*

Consider any two time-dependent variables  $X(t)$  and  $Y(t)$ . These may represent the position of two particles in one dimension, temperature and pressure, sales and advertising budget, and so on.

Assume that temporal variation of  $X(t)$  and  $Y(t)$  is governed by the equations:

$$\begin{aligned} X(t) &= \alpha X(t-1) + \beta Y(t-1) + \epsilon(t) \\ Y(t) &= \gamma X(t-1) + \delta Y(t-1) + \eta(t), \end{aligned} \tag{11.1}$$

with  $\epsilon(t)$  and  $\eta(t)$  being mutually and serially uncorrelated noise terms.

In this coordinate system, we find that the two components of the current state,  $X(t)$  and  $Y(t)$ , are uncorrelated conditioned on the components of the previous state,  $X(t-1)$  and  $Y(t-1)$ . Simultaneously, the components of the current state,  $X(t)$  and  $Y(t)$ , are correlated conditioned on the components of the future state,  $X(t+1)$  and  $Y(t+1)$ . Thus, according to Definition 2.8.1 (p. 58), the statistical time coincides with the physical time.

Now let us rotate the coordinates using the transformation

$$\begin{aligned} X'(t) &= aX(t) + bY(t) \\ Y'(t) &= cX(t) + dY(t). \end{aligned} \tag{11.2}$$

The governing physical equations remain the same as equation (11.1), but, written in the new coordinate system, they read

$$\begin{aligned} X'(t) &= \alpha'X'(t-1) + \beta'Y(t-1) + \epsilon'(t) \\ Y'(t) &= \gamma'X'_i(t-1) + \gamma'Y'(t-1) + \eta'(t). \end{aligned} \tag{11.3}$$

The primed coefficients can be obtained from the original (unprimed) coefficients by matrix multiplication. Likewise, we have:

$$\begin{aligned} \epsilon'(t) &= a\epsilon(t) + b\eta(t) \\ \eta'(t) &= c\epsilon(t) + d\eta(t). \end{aligned}$$

Since  $\epsilon(t)$  and  $\eta(t)$  are uncorrelated,  $\epsilon'(t)$  and  $\eta'(t)$  will be correlated, and we no longer have the condition that the components of the current state,  $X'(t)$  and  $Y'(t)$ , are uncorrelated conditioned on the components of the previous state,  $X'(t-1)$  and  $Y'(t-1)$ . Thus, the statistical time (if there is one) no longer runs along the physical time.

Now we need to show that we can choose the parameters  $a$ ,  $b$ ,  $c$ , and  $d$  in such a way as to have the statistical time run opposite to the physical time, namely, to make the components of the current state,  $X'(t)$  and  $Y'(t)$ , uncorrelated conditioned on the components of the future state,  $X'(t+1)$  and  $Y'(t+1)$ .

By inverting equation (11.3) we can express  $X'(t-1)$  and  $Y'(t-1)$  in terms of linear combinations of  $X'(t)$ ,  $Y'(t)$ ,  $\epsilon'(t)$ , and  $\eta'(t)$ . Clearly, since  $e(t)$  and  $h(t)$  are uncorrelated, we can choose  $a$ ,  $b$ ,  $c$ ,  $d$  in such a way that the noise term appearing in the  $X'(t-1)$  equation is uncorrelated with the one appearing in the  $Y'(t-1)$  equation. (This is better demonstrated in matrix calculus.)

Thus, the general principle for selecting the alternative coordinate system is to diagonalize the noise correlation matrix in the reverse direction.

I hope that readers will undertake the challenge of testing the Temporal Bias Conjecture (p. 59):

“In most natural phenomenon, the physical time coincides with at least one statistical time.”

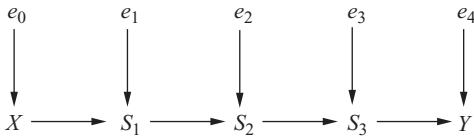
Alex Balke (personal communication) tried to test it with economic time series, but the results were not too conclusive, for lack of adequate data. I still believe the conjecture to be true, and I hope readers will have better luck.

## 11.3 ESTIMATING CAUSAL EFFECTS

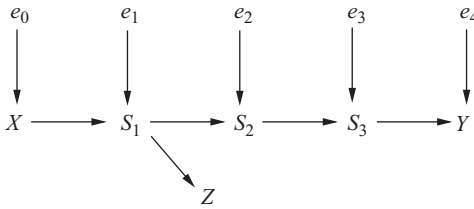
### 11.3.1 The Intuition behind the Back-Door Criterion (Chapter 3, p. 79)

#### *Question to Author:*

In the definition of the back-door condition (p. 79, Definition 3.3.1), the exclusion of  $X$ 's descendants (Condition (i)) seems to be introduced as an after fact, just because we get into trouble if we don't. Why can't we get it from first principles; first define sufficiency of  $Z$  in terms of the goal of removing bias, and then show that, to achieve this goal, we neither want nor need descendants of  $X$  in  $Z$ .



**Figure 11.4** Showing the noise factors on the path from  $X$  to  $Y$ .



**Figure 11.5** Conditioning on  $Z$  creates dependence between  $X$  and  $e_1$ , which biases the estimated effect of  $X$  on  $Y$ .

**Author’s Answer:**

The exclusion of descendants from the back-door criterion is indeed based on first principles, in terms of the goal of removing bias. The principles are as follows: We wish to measure a certain quantity (causal effect) and, instead, we measure a dependency  $P(y | x)$  that results from all the paths in the diagram; some are spurious (the back-door paths), and some are genuinely causal (the directed paths from  $X$  to  $Y$ ). Thus, to remove bias, we need to modify the measured dependency and make it equal to the desired quantity. To do this systematically, we condition on a set  $Z$  of variables while ensuring that:

1. We block all spurious paths from  $X$  to  $Y$ ,
2. We leave all directed paths unperturbed,
3. We create no new spurious paths.

Principles 1 and 2 are accomplished by blocking all back-door paths and only those paths, as articulated in condition (ii). Principle 3 requires that we do not condition on descendants of  $X$ , even those that do not block directed paths, because such descendants may create new spurious paths between  $X$  and  $Y$ . To see why, consider the graph

$$X \rightarrow S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow Y.$$

The intermediate variables,  $S_1, S_2, \dots$ , (as well as  $Y$ ) are affected by noise factors  $e_0, e_1, e_2, \dots$  which are not shown explicitly in the diagram. However, under magnification, the chain unfolds into the graph in Figure 11.4.

Now imagine that we condition on a descendant  $Z$  of  $S_1$  as shown in Figure 11.5. Since  $S_1$  is a collider, this creates dependency between  $X$  and  $e_1$  which is equivalent to a back-door path

$$X \leftrightarrow e_1 \rightarrow S_1 \rightarrow S_2 \rightarrow S_3 \rightarrow Y.$$

By principle 3, such paths should not be created, for it introduces spurious dependence between  $X$  and  $Y$ .

Note that a descendant  $Z$  of  $X$  that is not also a descendant of some  $S_i$  escapes this exclusion; it can safely be conditioned on without introducing bias (though it may decrease the efficiency of the associated estimator of the causal effect of  $X$  on  $Y$ ). Section

11.3.3 provides an alternative proof of the back-door criterion where the need to exclude descendants of  $X$  is even more transparent.

It is also important to note that the danger of creating new bias by adjusting for wrong variables can threaten randomized trials as well. In such trials, investigators may wish to adjust for covariates despite the fact that, asymptotically, randomization neutralizes both measured and unmeasured confounders. Adjustment may be sought either to improve precision (Cox 1958, pp. 48–55), or to match imbalanced samples, or to obtain covariate-specific causal effects. Randomized trials are immune to adjustment-induced bias when adjustment is restricted to pre-treatment covariates, but adjustment for post-treatment variables may induce bias by the mechanism shown in Figure 11.5 or, more severely, when correlation exists between the adjusted variable  $Z$  and some factor that affects outcome (e.g.,  $e_4$  in Figure 11.5).

As an example, suppose treatment has a side effect (e.g., headache) in patients who are predisposed to disease  $Y$ . If we wish to adjust for disposition and adjust instead for its proxy, headache, a bias would emerge through the spurious path: treatment  $\rightarrow$  headache  $\leftarrow$  predisposition  $\rightarrow$  disease. However, if we are careful never to adjust for any consequence of treatment (not only those that are on the causal pathway to disease), no bias will emerge in randomized trials.

#### ***Further Questions from This Reader:***

This explanation for excluding descendants of  $X$  is reasonable, but it has two shortcomings:

1. It does not address cases such as

$$X \leftarrow C \rightarrow Y \rightarrow F,$$

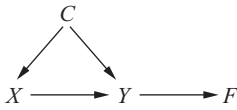
which occur frequently in epidemiology, and where tradition permits the adjustment for  $Z = \{C, F\}$ .

2. The explanation seems to redefine confounding and sufficiency to represent something different from what they have meant to epidemiologists in the past few decades. Can we find something in graph theory that is closer to their traditional meaning?

#### ***Author's Answer***

1. Epidemiological tradition permits the adjustment for  $Z = (C, F)$  for the task of testing whether  $X$  has a causal effect on  $Y$ , but not for estimating the magnitude of that effect. In the former case, while conditioning on  $F$  creates a spurious path between  $C$  and the noise factor affecting  $Y$ , that path is blocked upon conditioning on  $C$ . Thus, conditioning on  $Z = \{C, F\}$  leaves  $X$  and  $Y$  independent. If we happen to measure such dependence in any stratum of  $Z$ , it must be that the model is wrong, i.e., either there is a direct causal effect of  $X$  on  $Y$ , or some other paths exist that are not shown in the graph.

Thus, if we wish to test the (null) hypothesis that there is no causal effect of  $X$  on  $Y$ , adjusting for  $Z = \{C, F\}$  is perfectly legitimate, and the graph shows it (i.e.,  $C$  and  $F$  are nondescendant of  $X$ ). However, adjusting for  $Z$  is not legitimate for assessing the causal effect of  $X$  on  $Y$  when such effect is suspected,



**Figure 11.6** Graph applicable for accessing the effect of  $X$  on  $Y$ .

because the graph applicable for this task is given in Figure 11.6;  $F$  becomes a descendant of  $X$ , and is excluded by the back-door criterion.

2. If the explanation of confounding and sufficiency sounds at variance with traditional epidemiology, it is only because traditional epidemiologists did not have proper means of expressing the operations of blocking or creating dependencies. They might have had a healthy intuition about dependencies, but graphs translate this intuition into a formal system of closing and opening paths.

We should also note that before 1985, causal analysis in epidemiology was in a state of confusion, because the healthy intuitions of leading epidemiologists had to be expressed in the language of associations – an impossible task. Even the idea that confounding stands for “bias,” namely, a “difference between two dependencies, one that we wish to measure, the other that we do measure,” was resisted by many (see Chapter 6), because they could not express the former mathematically.<sup>3</sup>

Therefore, instead of finding “something in graph language that is closer to traditional meaning,” we can do better: explicate what that “traditional meaning” ought to have been.

In other words, traditional meaning was informal and occasionally misguided, while graphical criteria are formal and mathematically proven.

Chapter 6 (pp. 183, 194) records a long history of epidemiological intuitions, some by prominent epidemiologists, that have gone astray when confronted with questions of confounding and adjustment (see Greenland and Robins 1986; Wickramaratne and Holford 1987; Weinberg 1993). Although most leading epidemiologists today are keenly attuned to modern developments in causal analysis, (e.g., Glymour and Greenland 2008), epidemiological folklore is still permeated with traditional intuitions that are highly suspect. (See Section 6.5.2.)

In summary, graphical criteria, as well as principles 1–3 above, give us a sensible, friendly, and unambiguous interpretation of the “traditional meaning of epidemiological concepts.”

### 11.3.2 Demystifying “Strong Ignorability”

Researchers working within the confines of the potential-outcome language express the condition of “zero bias” or “no-confounding” using an independence relationship called

<sup>3</sup> Recall that Greenland and Robins (1986) were a lone beacon of truth for many years, and even they had to resort to the language of “exchangeability” to define “bias,” which discouraged intuitive interpretations of confounding (see Section 6.5.3). Indeed, it took epidemiologists another six years (Weinberg 1993) to discover that adjusting for factors affected by the exposure (as in Figure 11.5) would introduce bias.

“strong ignorability” (Rosenbaum and Rubin 1983). Formally, if  $X$  is a binary treatment (or action), strong ignorability is written as:

$$\{Y(0), Y(1)\} \perp\!\!\!\perp X \mid Z, \quad (11.4)$$

where  $Y(0)$  and  $Y(1)$  are the (unobservable) potential outcomes under actions  $do(X = 0)$  and  $do(X = 1)$ , respectively (see equation (3.51) for definition), and  $Z$  is a set of measured covariates. When “strong ignorability” holds,  $Z$  is *admissible*, or *deconfounding*, that is, treatment effects can be estimated without bias using the adjustment estimand, as shown in the derivation of equation (3.54).

Strong ignorability, as the derivation shows, is a convenient syntactic tool for manipulating counterfactual formulas, as well as a convenient way of formally assuming admissibility (of  $Z$ ) without having to justify it. However, as we have noted several times in this book, hardly anyone knows how to apply it in practice, because the counterfactual variables  $Y(0)$  and  $Y(1)$  are unobservable, and scientific knowledge is not stored in a form that allows reliable judgment about conditional independence of counterfactuals. It is not surprising, therefore, that “strong ignorability” is used almost exclusively as a surrogate for the assumption “ $Z$  is admissible,” that is,

$$P(y \mid do(x)) = \sum_z P(y \mid z, x)P(z), \quad (11.5)$$

and rarely, if ever, as a criterion to protect us from bad choices of  $Z$ .<sup>4</sup>

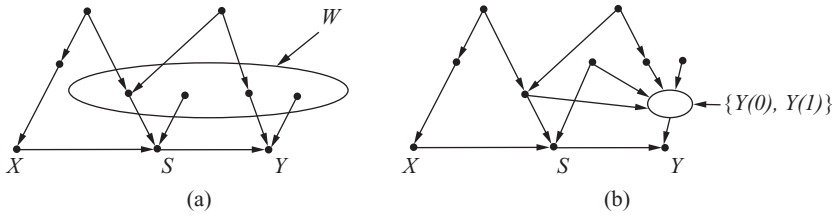
Readers enlightened by graphical models would recognize immediately that equation (11.4) must coincide with the back-door criterion (p. 79, Definition 3.3.1), since the latter too entails admissibility. This recognition allows us not merely to pose equation (11.4) as a claim, or an assumption, but also to reason about the cause–effect relationships that render it valid.

The question arises, however, whether the variables  $Y(0)$  and  $Y(1)$  could be represented in the causal graph in a way that would allow us to test equation (11.4) by graphical means, using  $d$ -separation. In other words, we seek a set  $W$  of nodes such that  $Z$  would  $d$ -separate  $X$  from  $W$  if and only if  $Z$  satisfies equation (11.4).

The answer follows directly from the rules of translation between graphs and potential outcome (Section 3.6.3). According to this translation,  $\{Y(0), Y(1)\}$  represents the sum total of all exogenous variables, latent as well as observed, which can influence  $Y$  through paths that avoid  $X$ . The reason is as follows: according to the structural definition of  $\{Y(0), Y(1)\}$  (equation (3.51)),  $Y(0)$  (similarly  $Y(1)$ ) represents the value of  $Y$  under a condition where all arrows entering  $X$  are severed, and  $X$  is held constant at  $X = 0$ . Statistical variations of  $Y(0)$  would therefore be governed by all exogenous ancestors of  $Y$  in the mutilated graphs with the arrows entering  $X$  removed.

In Figure 11.4, for example,  $\{Y(0), Y(1)\}$  will be represented by the exogenous variables  $\{e_1, e_2, e_3, e_4\}$ . In Figure 3.4, as another example,  $\{Y(0), Y(1)\}$  will be represented by the noise factors (not shown in the graph) that affect variables  $X_4, X_1, X_2, X_5$ , and  $X_6$ . However,

<sup>4</sup> In fact, in the rare cases where “strong ignorability” is used to guide the choice of covariates, the guidelines issued are wrong or inaccurate, perpetuating myths such as: “there is no reason to avoid adjustment for a variable describing subjects before treatment,” “a confounder is any variable associated with both treatment and disease,” and “strong ignorability requires measurement of all covariates related to both treatment and outcome” (citations withheld to spare embarrassment).



**Figure 11.7** Graphical interpretation of counterfactuals  $\{Y(0), Y(1)\}$  in the “strong ignorability” condition.

since variables  $X_4$  and  $X_5$  summarize (for  $Y$ ) the variations of their ancestors, a sufficient set for representing  $\{Y(0), Y(1)\}$  would be  $X_4, X_1$  and the noise factors affecting  $Y$  and  $X_6$ .

In summary, the potential outcomes  $\{Y(0), Y(1)\}$  are represented by the observed and unobserved parents<sup>5</sup> of all nodes on paths from  $X$  to  $Y$ . Schematically, we can represent these parents as in Figure 11.7(a). It is easy to see that, with this interpretation of  $\{Y(0), Y(1)\}$ , a set of covariates  $Z$   $d$ -separates  $W$  from  $X$  if and only if  $Z$  satisfies the back-door criterion.

It should be noted that the set of observable variables designated  $W$  in Figure 11.7(a) are merely surrogates of the unobservable counterfactuals  $\{Y(0), Y(1)\}$  for the purpose of confirming conditional independencies (e.g., equation (11.4)) in the causal graph (via  $d$ -separation.) A more accurate allocation of  $\{Y(0), Y(1)\}$  is given in Figure 11.7(b), where they are shown as (dummy) parents of  $Y$  that are functions of, though not identical to, the actual (observable) parents of  $Y$  and  $S$ .

Readers versed in structural equation modeling would recognize the graphical representations  $\{Y(0), Y(1)\}$  as a refinement of the classical econometric notion of “disturbance,” or “error term” (in the equation for  $Y$ ), and “strong ignorability” as the requirement that, for  $X$  to be “exogenous,” it must be independent of this “disturbance” (see Section 5.4.3). This notion fell into ill repute in the 1970s (Richard 1980) together with the causal interpretation of econometric equations, and I have predicted its re-acceptance (p. 170) in view of the clarity that graphical models shine on the structural equation formalism. Figure 11.7 should further help this acceptance.

Having translated “strong ignorability” into a simple separation condition in a model that encodes substantive process knowledge should demystify the nebulous concept of “strong ignorability” and invite investigators who speak “ignorability” to benefit from its graphical interpretation.

This interpretation permits researchers to understand what conditions covariates must fulfill before they eliminate bias, what to watch for and what to think about when covariates are selected, and what experiments we can do to test, at least partially, if we have the knowledge needed for covariate selection. Section 11.3.4 exemplifies such considerations.

One application where the symbiosis between the graphical and counterfactual frameworks has been useful is in estimating the effect of treatments on the treated:  $ETT = P(Y_{x'} = y | x)$  (see Sections 8.2.5 and 11.9.1). This counterfactual quantity (e.g., the probability that a treated person would recover if not treated, or the rate of disease among the exposed, had the exposure been avoided) is not easily analyzed in the  $do$ -calculus notation. The counterfactual notation, however, allows us to derive a

<sup>5</sup> The reason for explicitly including latent parents is explained in Section 11.3.1.

useful conclusion: Whenever a set of covariates  $Z$  exists that satisfies the back-door criterion, ETT can be estimated from observational studies. This follows directly from

$$(Y \perp\!\!\!\perp X | Z)_{G_{\underline{X}}} \implies Y_{x'} \perp\!\!\!\perp X | Z,$$

which allows us to write

$$\begin{aligned} \text{ETT} &= P(Y_{x'} = y | x) \\ &= \sum_z P(Y_{x'} = y | x, z)P(z | x) \\ &= \sum_z P(Y_{x'} = y | x', z)P(z | x) \\ &= \sum_z P(y | x', z)P(z | x). \end{aligned}$$

The graphical demystification of “strong ignorability” also helps explain why the probability of causation  $P(Y_{x'} = y' | x, y)$  and, in fact, any counterfactual expression conditioned on  $y$ , would not permit such a derivation and is, in general, non-identifiable (see Chapter 9).

### 11.3.3 Alternative Proof of the Back-Door Criterion

The original proof of the back-door criterion (Theorem 3.3.2) used an auxiliary intervention node  $F$  (Figure 3.2) and was rather indirect. An alternative proof is presented below, where the need for restricting  $Z$  to nondescendants of  $X$  is transparent.

#### *Proof of the Back-Door Criterion*

Consider a Markovian model  $G$  in which  $T$  stands for the set of parents of  $X$ . From equation (3.13), we know that the causal effect of  $X$  on  $Y$  is given by

$$P(y | \hat{x}) = \sum_{t \in T} P(y | x, t)P(t). \quad (11.6)$$

Now assume some members of  $T$  are unobserved. We seek another set  $Z$  of observed variables, to replace  $T$  so that

$$P(y | \hat{x}) = \sum_{z \in Z} P(y | x, z)P(z). \quad (11.7)$$

It is easily verified that (11.7) follow from (11.6) if  $Z$  satisfies:

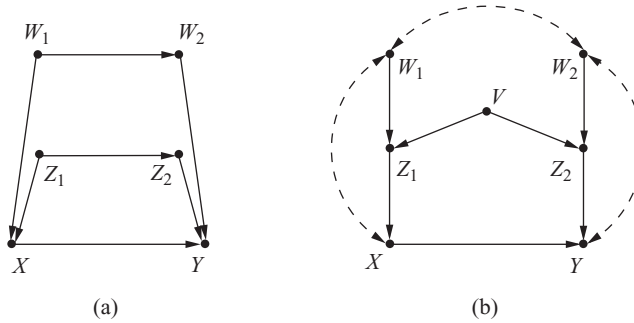
- (i)  $(Y \perp\!\!\!\perp T | X, Z)$
- (ii)  $(X \perp\!\!\!\perp Z | T)$ .

Indeed, conditioning on  $Z$ , (i) permits us to rewrite (11.6) as

$$P(y | \hat{x}) = \sum_t P(t) \sum_z P(y | z, x)P(z | t, x),$$

and (ii) further yields  $P(z | t, x) = P(z | t)$ , from which (11.7) follows.

It is now a purely graphical exercise to prove that the back-door criterion implies (i) and (ii). Indeed, (ii) follows directly from the fact that  $Z$  consists of nondescendants of  $X$ , while the blockage of all back-door paths by  $Z$  implies  $(Y \perp\!\!\!\perp T | X, Z)_G$ , hence (i). This follows from observing that any path from  $Y$  to  $T$  in  $G$  that is unblocked by  $\{X, Z\}$  can be extended to a back-door path from  $Y$  to  $X$ , unblocked by  $Z$ .



**Figure 11.8** (a)  $S_1 = \{Z_1, W_2\}$  and  $S_2 = \{Z_2, W_1\}$  are each admissible yet not satisfying  $C_1$  or  $C_2$ . (b) No subset of  $C = \{Z_1, Z_2, W_1, W_2, V\}$  is admissible.

**On Recognizing Admissible Sets of Deconfounders**

Note that conditions (i) and (ii) allow us to recognize a set  $Z$  as *admissible* (i.e., satisfying equation (11.7)) starting from any other admissible  $T$ , not necessarily the parents of  $X$ . The parenthood status of  $T$  was used merely to establish (11.6) but played no role in replacing  $T$  with  $Z$  to establish (11.7). Still, starting with the parent set  $T$  has the unique advantage of allowing us to recognize *every* other admissible set  $Z$  via (i) and (ii). For any other starting set,  $T$ , there exists an admissible  $Z$  that does not satisfy (i) and (ii). For an obvious example, choosing  $X$ 's parents for  $Z$  would violate (i) and (ii) because no set can  $d$ -separate  $X$  from its parents as would be required by (i).

Note also that conditions (i) and (ii) are purely statistical, invoking no knowledge of the graph or any other causal assumption. It is interesting to ask, therefore, whether there are general independence conditions, similar to (i) and (ii), that connect any two admissible sets,  $S_1$  and  $S_2$ . A partial answer is given by the Stone–Robins criterion (page 187) for the case where  $S_1$  is a subset of  $S_2$ ; another is provided by the following observation.

Define two subsets,  $S_1$  and  $S_2$ , as *c-equivalent* (“ $c$ ” connotes “confounding”) if the following equality holds:

$$\sum_{s_1} P(y | x, s_1)P(s_1) = \sum_{s_2} P(y | x, s_2)P(s_2). \tag{11.8}$$

This equality guarantees that, if adjusted for, sets  $S_1$  and  $S_2$  would produce the same bias relative to estimating the causal effect of  $X$  on  $Y$ .

**Claim:** A sufficient condition for  $c$ -equivalence of  $S_1$  and  $S_2$  is that either one of the following two conditions holds:

$$\begin{aligned} C_1 : X \perp\!\!\!\perp S_2 | S_1 & \quad \text{and} \quad Y \perp\!\!\!\perp S_1 | S_2, X \\ C_2 : X \perp\!\!\!\perp S_1 | S_2 & \quad \text{and} \quad Y \perp\!\!\!\perp S_2 | S_1, X. \end{aligned}$$

$C_1$  permits us to derive the right-hand side of equation (11.8) from the left-hand side, while  $C_2$  permits us to go the other way around. Therefore, if  $S_1$  is known to be admissible, the admissibility of  $S_2$  can be confirmed by either  $C_1$  or  $C_2$ . This broader condition allows us, for example, to certify  $S_2 = PA_X$  as admissible from any other admissible set  $S_1$ , since condition  $C_2$  would be satisfied by any such choice.

This broader condition still does not characterize *all*  $c$ -equivalent pairs,  $S_1$  and  $S_2$ . For example, consider the graph in Figure 11.8(a), in which each of  $S_1 = \{Z_1, W_2\}$  and

$S_2 = \{Z_2, W_2\}$  is admissible (by virtue of satisfying the back-door criterion), hence  $S_1$  and  $S_2$  are  $c$ -equivalent. Yet neither  $C_1$  nor  $C_2$  holds in this case.

A natural attempt would be to impose the condition that  $S_1$  and  $S_2$  each be  $c$ -equivalent to  $S_1 \cup S_2$  and invoke the criterion of Stone (1993) and Robins (1997) for the required set-subset equivalence. The resulting criterion, while valid, is still not complete; there are cases where  $S_1$  and  $S_2$  are  $c$ -equivalent yet not  $c$ -equivalent to their union. A theorem by Pearl and Paz (2008) broadens this condition using irreducible sets.

Having given a conditional-independence characterization of  $c$ -equivalence does not solve, of course, the problem of identifying admissible sets; the latter is a causal notion and cannot be given statistical characterization.

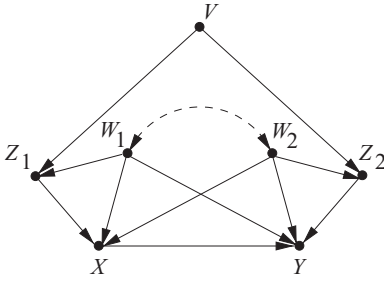
The graph depicted in Figure 11.8(b) demonstrates the difficulties commonly faced by social and health scientists. Suppose our target is to estimate  $P(y | do(x))$  given measurements on  $\{X, Y, Z_1, Z_2, W_1, W_2, V\}$ , but having no idea of the underlying graph structure. The conventional wisdom is to start with all available covariates  $C = \{Z_1, Z_2, W_1, W_2, V\}$ , and test if a proper subset of  $C$  would yield an equivalent estimand upon adjustment. Statistical methods for such reduction are described in Greenland et al. (1999b), Geng et al. (2002), and Wang et al. (2008). For example,  $\{Z_1, V\}$ ,  $\{Z_2, V\}$ , or  $\{Z_1, Z_2\}$  can be removed from  $C$  by successively applying conditions  $C_1$  and  $C_2$ . This reduction method would produce three irreducible subsets,  $\{Z_1, W_1, W_2\}$ ,  $\{Z_2, W_1, W_2\}$ , and  $\{V, W_1, W_2\}$ , all  $c$ -equivalent to the original covariate set  $C$ . However, none of these subsets is admissible for adjustment, because none (including  $C$ ) satisfies the back-door criterion. While a theorem due to Tian et al. (1998) assures us that any  $c$ -equivalent subset of a set  $C$  can be reached from  $C$  by a step-at-a-time removal method, going through a sequence of  $c$ -equivalent subsets, the problem of covariate selection is that, lacking the graph structure, we do not know which (if any) of the many subsets of  $C$  is admissible. The next subsection discusses how external knowledge, as well as more refined analysis of the data at hand, can be brought to bear on the problem.

### 11.3.4 Data vs. Knowledge in Covariate Selection

What then can be done in the absence of a causal graph? One way is to postulate a plausible graph, based on one's understanding of the domain, and check if the data refutes any of the statistical claims implied by that graph. In our case, the graph of Figure 11.8(b) advertises several such claims, cast as conditional independence constraints, each associated with a missing arrow in the graph:

$$\begin{array}{lll} V \perp\!\!\!\perp X | Z_1, W_1 & V \perp\!\!\!\perp W_2 & Z_2 \perp\!\!\!\perp W_1 | W_2 \\ V \perp\!\!\!\perp Y | X, Z_2, W_2 & Z_1 \perp\!\!\!\perp Z_2 | V, W_1, W_2 & Z_1 \perp\!\!\!\perp W_2 | W_1 \\ V \perp\!\!\!\perp W_1 & X \perp\!\!\!\perp Z_2 | Z, W_1, W_2 & X \perp\!\!\!\perp \{V, Z_2\} | Z_1, W_1, W_2. \end{array}$$

Satisfying these constraints does not establish, of course, the validity of the causal model postulated because, as we have seen in Chapter 2, alternative models may exist which satisfy the same independence constraints yet embody markedly different causal structures, hence, markedly different admissible sets and effect estimands. A trivial example would be a complete graph, with arbitrary orientation of arrows which, with a clever choice of parameters, can emulate any other graph. A less trivial example, one that is not sensitive to choice of parameters, lies in the class of equivalent structures, in



**Figure 11.9** A model that is statistically indistinguishable from that of Figure 11.8 (b), in which the irreducible sets  $\{Z_1, W_1, W_2\}$ ,  $\{W_1, W_2, V\}$ , and  $\{W_1, W_2, Z_2\}$  are admissible.

which all conditional independencies emanate from graph separations. The search techniques developed in Chapter 2 provide systematic ways of representing all equivalent models compatible with a given set of conditional independence relations.

For example, the model depicted in Figure 11.9 is indistinguishable from that of Figure 11.8(b), in that it satisfies all the conditional independencies implied by the latter, and no others.<sup>6</sup> However, in contrast to Figure 11.8(b), the sets  $\{Z_1, W_1, W_2\}$ ,  $\{V, W_1, W_2\}$ , and  $\{Z_2, W_1, W_2\}$  are admissible. Adjusting for the latter would remove bias if the correct model is Figure 11.9 and might produce bias if the correct model is Figure 11.8(b).

Is there a way of telling the two models apart? Although the notion of “observational equivalence” precludes discrimination by statistical means, substantive causal knowledge may provide discriminating information. For example, the model of Figure 11.9 can be ruled out if we have good reasons to believe that variable  $W_2$  cannot have any influence on  $X$  (e.g., it may occur *later* than  $X$ ), or that  $W_1$  could not possibly have direct effect on  $Y$ .

The power of graphs lies in offering investigators a transparent language to reason about, to discuss the plausibility of such assumptions and, when consensus is not reached, to isolate differences of opinion and identify what additional observations would be needed to resolve differences. This facility is lacking in the potential-outcome approach where, for most investigators, “strong ignorability” remains a mystical black box.

In addition to serving as carriers of substantive judgments, graphs also offer one the ability to reject large classes of models without testing each member of the class. For example, all models in which  $V$  and  $W_1$  are the sole parents of  $X$ , thus rendering  $\{V, W_1\}$  (as well as  $C$ ) admissible, could be rejected at once if the condition  $X \perp\!\!\!\perp Z_1 \mid V, W_1$  does not hold in the data.

In Chapter 3, for example, we demonstrated how the measurement of an additional variable, mediating between  $X$  and  $Y$ , was sufficient for identifying the causal effect of  $X$  on  $Y$ . This facility can also be demonstrated in Figure 11.8(b); measurement of a variable  $Z$  judged to be on the pathway between  $X$  and  $Y$  would render  $P(y \mid do(x))$  identifiable and estimable through equation (3.29). This is predicated, of course, on Figure 11.8(b) being the correct data-generating model. If, on the other hand, it is Figure 11.9 that represents the correct model, the causal effect would be given by

$$\begin{aligned} P(y \mid do(x)) &= \sum_{pa_X} P(y \mid pa_X, x) P(pa_X) \\ &= \sum_{z_1, w_1, w_2} P(y \mid x, z_1, w_1, w_2) P(z_1, w_1, w_2), \end{aligned}$$

<sup>6</sup> Semi-Markovian models may also be distinguished by functional relationships that are not expressible as conditional independencies (Verma and Pearl 1990; Tian and Pearl 2002b; Shpitser and Pearl 2008). We do not consider these useful constraints in this example.

which might or might not agree with equation (3.29). In the latter case, we would have good reason to reject the model in Figure 11.9 as inconsistent, and seek perhaps additional measurements to confirm or refute Figure 11.8(b).

Auxiliary experiments may offer an even more powerful discriminatory tool than auxiliary observations. Consider variable  $W_1$  in Figure 11.8(b). If we could conduct a controlled experiment with  $W_1$  randomized, instead of  $X$ , the data obtained would enable us to estimate the causal effect of  $X$  on  $Y$  with no bias (see Section 3.4.4). At the very least, we would be able to discern whether  $W_1$  is a parent of  $X$ , as in Figure 11.9, or an indirect ancestor of  $X$ , as in Figure 11.8(b).

In an attempt to adhere to traditional statistical methodology, some causal analysts have adopted a method called “sensitivity analysis” (e.g., Rosenbaum 2002, pp. 105–170), which gives the impression that causal assumptions are not invoked in the analysis. This, of course, is an illusion. Instead of drawing inferences by assuming the absence of certain causal relationships in the model, the analyst tries such assumptions and evaluates how strong alternative causal relationships must be in order to explain the observed data. The result is then submitted to a judgment of plausibility, the nature of which is no different from the judgments invoked in positing a model like the one in Figure 11.9. In its richer setting, sensitivity analysis amounts to loading a diagram with causal relationships whose strength is limited by plausibility judgments and, given the data, attempting to draw conclusions without violating those plausibility constraints. It is a noble endeavor, which thus far has been limited to problems with a very small number of variables. The advent of diagrams promises to expand the applicability of this method to more realistic problems.

### 11.3.5 Understanding Propensity Scores

The method of propensity score (Rosenbaum and Rubin 1983), or propensity score matching (PSM), is the most developed and popular strategy for causal analysis in observational studies. It is not emphasized in this book, because it is an estimation method, designed to deal with the variability of finite samples, but does not add much to our understanding of the asymptotic, large-sample limits, which is the main focus of the book. However, due to the prominence of the propensity score method in causal analysis, and recent controversies surrounding its usage, we devote this section to explain where it falls in the grand scheme of graphical models, admissibility, identifiability, bias reduction, and the statistical vs. causal dichotomy.

The method of propensity score is based on a simple, yet ingenious, idea of purely statistical character. Assuming a binary action (or treatment)  $X$ , and an arbitrary set  $S$  of measured covariates, the propensity score  $L(s)$  is the probability that action  $X = 1$  will be chosen by a participant with characteristics  $S = s$ , or

$$L(s) = P(X = 1 \mid S = s). \quad (11.9)$$

What Rosenbaum and Rubin showed is that, viewing  $L(s)$  as a function of  $S$ , hence, as a random variable,  $X$  and  $S$  are independent given  $L(s)$ , that is,  $X \perp\!\!\!\perp S \mid L(s)$ . In words, all units that map into the same value of  $L(s)$  are comparable, or “balanced,” in the sense that, within each stratum of  $L$ , treated and untreated units have the same distribution of characteristics  $S$ .<sup>7</sup>

<sup>7</sup> This independence emanates from the special nature of the function  $L(s)$  and is not represented in the graph, i.e., if we depict  $L$  as a child of  $S$ ,  $L$  would not in general  $d$ -separate  $S$  from  $X$ .

To see the significance of this result, let us assume, for simplicity, that  $L(s)$  can be estimated separately and approximated by discrete strata  $L = \{l_1, l_2, \dots, l_k\}$ . The conditional independence  $X \perp\!\!\!\perp S \mid L(s)$ , together with the functional mapping  $S \rightarrow L$ , renders  $S$  and  $L$   $c$ -equivalent in the sense defined in Section 11.3.3, equation (11.8), namely, for any  $Y$ ,

$$\sum_s P(y \mid s, x)P(s) = \sum_l P(y \mid l, x)P(l). \tag{11.10}$$

This follows immediately by writing:<sup>8</sup>

$$\begin{aligned} \sum_l P(y \mid l, x)P(l) &= \sum_s \sum_l P(y \mid l, s, x)P(l)P(s \mid l, x) \\ &= \sum_s \sum_l P(y \mid s, x)P(l)P(s \mid l) \\ &= \sum_s P(y \mid s, x)P(s). \end{aligned}$$

Thus far we have not mentioned any causal relationship, nor the fact that  $Y$  is an outcome variable and that, eventually, our task would be to estimate the causal effect of  $X$  on  $Y$ . The  $c$ -equivalence of  $S$  and  $L$  merely implies that, if for any reason one wishes to estimate the “adjustment estimand”  $\sum_s P(y \mid s, x)P(s)$ , with  $S$  and  $Y$  two arbitrary sets of variables, then, instead of summing over a high-dimensional set  $S$ , one might as well sum over a one-dimensional vector  $L(s)$ . The asymptotic estimate, in the limit of a very large sample, would be the same in either method.

This  $c$ -equivalence further implies – and this is where causal inference first comes into the picture – that if one chooses to approximate the causal effect  $P(y \mid do(x))$  by the adjustment estimand  $E_s P(y \mid s, x)$ , then, asymptotically, the same approximation can be achieved using the estimand  $E_l P(y \mid l, x)$ , where the adjustment is performed over the strata of  $L$ . The latter has the advantage that, for finite samples, each of the strata is less likely to be empty and each is likely to contain both treated and untreated units.

The method of propensity score can thus be seen as an efficient estimator of the adjustment estimand, formed by an arbitrary set of covariates  $S$ ; it makes no statement regarding the appropriateness of  $S$ , nor does it promise to correct for any confounding bias, or to refrain from creating new bias where none exists.

In the special case where  $S$  is admissible, that is,

$$P(y \mid do(x)) = E_s P(y \mid s, x), \tag{11.11}$$

$L$  would be admissible as well, and we would then have an unbiased estimand of the causal effect,<sup>9</sup>

$$P(y \mid do(x)) = E_l P(y \mid l, x),$$

accompanied by an efficient method of estimating the right-hand side. Conversely, if  $S$  is inadmissible,  $L$  would be inadmissible as well, and all we can guarantee is that the bias produced by the former would be faithfully and efficiently reproduced by the latter.

<sup>8</sup> This also follows from the fact that condition  $C_2$  is satisfied by the substitution  $S_1 = S$  and  $S_2 = L(s)$ .

<sup>9</sup> Rosenbaum and Rubin (1983) proved the  $c$ -equivalence of  $S$  and  $L$  only for admissible  $S$ , which is unfortunate; it gives readers the impression that the propensity score matching somehow contributes to bias reduction.

### *The Controversy Surrounding Propensity Score*

Thus far, our presentation of propensity score leaves no room for misunderstanding, and readers of this book would find it hard to understand how a controversy could emerge from an innocent estimation method which merely offers an efficient way of estimating a statistical quantity that sometimes does, and sometimes does not, coincide with the causal quantity of interest, depending on the choice of  $S$ .

But a controversy has developed recently, most likely due to the increased popularity of the method and the strong endorsement it received from prominent statisticians (Rubin 2007), social scientists (Morgan and Winship 2007; Berk and de Leeuw 1999), health scientists (Austin 2007), and economists (Heckman 1992). The popularity of the method has in fact grown to the point where some federal agencies now expect program evaluators to use this approach as a substitute for experimental designs (Peikes et al. 2008). This move reflects a general tendency among investigators to play down the cautionary note concerning the required admissibility of  $S$ , and to interpret the mathematical proof of Rosenbaum and Rubin as a guarantee that, in each strata of  $L$ , matching treated and untreated subjects somehow eliminates confounding from the data and contributes therefore to overall bias reduction. This tendency was further reinforced by empirical studies (Heckman et al. 1998; Dehejia and Wahba 1999) in which agreement was found between propensity score analysis and randomized trials, and in which the agreement was attributed to the ability of the former to “balance” treatment and control groups on important characteristics. Rubin has encouraged such interpretations by stating: “This application uses propensity score methods to create subgroups of treated units and control units ... as if they had been randomized. The collection of these subgroups then ‘approximate’ a randomized block experiment with respect to the observed covariates” (Rubin 2007).

Subsequent empirical studies, however, have taken a more critical view of propensity score, noting with disappointment that a substantial bias is sometimes measured when careful comparisons are made to results of clinical studies (Smith and Todd 2005; Luellen et al. 2005; Peikes et al. 2008).

But why would anyone play down the cautionary note of Rosenbaum and Rubin when doing so would violate the golden rule of causal analysis: No causal claim can be established by a purely statistical method, be it propensity scores, regression, stratification, or any other distribution-based design. The answer, I believe, rests with the language that Rosenbaum and Rubin used to formulate the condition of admissibility, i.e., equation (1.11). The condition was articulated in the restricted language of potential-outcome, stating that the set  $S$  must render  $X$  “strongly ignorable,” i.e.,  $\{Y_1, Y_0\} \perp\!\!\!\perp X \mid S$ . As stated several times in this book, the opacity of “ignorability” is the Achilles’ heel of the potential-outcome approach – no mortal can apply this condition to judge whether it holds even in simple problems, with all causal relationships correctly specified, let alone in partially specified problems that involve dozens of variables.<sup>10</sup>

<sup>10</sup> Advocates of the potential outcome tradition are invited to inspect Figure 11.8(b) (or any model, or story, or toy-example of their choice) and judge whether any subset of  $C$  renders  $X$  “strongly ignorable.” This could easily be determined, of course, by the back-door criterion, but, unfortunately, graphs are still feared and misunderstood by some of the chief advocates of the potential-outcome camp (e.g., Rubin 2004, 2008b, 2009).

The difficulty that most investigators experience in comprehending what “ignorability” means, and what judgment it summons them to exercise, has tempted them to assume that it is automatically satisfied, or at least is likely to be satisfied, if one includes in the analysis as many covariates as possible. The prevailing attitude is that adding more covariates can cause no harm (Rosenbaum 2002, p. 76) and can absolve one from thinking about the causal relationships among those covariates, the treatment, the outcome and, most importantly, the confounders left unmeasured (Rubin 2009).

This attitude stands contrary to what students of graphical models have learned, and what this book has attempted to teach. The admissibility of  $S$  can be established only by appealing to the causal knowledge available to the investigator, and that knowledge, as we know from graph theory and the back-door criterion, makes bias reduction a non-monotonic operation, i.e., eliminating bias (or imbalance) due to one confounder may awaken and unleash bias due to dormant, unmeasured confounders. Examples abound (e.g., Figure 6.3) where adding a variable to the analysis not only is not needed, but would introduce irreparable bias. (Pearl 2009, Shrier 2009, Sjölander 2009).

Another factor inflaming the controversy has been the general belief that the bias-reducing potential of propensity score methods can be tested experimentally by running case studies and comparing effect estimates obtained by propensity scores to those obtained by controlled randomized experiments (Shadish and Cook 2009).<sup>11</sup> This belief is unjustified because the bias-reducing potential of propensity scores depends critically on the specific choice of  $S$  or, more accurately, on the cause–effect relationships among variables inside and outside  $S$ . Measuring significant bias in one problem instance (say, an educational program in Oklahoma) does not preclude finding zero bias in another (say, crime control in Arkansas), even under identical statistical distributions  $P(x, s, y)$ .

With these considerations in mind, one is justified in asking a social science type question: What is it about propensity scores that has inhibited a more general understanding of their promise and limitations?

Richard Berk, in *Regression Analysis: A Constructive Critique* (Berk 2004), recalls similar phenomena in social science, where immaculate ideas were misinterpreted by the scientific community: “I recall a conversation with Don Campbell in which he openly wished that he had never written Campbell and Stanley (1966). The intent of the justly famous book, *Experimental and Quasi-Experimental Designs for Research*, was to contrast randomized experiments to quasi-experimental approximations and to strongly discourage the latter. Yet the apparent impact of the book was to legitimize a host of quasi-experimental designs for a wide variety of applied social science. After I got to know Dudley Duncan late in his career, he said that he often thought that his influential book on path analysis, *Introduction to Structural Equation Models* was a big mistake. Researchers had come away from the book believing that fundamental policy questions about social inequality could be quickly and easily answered with path analysis.” (p. xvii)

---

<sup>11</sup> Such beliefs are encouraged by valiant statements such as: “For dramatic evidence that such an analysis can reach the same conclusion as an exactly parallel randomized experiment, see Shadish and Clark (2006, unpublished)” (Rubin 2007).

I believe that a similar cultural phenomenon has evolved around propensity scores.

It is not that Rosenbaum and Rubin were careless in stating the conditions for success. Formally, they were very clear in warning practitioners that propensity scores work only under “strong ignorability” conditions. However, what they failed to realize is that it is not enough to warn people against dangers they cannot recognize; to protect them from perilous adventures, we must also give them eyeglasses to spot the threats, and a meaningful language to reason about them. By failing to equip readers with tools (e.g., graphs) for recognizing how “strong ignorability” can be violated or achieved, they have encouraged a generation of researchers (including federal agencies) to assume that ignorability either holds in most cases, or can be made to hold by clever designs.

### 11.3.6 The Intuition behind *do*-Calculus

#### *Question to Author Regarding Theorem 3.4.1:*

In the inference rules of *do*-calculus (p. 85), the subgraph  $G_{\bar{X}}$ , represents the distribution prevailing under the operation  $do(X = x)$ , since all direct causes of  $X$  are removed. What distribution does the submodel  $G_{\underline{X}}$  represent, with the direct effects of  $X$  removed?

#### *Author’s Reply:*

The removal of direct effects is a purely graphical operation that leaves us with a graph in which all directed paths from  $X$  to  $Y$  are disconnected, while all back-door paths remain intact. So, if  $X$  and  $Y$  are  $d$ -connected in that graph, it must be due to (unblocked) confounding paths between the two. Conversely, if we find a set  $Z$  of nodes that  $d$ -separate  $X$  from  $Y$  in that graph, we are assured that  $Z$  blocks all back-door paths in the original graph. If we further condition on variables  $Z$ , we are assured, by the back-door criterion, that we have neutralized all confounders and that whatever dependence we measure after such conditioning must be due to the causal effect of  $X$  on  $Y$ , free of confoundings.

### 11.3.7 The Validity of *G*-Estimation

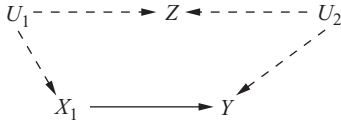
In Section 3.6.4 we introduced the *G*-estimation formula (3.63), together with the counterfactual independency (3.62),  $(Y(x) \perp\!\!\!\perp X_k \mid \bar{L}_k, \bar{X}_{k-1} = \bar{x}_{k-1})$ , which Robins proved to be a sufficient condition for (3.63). In general, condition (3.62) is both overrestrictive and lacks intuitive basis. A more general and intuitive condition leading to (3.63) is derived in (4.5) (p. 122), which reads as follows:

#### **(3.62\*) General Condition for *g*-Estimation (Sequential Deconfounding)**

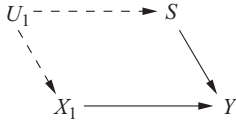
$P(y \mid g = x)$  is identifiable and is given by (3.63) if every action-avoiding back-door path from  $X_k$  to  $Y$  is blocked by some subset  $L_k$  of nondescendants of  $X_k$ . (By “action-avoiding” we mean a path containing no arrows entering an  $X$  variable later than  $X_k$ .)

This condition bears several improvements over (3.62), as demonstrated in the following three examples.

**Example 11.3.1** Figure 11.10 demonstrates cases where the *g*-formula (3.63) is valid with a subset  $L_k$  of the past but not with the entire past. Assuming  $U_1$  and  $U_2$  are



**Figure 11.10** Conditioning on the entire past  $L_1 = Z$  would invalidate  $g$ -estimation.



**Figure 11.11**  $g$ -estimation is rendered valid by including a non-predecessor  $S$ .

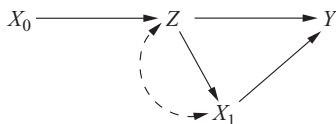
unobserved, and temporal order:  $U_1, Z, X_1, U_2, Y$ , we see that both (3.62) and (3.62\*), hence (3.63), are satisfied with  $L_1 = 0$ , while taking the whole past  $L_1 = Z$  would violate both.

**Example 11.3.2** Figure 11.11 demonstrates cases where defining  $L_k$  as the set of “nondescendants” of  $X_k$  (as opposed to temporal predecessors of  $X_k$ ) broadens (3.62). Assuming temporal order:  $U_1, X_1, S, Y$ , both (3.62) and (3.62\*) are satisfied with  $L_1 = S$ , but not with  $L_1 = 0$ .

**Example 11.3.3** (constructed by Ilya Shpitser in response to Eliezer Yudkowsky) Figure 11.12 demonstrates cases where (3.62) is not satisfied even with the new interpretation of  $L_k$ , but the graphical condition (3.62\*) is. It is easy to see that (3.62\*) is satisfied; all back-door action-avoiding paths from  $X_1$  to  $Y$  are blocked by  $\{X_0, Z\}$ . At the same time, it is possible to show (using the Twin Network Method, p. 213) that  $Y(x_0, x_1)$  is not independent of  $X_1$ , given  $Z$  and  $X_0$ . (In the twin network model there is a  $d$ -connected path from  $X_1$  to  $Y(x_0, x_1)$ , as follows:  $X_1 \leftarrow Z \leftrightarrow Z^* \rightarrow Y^*$ . Therefore, (3.62) is not satisfied for  $Y(x_0, x_1)$  and  $X_1$ .)

This example is another demonstration of the weakness of the potential-outcome language initially taken by Robins in deriving (3.63). The counterfactual condition (3.62) that legitimizes the use of the  $g$ -estimation formula is opaque, evoking no intuitive support. Epidemiologists who apply this formula are doing so under no guidance of substantive medical knowledge. Fortunately, graphical methods are rapidly making their way into epidemiological practice (Greenland et al. 1999a; Robins 2001; Hernán et al. 2002; Greenland and Brumback 2002; Kaufman et al. 2005; Petersen et al. 2006; VanderWeele and Robins 2007) as more and more researchers begin to understand the assumptions behind  $g$ -estimation. With the added understanding that structural equation models subsume, unify, and underlie the graphical, counterfactual, potential outcome and sufficient-component (Rothman 1976) approaches to causation,<sup>12</sup> epidemiology stands a good chance of becoming the first discipline to fully liberate itself from past dogmas and “break through our academically enforced reluctance to think directly about causes (Weinberg 2007).”

<sup>12</sup> This unification has not been sufficiently emphasized by leading epidemiologists (Greenland and Brumback 2002), economists (Heckman and Vytlačil 2007), and social scientists (Morgan and Winship 2007), not to mention statisticians (Cox and Wermuth 2004; Rubin 2005). With all due respect to multiculturalism, all approaches to causation are variants or abstractions of the structural equation model.



**Figure 11.12** A graph for which  $g$ -estimation is valid while Robins’ condition (3.62) is violated.

### 11.4 POLICY EVALUATION AND THE *do*-OPERATOR

#### 11.4.1 Identifying Conditional Plans (Section 4.2, p. 113)

**Question to Author:**

Section 4.2 of the book (p. 113) gives an identification condition and estimation formula for the effect of a conditional action, namely, the effect of an action  $do(X = g(z))$  where  $Z = z$  is a measurement taken prior to the action. Is this equation generalizable to the case of several actions, i.e., conditional plan?

The difficulty seen is that this formula was derived on the assumption that  $X$  does not change the value of  $Z$ . However, in a multi-action plan, some actions in  $X$  could change observations  $Z$  that guide future actions. We do not have notation for distinguishing post-intervention from pre-intervention observations. Absent such notation, it is not clear how conditional plans can be expressed formally and submitted to the *do*-calculus for analysis.

**Author’s Reply (with Ilya Shpitser):**

A notational distinction between post-intervention pre-intervention observations is introduced in Chapter 7 using the language of counterfactuals. The case of conditional plans, however, can be handled without resorting to richer notation. The reason is that the observations that dictate the choice of an action are not changed by that action, while those that have been changed by previous actions are well captured by the  $P(y | do(x), z)$  notation.

To see that this is the case, however, we will first introduce counterfactual notation, and then show that it can be eliminated from our expression. We will use bold letters to denote sets, and normal letters to denote individual elements. Also, capital letters will denote random variables, and small letters will denote possible values these variables could attain. We will write  $Y_x$  to mean ‘the value  $Y$  attains if we set variables  $X$  to values  $x$ .’ Similarly,  $Y_{X_g}$  is taken to mean ‘the value  $Y$  attains if we set variables  $X$  to whatever values they would have attained under the stochastic policy  $g$ .’ Note that  $Y_x$  and  $Y_{X_g}$  are both random variables, just as the original variable  $Y$ .

Say we have a set of  $K$  action variables  $X$  that occur in some temporal order. We will indicate the time at which a given variable is acted on by a superscript, so a variable  $X^i$  occurs before  $X^j$  if  $i < j$ . For a given  $X^i$ , we denote  $X^{<i}$  to be the set of action variables preceding  $X^i$ .

We are interested in the probability distribution of a set of outcome variables  $Y$ , under a policy that sets the values of each  $X^i \in X$  to the output of functions  $g_i$  (known in advance) which pay attention to some set of prior variables  $Z_i$ , as well as the previous interventions on  $X^{<i}$ . At the same time, the variables  $Z^i$  are themselves affected by previous interventions. To define this recursion appropriately, we use an inductive definition. The base case is  $X_g^1 = g_1(Z_1)$ . The inductive case is  $X_g^i = g_i(Z_{X_g^{<i}}, X_g^{<i})$ . Here the

subscript  $g$  represents the policy we use, in other words,  $g = \{g_i \mid i = 1, 2, \dots, K\}$ . We can now write the quantity of interest:

$$P(Y_{X_g} = y) = P(Y = y \mid do(X^1 = X_g^1), do(X^2 = X_g^2), \dots, do(X^K = X_g^K)).$$

Let  $Z_g = \cup_i Z_{X_g^i < i}$ . The key observation here is that if we observe  $Z_g$  to take on particular values,  $X_g$  collapse to unique values as well because  $X_g$  is a function of  $Z_g$ . We let  $x_z = \{x_z^1, \dots, x_z^K\}$  be the values attained by  $X_g$  in the situation where  $Z_g$  has been observed to equal  $z = \{z_1, \dots, z_K\}$ . We note here that if we know  $z$ , we can compute  $x_z$  in advance, because the functions  $g_i$  are fixed in advance and known to us. However, we don't know what values  $Z_g$  might obtain, so we use case analysis to consider all possible value combinations. We then obtain:

$$P(Y_{X_g} = y) = \sum_{z^1, \dots, z^K} P(Y = y \mid do(X = x_z), Z^1 = z^1, \dots, Z_{x_z < K}^K = z^K) \\ P(Z^1 = z^1, \dots, Z_{x_z < K}^K = z^K \mid do(X = x_z)).$$

Here we note that  $Z_i$  cannot depend on subsequent interventions. So we obtain

$$\sum_z P(Y = y \mid do(X = x_z), Z_{x_z}^1 = z^1, \dots, Z_{x_z}^K = z^K) P(Z^1 = z^1, \dots, Z^K = z^K \mid do(X = x_z)).$$

Now we note that the subscripts in the first and second terms are redundant, since the  $do(x_z)$  already implies such subscripts for all variables in the expression. Thus we can rewrite the target quantity as

$$\sum_z P(Y = y \mid do(X = x_z), Z^1 = z^1, \dots, Z^K = z^K) P(Z^1 = z^1, \dots, Z^K = z^K \mid do(X = x_z))$$

or, more succinctly,

$$\sum_z P(y \mid do(x_z), z) P(z \mid do(x_z)).$$

We see that we can compute this expression from  $P(y \mid do(x), z)$  and  $P(z \mid do(x))$ , where  $Y, X, Z$  are disjoint sets.

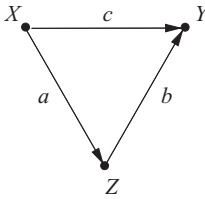
To summarize, though conditional plans are represented naturally by nested counterfactual expressions, their identification can nevertheless be reduced to identification of conditional interventional distributions of the form  $P(y \mid do(x), z)$  (possibly with  $z$  being empty). Complete conditions for identifying these quantities from a joint distribution in a given graph  $G$  are given in Shpitser and Pearl (2006a,b).

### 11.4.2 The Meaning of Indirect Effects

**Question to Author:**

I am teaching a course in latent variable modeling (to biostatistics and other public health students) and was yesterday introducing path analysis concepts, including direct and indirect effects.

I showed how to calculate indirect effects by taking the product of direct paths. Then a student asked about how to interpret the indirect effect, and I gave the answer that I always give, that the indirect effect  $ab$  (in the simple model of Fig. 11.13) is the effect that a change in  $X$  has on  $Y$  through its relationship with  $Z$ .



**Figure 11.13** Demonstrating an indirect effect of  $X$  on  $Y$  via  $Z$ .

After chewing on this for a second, the student asked the following:

**Student:** “The interpretation of the  $b$  path is:  $b$  is the increase we would see in  $Y$  given a unit increase in  $Z$  while holding  $X$  fixed, right?”

**Me:** “That’s right.”

**Student:** “Then what is being held constant when we interpret an indirect effect?”

**Me:** “Not sure what you mean.”

**Student:** “You said the interpretation of the indirect effect  $ab$  is:  $ab$  is the increase we would see in  $Y$  given a one unit increase in  $X$  through its causal effect on  $Z$ . But since  $b$  (the direct effect from  $Z$  to  $Y$ ) requires  $X$  to be held constant, how can it be used in a calculation that is also requiring  $X$  to change one unit.”

**Me:** “Hmm. Very good question. I’m not sure I have a good answer for you. In the case where the direct path from  $X$  to  $Y$  is zero, I think we have no problem, since the relationship between  $Z$  and  $Y$  then has nothing to do with  $X$ . But you are right, here if “ $c$ ” is nonzero then we must interpret  $b$  as the effect of  $Z$  on  $Y$  when  $X$  is held constant. I understand that this sounds like it conflicts with the interpretation of the  $ab$  indirect effect, where we are examining what a change in  $X$  will cause. How about I get back to you. As I have told you before, the calculations here aren’t hard, its trying to truly understand what your model means that’s hard.”

### ***Author’s Reply:***

Commend your student on his/her inquisitive mind. The answer can be formulated rather simply (see Section 4.5.5, which was appended to the second edition):

The indirect effect of  $X$  on  $Y$  is the increase we would see in  $Y$  while holding  $X$  constant and increasing  $Z$  to whatever value  $Z$  would attain under a unit increase of  $X$ .

Incidentally, the definition of  $b$  (the direct effect of  $Z$  on  $Y$ ) does not “require  $X$  to be held constant”; it requires merely that the increase in  $Z$  be produced by intervention, and not in response to other variations in the system. See discussion on p. 97 and equation (5.24) (p. 161).

### ***Author’s Afterthought:***

This question represents one of several areas where standard education in structural equation models (SEM) can stand reform. While some SEM textbooks give a cursory mention of the interpretation of structural parameters as effect coefficients, this interpretation is not taken very seriously by authors, teachers, and students. Writing in 2008, I find that the bulk of SEM education still focuses on techniques of statistical estimation and model fitting, and

one can hardly find a serious discussion of what the model means, once it is fitted and estimated (see Section 11.5.3 for SEM survival kit).<sup>13</sup>

The weakness of this educational tradition surfaces when inquisitive students ask questions that deviate slightly from standard LISREL routines, the answers to which hinge on the causal interpretation of structural coefficients and structural equations. For example:

1. Why should we define the total effect the way we do? (i.e., as the sum of products of certain direct effects). Is this an arbitrary definition, or is it compelled by the causal interpretation of the path coefficients?
2. Why should we define the indirect effect as the difference between the total and direct effects?
3. How can we define direct effect in nonlinear systems or in systems involving dichotomous variables?
4. How should we, in a meaningful way, define effects in systems involving feedback loops (i.e., reciprocal causation) so as to avoid the pitfalls of erroneous definitions quoted in SEM textbooks? (see p. 164)
5. Would our assessment of direct and total effects remain the same if we were to take some measurements prior to implementing the action whose effect we attempt to estimate?

Readers will be pleased to note that these questions can be given formal answers, as in Sections 4.5.4, 4.5.5, 11.5.2, 11.5.3, and 11.7.1.

On a personal note, my interest in direct and indirect effects was triggered by a message from Jacques Hageaars, who wrote (September 15, 2000): “Indirect effects do occupy an important place in substantive theories. Many social science theories ‘agree’ on the input (background characteristics) and output (behavioral) variables, but differ exactly with regard to the intervening mechanisms. To take a simple example, we know that the influence of Education on Political Preferences is mediated through ‘economic status’ (higher educated people get the better jobs and earn more money) and through a ‘cultural mechanism’ (having to do with the contents of the education and the accompanying socialization processes at school). We need to know and separate the nature and consequences of these two different processes, that is, we want to know the signs and the magnitudes of the indirect effects. In the parametric linear version of structural equation models, there exists a ‘calculus of path coefficients’ in which we can write total effects in terms of direct and several indirect effects. But this is not possible in the general nonparametric cases and not, e.g., in the log-linear parametric version. For systems of logic models there does not exist a comparable ‘calculus of path coefficients’ as has been remarked long ago. However, given its overriding theoretical importance, the issue of indirect effects cannot be simply neglected.”

Stimulated by these comments, and armed with the notation of nested counterfactuals, I set out to formalize the legal definition of hiring discrimination given on page 147, and

---

<sup>13</sup> The word “causal” does not appear in the index of any of the post-2000 SEM textbooks that I have examined.

was led to the results of Sections 4.5.4 and 4.5.5, some already anticipated by Robins and Greenland (1992). Enlightened by these results, I was compelled and delighted to retract an earlier statement made on page 165 of the first edition of *Causality*: “indirect effects lack intrinsic operational meaning” because they cannot be isolated using the  $do(x)$  operator. While it is true that indirect effects cannot be isolated using the  $do(x)$  operator, they do possess intrinsic operational meaning. Policy-making implications of direct and indirect effects are further exemplified in Pearl (2001) and Petersen et al. (2006).

### 11.4.3 Can $do(x)$ Represent Practical Experiments?

#### *Question to Author:*

L.B.S., from the University of Arizona, questioned whether the  $do(x)$  operator can represent realistic actions or experiments: “Even an otherwise perfectly executed randomized experiment may yield perfectly misleading conclusions. A good example is a study involving injected vitamin E as a treatment for incubated children at risk for retrolental fibroplasia. The randomized experiment indicated efficacy for the injections, but it was soon discovered that the actual effective treatment was opening the pressurized, oxygen-saturated incubators several times per day to give the injections, thus lowering the barometric pressure and oxygen levels in the blood of the infants (Leonard, *Major Medical Mistakes*). Any statistical analysis would have been misleading in that case.”

S.M., from Georgia Institute of Technology, adds:

“Your example of the misleading causal effect shows the kind of thing that troubles me about the  $do(x)$  concept. You  $do(x)$  or don’t  $do(x)$ , but it may be something else that covaries with  $do(x)$  that is the cause and not the  $do(x)$  per se.”

#### *Author’s Reply:*

Mathematics deals with ideal situations, and it is the experimenter’s job to make sure that the experimental conditions approximate the mathematical ideal as closely as possible. The  $do(x)$  operator stands for doing  $X = x$  in an ideal experiment, where  $X$  and  $X$  alone is manipulated, not any other variable in the model.

In your example of the vitamin E injection, there is another variable being manipulated together with  $X$ , namely, the incubator cover,  $Z$ , which turns the experiment into a  $do(x, z)$  condition instead of  $do(x)$ . Thus, the experiment was far from ideal, and far even from the standard experimental protocol, which requires the use of a placebo. Had a placebo been used (to approximate the requirement of the  $do(x)$  operator), the result would not have been biased.

There is no way a model can predict the effect of an action unless one specifies what variables in the model are affected by the action, and how. The  $do(x)$  operator is a mathematical device that helps us specify explicitly and formally what is held constant, and what is free to vary in any given experiment. The  $do$ -calculus then helps us predict the logical ramifications of such specifications, assuming they are executed faithfully, assuming we have a valid causal model of the environment, and assuming we have data from other experiments conducted under well-specified conditions.

### 11.4.4 Is the *do*(*x*) Operator Universal?

#### *Question to Author (from Bill Shipley)*

In most experiments, the external manipulation consists of adding (or subtracting) some amount from *X* without removing preexisting causes of *X*. For example, adding 5 kg/h of fertilizer to a field, adding 5 mg/l of insulin to subjects, etc. Here, the preexisting causes of the manipulated variable still exert effects, but a new variable (*M*) is added.

The problem that I see with the *do*(*x*) operator as a general operator of external manipulation is that it requires two things: (1) removing any preexisting causes of *x* and (2) setting *x* to some value. This corresponds to some types of external manipulation, but not to all (or even most) external manipulations. I would introduce an *add*(*x* = *n*) operator, meaning “add, external to the preexisting causal process, an amount ‘*n*’ of *x*.” Graphically, this consists of augmenting the preexisting causal graph with a new edge, namely,  $M - n \rightarrow X$ . Algebraically, this would consist of adding a new term *-n* as a cause of *X* (Shipley 2000b).

#### *Author’s Answer:*

In many cases, your “additive intervention” indeed represents the only way we can intervene on a variable *X*; in others, it may represent the actual policy we wish evaluated. In fact, the general notion of intervention (p. 113) permits us to replace the equation of *X* by any other equation that fits the circumstances, not necessarily a constant  $X = x$ .

What you are proposing corresponds to replacing the old equation of *X*,  $x = f(pa_X)$ , with a new equation:  $x = f(pa_X) + n$ . This replacement can be represented using “instrumental variables,” since it is equivalent to writing  $x = f(pa_X) + I$  (where *I* is an instrument) and varying *I* from 0 to *n*.

There are three points to notice:

1. The additive manipulation *can* be represented in the *do*( ) framework – we merely apply the *do*( ) operator to the instrument *I*, and not to *X* itself. This is a different kind of manipulation that needs to be distinguished from *do*(*x*) because, as is shown below, the effect on *Y* may be different.
2. In many cases, scientists are not satisfied with estimating the effect of the instrument on *Y*, but are trying hard to estimate the effect of *X* itself, which is often more meaningful or more transportable to other situations. (See p. 261 for discussion of the effect of “intention to treat.”)
3. Consider the nonrecursive example where LISREL fails  $y = bx + e_1 + I, x = ay + e_2$ , (p. 164). If we interpret “total effects” as the response of *Y* to a unit change of the instrument *I*, then LISREL’s formula obtains: The effect of *I* on *Y* is  $b/(1 - ab)$ . However, if we adhere to the notion of “per unit change in *X*,” we get back the *do*-formula: The effect of *X* on *Y* is *b*, not  $b/(1 - ab)$ , even though the manipulation is done through an instrument. In other words, we change *I* from 0 to 1 and observe the changes in *X* and in *Y*; if we divide the change in *Y* by the change in *X*, we get *b*, not  $b/(1 - ab)$ .

To summarize: Yes, additive manipulation is sometimes what we need to model, and it can be done in the *do*(*x*) framework using instrumental variables. We still need to

distinguish, though, between the effect of the instrument and the effect of  $X$ . The former is not stable (p. 261), the latter is. LISREL's formula corresponds to the effect of an instrument, not to the effect of  $X$ .

***Bill Shipley Further Asked:***

Thanks for the clarification. It seems to me that the simplest, and most straightforward, way of modeling and representing manipulations of a causal system is to simply (1) modify the causal graph of the unmanipulated system to represent the proposed manipulation, (2) translate this new graph into structural equations, and (3) derive predictions (including conditional predictions) from the resulting equations; this is how I have treated the notion in my book. Why worry about  $do(x)$  at all? In particular, one can model quite sophisticated manipulations this way. For instance, one might well ask what would happen if one added an amount  $z$  to some variable  $x$  in the causal graph, in which  $z$  is dependent on some other variable in the graph.

***Author's Reply:***

The method you are proposing, to replace the current equation  $x = f(pa_X)$  with  $x = g(f(pa_X), I, z)$ , requires that we know the functional forms of  $f$  and  $g$ , as in linear systems or, alternatively, that the parents of  $X$  are observed, as in the Process Control example on page 74. These do not hold, however, in the non-parametric, partially observable settings of Chapters 3 and 4, which might render it impossible to predict the effect of the proposed intervention from data gathered prior to the intervention, a problem we called *identification*. Because pre-intervention statistics is not available for variable  $I$ , and  $f$  is unknown, there are semi-Markovian cases where  $P(y | do(x))$  is identifiable while  $P(y | do(x = g(f(pa_X), I, z)))$  is not; each case must be analyzed on its own merits. It is important, therefore, to impose certain standards on this vast space of potential interventions, and focus attention on those that could illuminate others.

Science thrives on standards, because standards serve (at least) two purposes: communication and theoretical focus. Mathematicians, for example, have decided that the derivative operator " $dy/dx$ " is a nice standard for communicating information about change, so that is what we teach in calculus, although other operators might also serve the purpose, for example,  $xdy/dx$  or  $(dy/dx)/y$ , etc. The same applies to causal analysis:

1. **Communication:** If we were to eliminate the term "treatment effect" from epidemiology, and replace it with detailed descriptions of how the effect was measured, we would practically choke all communication among epidemiologists. A standard was therefore established: what we measure in a controlled, randomized experiment will be called "treatment effect"; the rest will be considered variations on the theme. The "*do*-operator" represents this standard faithfully.

The same goes for SEM. Sewall Wright talked about "effect coefficients" and established them as the standard of "direct effect" in path analysis (before it got molested with regression jargon), with the help of which more elaborate effects can be constructed. Again, the "*do*-operator" is the basis for defining this standard.

2. **Theoretical focus:** Many of the variants of manipulations can be reduced to "*do*," or to several applications of "*do*." Theoretical results established for "*do*"

are then applicable to those variants. Example: your “add  $n$ ” manipulation is expressible as “*do*” on an instrument. Another example: questions of identification for expressions involving “*do*” are applicable to questions of identification for more sophisticated effects. On page 113, for example, we show that if the expression  $P(y \mid do(x), z)$  is identifiable, then so also is the effect of conditional actions  $P(y \mid do(x = g(z) \text{ if } Z = z))$ . The same goes for many other theoretical results in the book; they were developed for the “*do*”-operator, they borrow from each other, and they are applicable to many variants.

Finally, the “surgical” operation underlying the *do*-operator provides the appropriate formalism for interpreting counterfactual sentences (see p. 204), and counterfactuals are abundant in scientific discourse (see pp. 217–19). I have yet to see any competing candidate with comparable versatility, generality, formal power, and (not the least) conceptual appeal.

### 11.4.5 Causation without Manipulation!!!

#### *Question to Author*

In the analysis of direct effects, Section 4.5 invokes an example of sex discrimination in school admission and, in several of the formulas, gender is placed under the “hat” symbol or, equivalently, as an argument of the *do*-operator. How can gender be placed after the *do*-operator when it is a variable that cannot be manipulated?

#### *Author’s Reply*

Since Holland coined the phrase “No Causation without Manipulation” (Holland 1986), many good ideas have been stifled or dismissed from causal analysis. To suppress talk about how gender causes the many biological, social, and psychological distinctions between males and females is to suppress 90% of our knowledge about gender differences.

Surely we have causation without manipulation. The moon causes tides, race causes discrimination, and sex causes the secretion of certain hormones and not others. Nature is a society of mechanisms that relentlessly sense the values of some variables and determine the values of others; it does not wait for a human manipulator before activating those mechanisms.

True, manipulation is one way (albeit a crude one) for scientists to test the workings of mechanisms, but it should not in any way inhibit causal thoughts, formal definitions, and mathematical analyses of the mechanisms that propel the phenomena under investigation. It is for that reason, perhaps, that scientists invented counterfactuals; it permits them to state and conceive the realization of antecedent conditions without specifying the physical means by which these conditions are established.

The purpose of the “hat” symbol in Definition 4.5.1 is not to stimulate thoughts about possible ways of changing applicants’ gender, but to remind us that any definition concerned with “effects” should focus on causal links and filter out spurious associations from the quantities defined. True, in the case of gender, one can safely replace  $P(y \mid do(\widehat{female}))$  with  $P(y \mid female)$ , because the mechanism determining gender can safely be assumed to be independent of the background factors that influence  $Y$  (thus

ensuring no confounding). But as a general definition, and even as part of an instructive example, mathematical expressions concerned with direct effects and sex discrimination should maintain the hat symbol. If nothing else, placing “female” under the “hat” symbol should help propagate the long-overdue counter-slogan: “Causation without manipulation? You bet!”

#### 11.4.6 Hunting Causes with Cartwright

In her book *Hunting Causes and Using Them* (Cambridge University Press, 2007), Nancy Cartwright expresses several objections to the  $do(x)$  operator and the “surgery” semantics on which it is based (p. 72, p. 201). In so doing, she unveils several areas in need of systematic clarification; I will address them in turn.

Cartwright description of surgery goes as follows:

Pearl gives a precise and detailed semantics for counterfactuals. But what is the semantics a semantics of? The particular semantics Pearl develops is unsuited to a host of natural language uses of counterfactuals, especially those for planning and evaluation of the kind I have been discussing. That is because of the special way in which he imagines that the counterfactual antecedent will be brought about: by a precise incision that changes exactly the counterfactual antecedent and nothing else (except what follows causally from just that difference). But when we consider implementing a policy, this is not at all the question we need to ask. For policy and evaluation we generally want to know what would happen were the policy really set in place. And whatever we know about how it might be put in place, the one thing we can usually be sure of is that it will not be by a precise incision of the kind Pearl assumes.

Consider for example Pearl’s axiom of composition, which he proves to hold in all causal models – given his characterization of a causal model and his semantics for counterfactuals. This axiom states that ‘if we force a variable ( $W$ ) to a value  $w$  that it would have had without our intervention, then the intervention will have no effect on other variables in the system’ (p. 229). This axiom is reasonable if we envisage interventions that bring about the antecedent of the counterfactual in as minimal a way as possible. But it is clearly violated in a great many realistic cases. Often we have no idea whether the antecedent will in fact obtain or not, and this is true even if we allow that the governing principles are deterministic. We implement a policy to ensure that it will obtain – and the policy may affect a host of changes in other variables in the system, some envisaged and some not. (Cartwright 2007, pp. 246–7)

Cartwright’s objections can thus be summarized in three claims; each will be addressed separately.

1. In most studies we need to predict the effect of nonatomic interventions.
2. For policy evaluation “we generally want to know what would happen were the policy really set in place,” but, unfortunately, “the policy may affect a host of changes in other variables in the system, some envisaged and some not.”
3. Because practical policies are nonatomic, they cannot be evaluated from the atomic semantics of the  $do(x)$  calculus even if we *could* envisage the variables that are affected by the policy.

Let us start with claim (2) – the easiest one to disprove. This objection is identical to the one discussed in Section 11.4.3, the answer to which was: “There is no way a model can predict the effect of an action unless one specifies correctly what variables in the model are affected by the action, and how.” In other words, under the state of ignorance described in claim (2) of Cartwright, a policy evaluation study must end with a trivial answer: There is not enough information, hence, anything can happen. It is like pressing an unknown button in the dark, or trying to solve two equations with three unknowns. Moreover, the *do*-calculus can be used to test whether the state of ignorance in any given situation should justify such a trivial answer. Thus, it would be a mistake to assume that serious policy evaluation studies are conducted under such a state of ignorance; all policy analyses I have seen commence by assuming knowledge of the variables affected by the policy, and expressing that knowledge formally.

Claim (1) may apply in some cases, but certainly not in most; in many studies our goal is not to predict the effect of the crude, nonatomic intervention that we are about to implement but, rather, to evaluate an ideal, atomic policy that cannot be implemented given the available tools, but that represents nevertheless a theoretical relationship that is pivotal for our understanding of the domain.

An example will help. Smoking cannot be stopped by any legal or educational means available to us today; cigarette advertising can. That does not stop researchers from aiming to estimate “the effect of smoking on cancer,” and doing so from experiments in which they vary the instrument – cigarette advertisement – not smoking.

The reason they would be interested in the atomic intervention  $P(\text{cancer} \mid \text{do}(\text{smoking}))$  rather than (or in addition to)  $P(\text{cancer} \mid \text{do}(\text{advertising}))$  is that the former represents a stable biological characteristic of the population, uncontaminated by social factors that affect susceptibility to advertisement. With the help of this stable characteristic one can assess the effects of a wide variety of practical policies, each employing a different smoking-reduction instrument.

Finally, claim (3) is demonstratively disproved in almost every chapter of this book. What could be more nonatomic than a policy involving a sequence of actions, each chosen by a set of observations  $Z$  which, in turn, are affected by previous actions (see Sections 4.4 and 11.4.1)? And yet the effect of implementing such a complex policy can be predicted using the “surgical” semantics of the *do*-calculus in much the same way that properties of complex molecules can be predicted from atomic physics.

I once challenged Nancy Cartwright (Pearl 2003a), and I would like to challenge her again, to cite a single example of a policy that cannot either be specified and analyzed using the  $\text{do}(x)$  operators, or proven “unpredictable” (e.g., pressing an unknown button in the dark), again using the calculus of  $\text{do}(x)$  operators.

Ironically, shunning mathematics based on ideal atomic intervention may condemn scientists to ineptness in handling realistic non-atomic interventions.

Science and mathematics are full of auxiliary abstract quantities that are not directly measured or tested, but serve to analyze those that are. Pure chemical elements do not exist in nature, yet they are indispensable to the understanding of alloys and compounds. Negative numbers (let alone imaginary numbers) do not exist in isolation, yet they are essential for the understanding of manipulations on positive numbers.

The broad set of problems tackled (and solved) in this book testifies that, invariably, questions about interventions and experimentation, ideal as well as non-ideal, practical

as well as epistemological, can be formulated precisely and managed systematically using the atomic intervention as a primitive notion.

#### 11.4.7 The Illusion of Nonmodularity

In her critique of the *do*-operator, Cartwright invokes yet another argument – the failure of modularity, which allegedly plagues most mechanical and social systems.

In her words:

“When Pearl talked about this recently at LSE he illustrated this requirement with a Boolean input-output diagram for a circuit. In it, not only could the entire input for each variable be changed independently of that for each other, so too could each Boolean component of that input. But most arrangements we study are not like that. They are rather like a toaster or a carburetor.”

At this point, Cartwright provides a four-equation model of a car carburetor and concludes:

The gas in the chamber is the result of the pumped gas and the gas exiting the emulsion tube. How much each contributes is fixed by other factors: for the pumped gas both the amount of airflow and a parameter  $a$ , which is partly determined by the geometry of the chamber; and for the gas exiting the emulsion tube, by a parameter  $a'$ , which also depends on the geometry of the chamber. The point is this. In Pearl's circuit-board, there is one distinct physical mechanism to underwrite each distinct causal connection. But that is incredibly wasteful of space and materials, which matters for the carburetor. One of the central tricks for an engineer in designing a carburetor is to ensure that one and the same physical design – for example, the design of the chamber – can underwrite or ensure a number of different causal connections that we need all at once.

Just look back at my diagrammatic equations, where we can see a large number of laws all of which depend on the same physical features – the geometry of the carburetor. So no one of these laws can be changed on its own. To change any one requires a redesign of the carburetor, which will change the others in train. By design the different causal laws are harnessed together and cannot be changed singly. So modularity fails. (Cartwright 2007, pp. 15–16)

Thus, for Cartwright, a set of equations that share parameters is inherently nonmodular; changing one equation means modifying at least one of its parameters, and if this parameter appears in some other equation, it must change as well, in violation of modularity.

Heckman (2005, p. 44) makes similar claims: “Putting a constraint on one equation places a restriction on the entire set of internal variables.” “Shutting down one equation might also affect the parameters of the other equations in the system and violate the requirements of parameter stability.”

Such fears and warnings are illusory. Surgery, and the whole semantics and calculus built around it, does not assume that in the physical world we have the technology to incisively modify the mechanism behind each structural equation while leaving all others unaltered. Symbolic modularity does not assume physical modularity. Surgery is a symbolic operation which makes no claims about the physical means available to the experimenter, or about possible connections that might exist between the mechanisms involved.

Symbolically, one can surely change one equation without altering others and proceed to define quantities that rest on such “atomic” changes. Whether the quantities defined in this manner correspond to changes that can be physically realized is a totally different question that can only be addressed once we have a formal description of the interventions available to us. More importantly, shutting down an equation does not necessarily mean meddling with its parameters; it means overruling that equation, namely, leaving the equation intact but lifting the outcome variable from its influence.

A simple example will illustrate this point.

Assume we have two objects under free-fall conditions. The respective accelerations,  $a_1$  and  $a_2$ , of the two objects are given by the equations:

$$a_1 = g \tag{11.12}$$

$$a_2 = g, \tag{11.13}$$

where  $g$  is the earth’s gravitational pull. The two equations share a parameter,  $g$ , and appear to be nonmodular in Cartwright’s sense; there is indeed no physical way of changing the gravitational force on one object without a corresponding change on the other. However, this does not mean that we cannot intervene on object 1 without touching object 2. Assume we grab object 1 and bring it to a stop. Mathematically, the intervention amounts to replacing equation (11.12) by

$$a_1 = 0 \tag{11.14}$$

while leaving equation (11.13) intact. Setting  $g$  to zero in equation (11.12) is a symbolic surgery that does not alter  $g$  in the physical world but, rather, sets  $a_1$  to 0 by bringing object 1 under the influence of a new force,  $f$ , emanating from our grabbing hand. Thus, equation (11.14) is a result of two forces:

$$a_1 = g + f/m_1, \tag{11.15}$$

where  $f = -gm_1$ , which is identical to (11.14).

This same operation can be applied to Cartwright carburetor; for example, the gas outflow can be fixed without changing the chamber geometry by installing a flow regulator at the emulsion tube. It definitely applies to economic systems, where human agents are behind most of the equations; the left-hand side of the equations can be fixed by exposing agents to different information, rather than by changing parameters in the physical world. A typical example emerges in job discrimination cases (Section 4.5.3). To test the “effect of gender on hiring” one need not physically change the applicant’s gender; it is enough to change the employer’s awareness of the applicant’s gender. I have yet to see an example of an economic system which is not modular in the sense described here.

This operation of adding a term to the right-hand side of an equation to ensure constancy of the left-hand side is precisely how Haavelmo (1943) envisioned surgery in economic settings. Why his wisdom disappeared from the teachings of his disciples in 2008 is one of the great mysteries of economics (see Hoover (2004)); my theory remains (p. 138) that it all happened due to a careless choice of notation which crumbled under the ruthless invasion of statistical thinking in the early 1970s.

More on the confusion in econometrics and the reluctance of modern-day econometricians to reinstate Haavelmo’s wisdom is to be discussed in Section 11.5.4.

## 11.5 CAUSAL ANALYSIS IN LINEAR STRUCTURAL MODELS

### 11.5.1 General Criterion for Parameter Identification (Chapter 5, pp. 149–54)

#### *Question to Author:*

The parameter identification method described in Section 5.3.1 rests on repetitive applications of two basic criteria: (1) the single-door criterion of Theorem 5.3.1, and (2) the back-door criterion of Theorem 5.3.2. This method may require appreciable bookkeeping in combining results from various segments of the graph. Is there a single graphical criterion of identification that unifies the two theorems and thus avoids much of the bookkeeping involved?

#### *Author's Reply:*

A unifying criterion is described in the following lemma (Pearl 2004):

#### **Lemma 11.5.1** (*Graphical identification of direct effects*)

Let  $c$  stand for the path coefficient assigned to the arrow  $X \rightarrow Y$  in a causal graph  $G$ . Parameter  $c$  is identified if there exists a pair  $(W, Z)$ , where  $W$  is a single node in  $G$  (not excluding  $W = X$ ), and  $Z$  is a (possibly empty) set of nodes in  $G$ , such that:

1.  $Z$  consists of nondescendants of  $Y$ ,
2.  $Z$   $d$ -separates  $W$  from  $Y$  in the graph  $G_c$  formed by removing  $X \rightarrow Y$  from  $G$ ,
3.  $W$  and  $X$  are  $d$ -connected, given  $Z$ , in  $G_c$ .

Moreover, the estimand induced by the pair  $(W, Z)$  is given by:

$$c = \frac{\text{cov}(Y, W \mid Z)}{\text{cov}(X, W \mid Z)}.$$

The intuition is that, conditional on  $Z$ ,  $W$  acts as an instrumental variable relative to  $X \rightarrow Y$ . See also McDonald (2002a).

More general identification methods are reported in Brito and Pearl (2002a,b,c; 2006).

### 11.5.2 The Causal Interpretation of Structural Coefficients

#### *Question to Author:*

In response to assertions made in Sections 5.1 and 5.4 that a correct causal interpretation is conspicuously absent from SEM books and papers, including all 1970–99 texts in economics, two readers wrote that the “unit-change” interpretation is common and well accepted in the SEM literature. L.H. from the University of Alberta wrote:

Page 245 of L. Hayduk, *Structural Equation Modeling with LISREL: Essentials and Advances*, 1987, [states] that a slope can be interpreted as: the magnitude of the change in  $y$  that would be predicted to accompany a unit change in  $x$  with the other variables in the equation left untouched at their original values.

O.D. Duncan, *Introduction to Structural Equation Models* (1975) pages 1 and 2 are pretty clear on  $b$  as causal. More precisely, it says that “a change of one unit in  $x$  ... produces a change of  $b$  units in  $y$ ” (page 2). I suspect that H. M. Blalock’s book

‘Causal models in the social sciences,’ and D. Heise’s book ‘Causal analysis’ probably speak of  $b$  as causal.

S.M., from Georgia Tech, concurs:

Heise, author of *Causal Analysis* (1975), regarded the  $b$  of causal equations to be how much a unit change in a cause produced an effect in an effect variable. This is a well-accepted idea.

***Author’s Reply:***

The “unit change” idea appears, sporadically, in several SEM publications, yet, invariably, its appearance is marred by omissions, ambiguities, and subsequent misstatements that obscure the causal character of this idea and its ramifications.

The paragraph cited above (from Hayduk 1987) can serve to illustrate how the unit change idea is typically introduced in the SEM literature and how it *should be* introduced using modern understanding of causal modeling. The original paragraph reads:

The interpretation of structural coefficients as “effect coefficients” originates with ordinary regression equations like

$$X_0 = a + b_1X_1 + b_2X_2 + b_3X_3 + e$$

for the effects of variables  $X_1$ ,  $X_2$ , and  $X_3$  on variable  $X_0$ . We can interpret the estimate of  $b_1$  as the magnitude of the change in  $X_0$  that would be predicted to accompany a unit change INCREASE in  $X_1$  with  $X_2$  and  $X_3$  left untouched at their original values. We avoid ending with the phrase “held constant” because this phrase must be abandoned for models containing multiple equations, as we shall later see. Parallel interpretations are appropriate for  $b_2$  and  $b_3$ . (Hayduk 1987, p. 245)

This paragraph illustrates how two basic distinctions are often conflated in the SEM literature. The first is the distinction between structural coefficients and regressional (or statistical) estimates of those coefficients. We rarely find the former defined independently of the latter – a confusion that is rampant and especially embarrassing in econometric texts. The second is the distinction between “held constant” and “left untouched” or “found to remain constant,” for which the  $do(x)$  operator was devised. (Related distinctions: “doing” versus “seeing” and “interventional change” versus “natural change.”)

To emphasize the centrality of these distinctions I will now propose a concise revision of Hayduk’s paragraph:

***Proposed Revised Paragraph***

The interpretation of structural coefficients as “effect coefficients” bears some resemblance to, but differs fundamentally from, the interpretation of coefficients in regression equations like

$$X_0 = a + b_1X_1 + b_2X_2 + b_3X_3 + e. \tag{11.16}$$

If (11.16) is a regression equation, then  $b_1$  stands for the change in  $X_0$  that would be predicted to accompany a unit change in  $X_1$  in those situations where  $X_2$  and  $X_3$  remain

constant at their original values. We formally express this interpretation using conditional expectations:

$$\begin{aligned} b_1 &= E(X_0 \mid x_1 + 1, x_2, x_3) - E(X_0 \mid x_1, x_2, x_3) \\ &= R_{X_0 X_1 \cdot X_2 X_3}. \end{aligned} \quad (11.17)$$

Note that, as a regression equation, (11.16) is claimless; i.e., it cannot be falsified by any experiment and, from (11.17),  $e$  is automatically rendered uncorrelated with  $X_1$ ,  $X_2$ , and  $X_3$ .

In contrast, if equation (11.16) represents a structural equation, it makes empirical claims about the world (e.g., that other variables in the system do not affect  $X_0$  once we hold  $X_1$ ,  $X_2$ , and  $X_3$  fixed), and the interpretation of  $b_1$  must be modified in two fundamental ways. First, the phrase “a unit change in  $X_1$ ” must be qualified to mean “a unit interventional change in  $X_1$ ,” thus ruling out changes in  $X_1$  that are produced by other variables in the model (possibly correlated with  $e$ ). Second, the phrase “where  $X_2$  and  $X_3$  remain constant” must be abandoned and replaced by the phrase “if we hold  $X_2$  and  $X_3$  constant,” thus ensuring constancy even when  $X_2$  is affected by  $X_1$ .

Formally, these two modifications are expressed as:

$$b_1 = E(X_0 \mid do(x_1 + 1, x_2, x_3)) - E(X_0 \mid do(x_1, x_2, x_3)). \quad (11.18)$$

The phrase “left untouched at their original values” may lead to ambiguities. Leaving variables untouched permits those variables to vary (e.g., in response to the unit increase in  $X_1$  or other influences), in which case the change in  $X_0$  would correspond to the total effect

$$E(X_0 \mid do(x_1 + 1)) - E(X_0 \mid do(x_1)) \quad (11.19)$$

or to the marginal conditional expectation

$$E(X_0 \mid x_1 + 1) - E(X_0 \mid x_1), \quad (11.20)$$

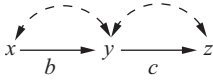
depending on whether the change in  $X_1$  is interventional or observational. None of (11.19) or (11.20) matches the meaning of  $b_1$  in equation (11.16), regardless of whether we treat (11.16) as a structural or a regression equation.

The interpretation expressed in (11.18) holds in *all* models, including those containing multiple equations, recursive and nonrecursive, regardless of whether  $e$  is correlated with other variables in the model and regardless of whether  $X_2$  and  $X_3$  are affected by  $X_1$ . In contrast, expression (11.17) coincides with (11.18) only under very special circumstances (defined by the single-door criterion of Theorem 5.3.1). It is for this reason that we consider (11.18), not (11.17), to be an “interpretation” of  $b_1$ ; (11.17) interprets the “regression estimate” of  $b_1$  (which might well be biased), while (11.18) interprets  $b_1$  itself.

### 11.5.3 Defending the Causal Interpretation of SEM (or, SEM Survival Kit)

#### *Question to Author:*

J. Wilson from Surrey, England, asked about ways of defending his Ph.D. thesis before examiners who do not approve of the causal interpretation of structural equation models (SEM). He complained about “the complete lack of emphasis in PhD programmes on



**Figure 11.14** The graph underlying equations (11.21)–(11.22).

how to defend causal interpretations and policy implications in a viva when SEM is used ... if only causality had been fully explained at the beginning of the programme, then each of the 70,000 words used in my thesis would have been carefully measured to defend first the causal assumptions, then the data, and finally the interpretations ... (I wonder how widespread this problem is?) Back to the present and urgent task of trying to satisfy the examiners, especially those two very awkward Stat Professors – they seem to be trying to outdo each other in nastiness.”

### **Author’s Reply:**

The phenomenon that you complain about is precisely what triggered my writing of Chapter 5 – the causal interpretation of SEM is still a mystery to most SEMs researchers, leaders, educators, and practitioners. I have spent hours on SEMNET Discussion List trying to rectify the current neglect, but it is only students like yourself who can turn things around and help reinstate the causal interpretation to its central role in SEM research.

As to your concrete question – how to defend the causal interpretation of SEM against nasty examiners who oppose such interpretation – permit me to assist by sketching a hypothetical scenario in which you defend the causal interpretation of your thesis in front of a hostile examiner, Dr. EX. (Any resemblance to Dr. EX is purely coincidental.)

### **A Dialogue with a Hostile Examiner or SEM Survival Kit**

For simplicity, let us assume that the model in your thesis consists of just two-equations,

$$y = bx + e_1 \quad (11.21)$$

$$z = cy + e_2, \quad (11.22)$$

with  $e_2$  uncorrelated with  $x$ . The associated diagram is given in Figure 11.14. Let us further assume that the target of your thesis was to estimate parameter  $c$ , that you have estimated  $c$  satisfactorily to be  $c = 0.78$  using the best SEM methods, and that you have given a causal interpretation to your finding.

Now your nasty examiner, Dr. EX, arrives and questions your interpretation.

**Dr. EX:** What do you mean by “ $c$  has a causal interpretation”?

**You:** I mean that a unit change in  $y$  will bring about a  $c$  units change in  $E(Z)$ .

**Dr. EX:** The words “change” and “bring about” make me uncomfortable; let’s be scientific. Do you mean  $E(Z|y) = cy + a$ ?? I can understand this last expression, because the conditional expectation of  $Z$  given  $y$ ,  $E(Z|y)$ , is well defined mathematically, and I know how to estimate it from data. But “change” and “bring about” is jargon to me.

**You:** I actually mean “change,” not “an increase in conditional expectation,” and by “change” I mean the following: If we had the physical means of fixing  $y$  at some

constant  $y_1$ , and of changing that constant from  $y_1$  to  $y_2$ , then the observed change in  $E(Z)$  would be  $c(y_2 - y_1)$ .

**Dr. EX:** Well, well, aren't we getting a bit metaphysical here? I never heard about "fixing" in my statistics classes.

**You:** Oh, sorry, I did not realize you have statistics background. In that case, let me rephrase my interpretation a bit, to read as follows: If we had the means of conducting a controlled randomized experiment, with  $y$  randomized, then if we set the control group to  $y_1$  and the experimental group to  $y_2$ , the observed difference in  $E(Z)$  would be  $E(Z_2) - E(Z_1) = c(y_2 - y_1)$  regardless of what values  $y_1$  and  $y_2$  we choose. ( $Z_1$  and  $Z_2$  are the measurements of  $z$  under the control and experimental groups, respectively.)<sup>14</sup>

**Dr. EX:** That sounds much closer to what I can understand. But I am bothered by a giant leap that you seem to be making. Your data was nonexperimental, and in your entire study you have not conducted a single experiment. Are you telling us that your SEM exercise can take data from an observational study, do some LISREL analysis on it, and come up with a prediction of what the outcome of a controlled randomized experiment will be? You've got to be kidding!! Do you know how much money can be saved nationwide if we could replace experimental studies with SEM magic?

**You:** This is not magic, Dr. EX, it is plain logic. The input to my LISREL analysis was more than just nonexperimental data. The input consisted of two components: (1) data, (2) causal assumptions; my conclusion logically follows from the two. The second component is absent in standard experimental studies, and that is what makes them so expensive.

**Dr. EX:** What kind of assumptions? "Causal"? I never heard of such strangers. Can you express them mathematically the way we normally express assumptions – say, in the form of conditions on the joint density, or properties of the covariance matrix?

**You:** Causal assumptions are of a different kind; they cannot be written in the vocabulary of density functions or covariance matrices. Instead, they are expressed in my causal model.

**Dr. EX:** Looking at your model, equations (11.21)–(11.22), I do not see any new vocabulary; all I see is equations.

**You:** These are not ordinary algebraic equations, Dr. EX. These are "structural equations," and if we read them correctly, they convey a set of assumptions with which you are familiar, namely, assumptions about the outcomes of hypothetical randomized experiments conducted on the population – we call them "causal" or "modeling" assumptions, for want of better words, but they can be understood as assumptions about the behavior of the population under various randomized experiments.

**Dr. EX:** Wait a minute! Now that I begin to understand what your causal assumptions are, I am even more puzzled than before. If you allow yourself to make assumptions about the behavior of the population under randomized experiments, why go through the trouble of conducting a study? Why not make the assumption directly that in a randomized experiment, with  $y$  randomized, the observed difference in  $E(Z)$  should be  $c'(y_2 - y_1)$ , with  $c'$  just any convenient number, and save yourself agonizing months of data collection and analysis. He who believes your other untested assumptions should also believe your  $E(Z_2) - E(Z_1) = c'(y_2 - y_1)$  assumption.

<sup>14</sup> Just in case Dr. EX asks: "Is that the only claim?" you should add: Moreover, I claim that the distribution of the random variable  $Z_1 - cy_1$  will be the same as that of the variable  $Z_2 - cy_2$ .

**You:** Not so, Dr. EX. The modeling assumptions with which my program begins are much milder than the assertion  $E(Z_2) - E(Z_1) = 0.78(y_2 - y_1)$  with which my study concludes. First, my modeling assumptions are qualitative, while my conclusion is quantitative, making a commitment to a specific value of  $c = 0.78$ . Second, many researchers (including you, Dr. EX) would be prepared to accept my assumptions, not my conclusion, because the former conforms to commonsense understanding and general theoretical knowledge of how the world operates. Third, the majority of my assumptions can be tested by experiments that do not involve randomization of  $y$ . This means that if randomizing  $y$  is expensive, or infeasible, we still can test the assumptions by controlling other, less formidable variables. Finally, though this is not the case in my study, modeling assumptions often have some statistical implications that can be tested in nonexperimental studies, and, if the test turns out to be successful (we call it “fit”), it gives us further confirmation of the validity of those assumptions.

**Dr. EX:** This is getting interesting. Let me see some of those “causal” or modeling assumptions, so I can judge how mild they are.

**You:** That’s easy, have a look at our model, Figure 11.14, where

- $z$  – student’s score on the final exam,
- $y$  – number of hours the student spent on homework,
- $x$  – weight of homework (as announced by the teacher) in the final grade.

When I put this model down on paper, I had in mind two randomized experiments, one where  $x$  is randomized (i.e., teachers assigning weight at random), the second where the actual time spent on homework ( $y$ ) is randomized. The assumptions I made while thinking of those experiments were:

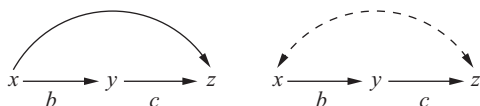
1. Linearity and exclusion for  $y$ :  $E(Y_2) - E(Y_1) = b(x_2 - x_1)$ , with  $b$  unknown ( $Y_2$  and  $Y_1$  are the time that would be spent on homework under announced weights  $x_2$  and  $x_1$ , respectively.) Also, by excluding  $z$  from the equation, I assumed that the score  $z$  would not affect  $y$ , because  $z$  is not known at the time  $y$  is decided.
2. Linearity and exclusion for  $z$ :  $E(Z_2) - E(Z_1) = c(y_2 - y_1)$  for all  $x$ , with  $c$  unknown. In words,  $x$  has no effect on  $z$ , except through  $y$ .

In addition, I made qualitative assumptions about unmeasured factors that govern  $x$  under nonexperimental conditions; I assumed that there are no common causes for  $x$  and  $z$ .

Do you, Dr. EX, see any objection to any of these assumptions?

**Dr. EX:** Well, I agree that these assumptions are milder than a blunt, unsupported declaration of your thesis conclusion,  $E(Z_2) - E(Z_1) = 0.78(y_2 - y_1)$ , and I am somewhat amazed that such mild assumptions can support a daring prediction about the actual effect of homework on score (under experimental setup). But I am still unhappy with your common cause assumption. It seems to me that a teacher who emphasizes the importance of homework would also be an inspiring, effective teacher, so  $e_2$  (which includes factors such as quality of teaching) should be correlated with  $x$ , contrary to your assumption.

**You:** Dr. EX, now you begin to talk like an SEM researcher. Instead of attacking the method and its philosophy, we are beginning to discuss substantive issues – e.g., whether it is reasonable to assume that a teacher’s effectiveness is uncorrelated with the weight



**Figure 11.15** Statistically equivalent models to Figure 11.14.

that the teacher assigns to homework. I personally have had great teachers that could not care less about homework, and conversely so.

But this is not what my thesis is about. I am not claiming that teachers' effectiveness is uncorrelated with how they weigh homework; I leave that to other researchers to test in future studies (or might it have been tested already?). All I am claiming is: Those researchers who are willing to accept the assumption that teachers' effectiveness is uncorrelated with how they weigh homework will find it interesting to note that this assumption, coupled with the data, logically implies the conclusion that an increase of one homework-hour per day causes an (average) increase of 0.78 grade points in student's score. And this claim can be verified empirically if we are allowed a controlled experiment with randomized amounts of homework ( $y$ ).

**Dr. EX:** I am glad you do not insist that your modeling assumptions are true; you merely state their plausibility and explicate their ramifications. I cannot object to that. But I have another question. You said that your model does not have any statistical implications, so it cannot be tested for fitness to data. How do you know that? And doesn't this bother you?

**You:** I know it by just looking at the graph and examining the missing links. A criterion named  $d$ -separation (see Section 11.1.2, " $d$ -separation without tears") permits students of SEM to glance at a graph and determine whether the corresponding model implies any constraint in the form of a vanishing partial correlation between variables. Most statistical implications (though not all) are of this nature. The model in our example does not imply any constraint on the covariance matrix, so it can fit perfectly any data whatsoever. We call this model "saturated," a feature that some SEM researchers, unable to shake off statistical-testing traditions, regard as a fault of the model. It isn't. Having a saturated model at hand simply means that the investigator is not willing to make implausible causal assumptions, and that the mild assumptions he/she is willing to make are too weak to produce statistical implications. Such a conservative attitude should be commended, not condemned. Admittedly, I would be happy if my model were not saturated – say, if  $e_1$  and  $e_2$  were uncorrelated. But this is not the case at hand; common sense tells us that  $e_1$  and  $e_2$  are correlated, and it also shows in the data. I tried assuming  $cov(e_1, e_2) = 0$ , and I got terrible fit. Am I going to make unwarranted assumptions just to get my model "knighted" as "nonsaturated"? No! I would rather make reasonable assumptions, get useful conclusions, and report my results side by side with my assumptions.

**Dr. EX:** But suppose there is another saturated model, based on equally plausible assumptions, yet leading to a different value of  $c$ . Shouldn't you be concerned with the possibility that some of your initial assumptions are wrong, hence that your conclusion  $c = 0.78$  is wrong? There is nothing in the data that can help you prefer one model over the other.

**You:** I am concerned indeed, and, in fact, I can immediately enumerate the structures of all such competing models; the two models in Figure 11.15 are examples, and many

more. (This too can be done using the  $d$ -separation criterion; see pp. 145–8.) But note that the existence of competing models does not in any way weaken my earlier stated claim: “Researchers who accept the qualitative assumptions of model  $M$  are compelled to accept the conclusion  $c = 0.78$ .” This claim remains logically invincible. Moreover, the claim can be further refined by reporting the conclusions of each contending model, together with the assumptions underlying that model. The format of the conclusion will then read:

If you accept assumption set  $A_1$ , then  $c = c_1$  is implied,  
 If you accept assumption set  $A_2$ , then  $c = c_2$  is implied,

and so on.

**Dr. EX:** I see, but still, in case we wish to go beyond these conditional statements and do something about deciding among the various assumption sets, are there no SEM methods to assist one in this endeavor? We, in statistics, are not used to facing problems with two competing hypotheses that cannot be submitted to some test, however feeble.

**You:** This is a fundamental difference between statistical data analysis and SEM. Statistical hypotheses, by definition, are testable by statistical methods. SEM models, in contrast, rest on *causal* assumptions, which, also by definition (see p. 39), cannot be given statistical tests. If our two competing models are saturated, we know in advance that there is nothing more we can do but report our conclusions in a conditional format, as listed above. If, however, the competition is among equally plausible yet statistically distinct models, then we are facing the century-old problem of model selection, where various selection criteria such as AIC have been suggested for analysis. However, the problem of model selection is now given a new, causal twist – our mission is not to maximize fitness, or to maximize predictive power, but rather to produce the most reliable estimate of causal parameters such as  $c$ . This is a new arena altogether (see Pearl 2004).

**Dr. EX:** Interesting. Now I understand why my statistician colleagues got so totally confused, mistrustful, even antagonistic, upon encountering SEM methodology (e.g., Freedman 1987; Holland 1988; Wermuth 1992). One last question. You started talking about randomized experiments only after realizing that I am a statistician. How would you explain your SEM strategy to a nonstatistician?

**You:** I would use plain English and say: “If we have the physical means of fixing  $y$  at some constant  $y_1$ , and of changing that constant from  $y_1$  to  $y_2$ , then the observed change in  $E(Z)$  would be  $c(y_2 - y_1)$ .” Most people understand what “fixing” means, because this is on the mind of policy makers. For example, a teacher interested in the effect of homework on performance does not think in terms of randomizing homework. Randomization is merely an indirect means for predicting the effect of fixing.

Actually, if the person I am talking to is really enlightened (and many statisticians are), I might even resort to counterfactual vocabulary and say, for example, that a student who scored  $z$  on the exam after spending  $y$  hours on homework would have scored  $z + c$  had he/she spent  $y + 1$  hours on homework. To be honest, this is what I truly had in mind when writing the equation  $z = cy + e_2$ , where  $e_2$  stood for all other characteristics of the student that were not given variable names in our model and that are not affected by  $y$ . I did not even think about  $E(Z)$ , only about  $z$  of a typical student. Counterfactuals are the most precise linguistic tool we have for expressing the meaning

of scientific relations. But I refrain from mentioning counterfactuals when I talk to statisticians because, and this is regrettable, statisticians tend to suspect deterministic concepts, or concepts that are not immediately testable, and counterfactuals are such concepts (Dawid 2000; Pearl 2000).

**Dr. EX:** Thanks for educating me on these aspects of SEM. No further questions.

**You:** The pleasure is mine.

#### 11.5.4 Where Is Economic Modeling Today? – Courting Causes with Heckman

Section 5.2 of this book decries the decline in the understanding of structural equation modeling in econometric in the past three decades (see also Hoover 2003, “Lost Causes”) and attributes this decline to a careless choice of notation which blurred the essential distinction between algebraic and structural equations. In a series of articles (Heckman 2000, 2003, 2005; Heckman and Vytlačil 2007), James Heckman has set out to overturn this perception, reclaim causal modeling as the central focus of economic research, and reestablish economics as an active frontier in causal analysis. This is not an easy task by any measure. To adopt the conceptual and technical advances that have emerged in neighboring disciplines would amount to admitting decades of neglect in econometrics, while to dismiss those advances would necessitate finding them econometric surrogates. Heckman chose the latter route, even though most modern advances in causal modeling are rooted in the ideas of economists such as Haavelmo (1943), Marschak (1950), and Strotz and Wold (1960).

One step in Heckman’s program was to reject the *do*-operator and the “surgery” semantics upon which it is based, thus depriving economists of the structural semantics of counterfactuals developed in this book (especially Chapter 7), which unifies traditional econometrics with the potential-outcome approach. Heckman’s reasons for rejecting surgery are summarized thus:

Controlled variation in external (forcing) variables is the key to defining causal effects in nonrecursive models ... Pearl defines a causal effect by ‘shutting one equation down’ or performing ‘surgery’ in his colorful language. He implicitly assumes that ‘surgery,’ or shutting down an equation in a system of simultaneous equations, uniquely fixes one outcome or internal variable (the consumption of the other person in my example). In general, it does not. Putting a constraint on one equation places a restriction on the entire set of internal variables. In general, no single equation in a system of simultaneous equations uniquely determines any single outcome variable. Shutting down one equation might also affect the parameters of the other equations in the system and violate the requirements of parameter stability. (Heckman and Vytlačil 2007)

Clearly, Heckman’s objections are the same as Cartwright’s (Section 11.4.6):

1. Ideal surgery may be technically infeasible,
2. Economic systems are nonmodular.

We have repudiated these objections in four previous subsections (11.4.3–11.4.6) which readers can easily reapply to deconstruct Heckman’s arguments. It is important to reemphasize, though, that, as in the case of Cartwright, these objections emanate from conflating the task of definition (of counterfactuals) with those of identification and

practical estimation, a frequent confusion among researchers which Heckman (2005) sternly warns readers to avoid.

This conflation is particularly visible in Heckman’s concern that “shutting down one equation might also affect the parameters of the other equations in the system.” In the physical world, attempting to implement the conditions dictated by a “surgery” may sometimes affect parameters in other equations, and, as we shall see, the same applies to Heckman’s proposal of “external variation.” However, we are dealing here with symbolic, not physical, manipulations. Our task is to formulate a meaningful mathematical definition of “the causal effect of one variable on another” in a symbolic system called a “model.” This permits us to manipulate symbols at will, while ignoring the technical feasibility of these manipulations. Implementational considerations need not enter the discussion of *definition*.

**A New Definition of Causal Effects: “External Variation”**

Absent surgery semantics, Heckman and Vytlacil (HV) set out to configure a new definition of causal effects, which, hopefully, would be free of the faults they discovered in the surgery procedure, by basing it on “external-variations,” instead of shutting down equations. It is only unfortunate that their new definition, the cornerstone of their logic of counterfactuals, is not given an explicit formal exposition: it is relegated to a semiformal footnote (HV, p. 77) that even a curious and hard-working reader would find difficult to decipher. The following is my extrapolation of HV’s definition as it applies to multi-equations and nonlinear systems.

Given a system of equations:

$$Y_i = f_i(Y, X, U) \quad i = 1, 2, \dots, n,$$

where  $X$  and  $U$  are sets of observed and unobserved external variables, respectively, the causal effect of  $Y_j$  on  $Y_k$  is computed in four steps:

1. Choose any member  $X_t$  of  $X$  that appears in  $f_j$ . If none exists, exit with failure.
2. If  $X_t$  appears in any other equation as well, consider excluding it from that equation (e.g., set its coefficient to zero if the equation is linear or replace  $X_t$  by a constant).<sup>15</sup>

3. Solve for the reduced form

$$Y_i = g_i(X, U) \quad i = 1, 2, \dots, n \tag{11.23}$$

of the resulting system of equations.

4. The causal effect of  $Y_j$  on  $Y_k$  is given by the partial derivative:

$$dY_k/dY_j = dg_k/dX_t : dg_j/dX_t. \tag{11.24}$$

**Example 11.5.2** Consider a system of three equations:

$$\begin{aligned} Y_1 &= aY_2 + cY_3 + eX + U_1 \\ Y_2 &= bY_1 + X + U_2 \\ Y_3 &= dY_1 + U_3. \end{aligned}$$

---

<sup>15</sup> It is not clear what conditions (if any) would forbid one from setting  $e = 0$ , in example 11.5.2, or ignoring  $X$  altogether and adding a dummy variable  $X'$  to the second equation. HV give the impression that deciding on whether  $e$  can be set to 0 requires deep understanding of the problem at hand; if this is their intention, it need not be.

*Needed: the causal effect of  $Y_2$  on  $Y_1$ .*

The system has one external variable,  $X$ , which appears in the first two equations. If we can set  $e = 0$ ,  $x$  will appear in the equation of  $Y_2$  only, and we can then proceed to Step 3 of the “external variation” procedure. The reduced form of the modified model yields:

$$dY_1/dX = a/(1 - ba - cd) \quad dY_2/dX = (1 - cd)/(1 - ab - cd),$$

and the causal effect of  $Y_1$  on  $Y_2$  calculates to:

$$dY_1/dY_2 = a/(1 - cd).$$

In comparison, the surgery procedure constructs the following modified system of equations:

$$\begin{aligned} Y_1 &= aY_2 + cY_3 + eX + U_1 \\ Y_2 &= y_2 \\ Y_3 &= dY_1 + U_3, \end{aligned}$$

from which we obtain for the causal effect of  $Y_2$  on  $Y_1$ ;

$$dY_1/dy_2 = a/(1 - cd),$$

an expression identical to that obtained from the “external variation” procedure.

It is highly probable that the two procedures always yield identical results, which would bestow validity and conceptual clarity on the “external variation” definition.

### 11.5.5 External Variation vs. Surgery

In comparing their definition to the one provided by the surgery procedure, HV write (p. 79): “Shutting down an equation or fiddling with the parameters ... is not required to define causality in an interdependent, nonrecursive system or to identify causal parameters. The more basic idea is exclusion of different external variables from different equations which, when manipulated, allow the analyst to construct the desired causal quantities.”

I differ with HV on this issue. I believe that “surgery” is the more basic idea, more solidly motivated, and more appropriate for policy evaluation tasks. I further note that basing a definition on exclusion and external variation suffers from the following flaws:

1. In general, “exclusion” involves the removal of a variable from an equation and amounts to “fiddling with the parameters.” It is, therefore, a form of “surgery” – a modification of the original system of equations – and would be subject to the same criticism one may raise against “surgery.” Although we have refuted such criticism in previous sections, we should nevertheless note that if it ever has a grain of validity, the criticism would apply equally to both methods.
2. The idea of relying exclusively on external variables to reveal internal cause–effect relationships has its roots in the literature on *identification* (e.g., as in the studies of “instrumental variables”) when such variables act as “nature’s experiments.” This restriction, however, is unjustified in the context

of *defining* causal effect, since “causal effects” are meant to quantify effects produced by *new* external manipulations, not necessarily those shown explicitly in the model and not necessarily those operating in the data-gathering phase of the study. Moreover, every causal structural equation model, by its very nature, provides an implicit mechanism for emulating such external manipulations, via surgery.

Indeed, most policy evaluation tasks are concerned with *new* external manipulations which exercise direct control over endogenous variables. Take, for example, a manufacturer deciding whether to double the current price of a given product after years of letting the price track the cost, i.e.,  $price = f(cost)$ . Such a decision amounts to removing the equation  $price = f(cost)$  in the model at hand (i.e., the one responsible for the available data) and replacing it with a constant equal to the new price. This removal emulates faithfully the decision under evaluation, and attempts to circumvent it by appealing to “external variables” are artificial and hardly helpful.

As another example, consider the well-studied problem (Heckman 1992) of evaluating the impact of terminating an educational program for which students are admitted based on a set of qualifications. The equation  $admission = f(qualifications)$  will no longer hold under program termination, and no external variable can simulate the new condition (i.e.,  $admission = 0$ ) save for one that actually neutralizes (or “ignores,” or “shuts down”) the equation  $admission = f(qualifications)$ .

It is also interesting to note that the method used in Haavelmo (1943) to define causal effects is mathematically equivalent to surgery, not to external variation. Instead of replacing the equation  $Y_j = f_j(Y, X, U)$  with  $Y_j = y_j$ , as would be required by surgery, Haavelmo writes  $Y_j = f_j(Y, X, U) + x_j$ , where  $X_j$  is chosen so as to make  $Y_j$  constant,  $Y_j = y_j$ . Thus, since  $X_j$  liberates  $Y_j$  from any residual influence of  $f_j(Y, X, U)$ , Haavelmo’s method is equivalent to that of surgery. Heckman’s method of external variation leaves  $Y_j$  under the influence  $f_j$ .

3. Definitions based on external variation have the obvious flaw that the target equation may not contain any observable external variable. In fact, in many cases the set of observed external variables in the system is empty (e.g., Figure 3.5). Additionally, a definition based on a ratio of two partial derivatives does not generalize easily to nonlinear systems with discrete variables. Thus, those who seriously accept Heckman’s definition would be deprived of the many identification techniques now available for instrumentless models (see Chapters 3 and 4) and, more seriously yet, would be unable to even ask whether causal effects are identified in any such model – identification questions are meaningless for undefined quantities.

Fortunately, liberated by the understanding that definitions can be based on purely symbolic manipulations, we can modify Heckman’s proposal and *add* fictitious external variables to any equation we desire. The added variables can then serve to define causal effects in a manner similar to the steps in equations (11.23) and (11.24) (assuming continuous variables). This brings us closer to surgery, with the one basic difference of leaving  $Y_j$  under the influence of  $f_j(Y, X, U)$ .

Having argued that definitions based on “external variation” are conceptually ill-motivated, we now explore whether they can handle noncausal systems of equations.

### *Equation Ambiguity in Noncausal Systems*

Several economists (Leroy 2002; Neuberger 2003; Heckman and Vytlacil 2007) have criticized the *do*-operator for its reliance on *causal*, or directional, structural equations, where we have a one-to-one correspondence between variables and equations. HV voice this criticism thus: “In general, no single equation in a system of simultaneous equations uniquely determines any single outcome variable” (Heckman and Vytlacil 2007, p. 79).

One may guess that Heckman and Vytlacil refer here to systems containing nondirectional equations, namely, equations in which the equality sign does not stand for the non-symmetrical relation “is determined by” or “is caused by” but for symmetrical algebraic equality. In econometrics, such noncausal equations usually convey equilibrium or resource constraints; they impose equality between the two sides of the equation but do not endow the variable on the left-hand side with the special status of an “outcome” variable.

The presence of nondirectional equations creates ambiguity in the surgical definition of the counterfactual  $Y_x$ , which calls for replacing the equation *determining*  $X$  with the constant equation  $X = x$ . If  $X$  appears in several equations, and if the position of  $X$  in the equation is arbitrary, then each one of those equations would be equally qualified for replacement by  $X = x$ , and the value of  $Y_x$  (i.e., the solution for  $Y$  after replacement) would be ambiguous.

Note that symmetrical equalities differ structurally from reciprocal causation in directional nonrecursive systems (i.e., systems with feedback, as in Figure 7.4), since, in the latter, each variable is an “outcome” of precisely one equation. Symmetrical constraints can nevertheless be modeled as the solution of a dynamic feedback system in which equilibrium is reached almost instantaneously (Lauritzen and Richardson 2002; Pearl 2003a).

Heckman and Vytlacil create the impression that equation ambiguity is a flaw of the surgery definition and does not plague the exclusion-based definition. This is not the case. In a system of nondirectional equations, we have no way of knowing which external variable to exclude from which equation to get the right causal effect.

For example: Consider a nonrecursive system of two equations that is discussed in HV, p. 75:

$$Y_1 = a_1 + c_{12}Y_2 + b_{11}X_1 + b_{12}X_2 + U_1 \quad (11.25)$$

$$Y_2 = a_2 + c_{21}Y_1 + b_{21}X_1 + b_{22}X_2 + U_2. \quad (11.26)$$

Suppose we move  $Y_1$  to the l.h.s. of (11.26) and get:

$$Y_1 = [a_2 - Y_2 + b_{21}X_1 + b_{22}X_2 + U_2]/c_{21}. \quad (11.27)$$

To define the causal effect of  $Y_2$  on  $Y_1$ , we now have a choice of excluding  $X_2$  from (11.25) or from (11.27). The former yields  $c_{12}$ , while the latter yields  $1/c_{21}$ . We see that the ambiguity we have in choosing an equation for surgery translates into ambiguity in choosing an equation and an external variable for exclusion.

Methods of breaking this ambiguity were proposed by Simon (1953) and are discussed on pages 226–8.

## Summary

The idea of constructing causal quantities by exclusion and manipulation of external variables, while soundly motivated in the context of identification problems, has no logical basis when it comes to model-based definitions. Definitions based on surgery, on the other hand, enjoy generality, semantic clarity, and computational simplicity.

So, where does this leave econometric modeling? Is the failure of the “external variable” approach central or tangential to economic analysis and policy evaluation?

In almost every one of his recent articles James Heckman stresses the importance of counterfactuals as a necessary component of economic analysis and the hallmark of econometric achievement in the past century. For example, the first paragraph of the HV article reads: “they [policy comparisons] require that the economist construct counterfactuals. Counterfactuals are required to forecast the effects of policies that have been tried in one environment but are proposed to be applied in new environments and to forecast the effects of new policies.” Likewise, in his *Sociological Methodology* article (2005), Heckman states: “Economists since the time of Haavelmo (1943, 1944) have recognized the need for precise models to construct counterfactuals... The econometric framework is explicit about how counterfactuals are generated and how interventions are assigned...”

And yet, despite the proclaimed centrality of counterfactuals in econometric analysis, a curious reader will be hard pressed to identify even one econometric article or textbook in the past 40 years in which counterfactuals or causal effects are formally defined. Needed is a procedure for computing the counterfactual  $Y(x, u)$  in a well-posed, fully specified economic model, with  $X$  and  $Y$  two arbitrary variables in the model. By rejecting Haavelmo’s definition of  $Y(x, u)$ , based on surgery, Heckman commits econometrics to another decade of division and ambiguity, with two antagonistic camps working in almost total isolation.

Economists working within the potential-outcome framework of the Neyman-Rubin model take counterfactuals as primitive, unobservable variables, totally detached from the knowledge encoded in structural equation models (e.g., Angrist 2004; Imbens 2004). Even those applying propensity score techniques, whose validity rests entirely on the causal assumption of “ignorability,” or unconfoundedness, rarely know how to confirm or invalidate that assumption using structural knowledge (see Section 11.3.5). Economists working within the structural equation framework (e.g., Kennedy 2003; Mittelhammer et al. 2000; Intriligator et al. 1996) are busy estimating parameters while treating counterfactuals as metaphysical ghosts that should not concern ordinary mortals. They trust leaders such as Heckman to define precisely what the policy implications are of the structural parameters they labor to estimate, and to relate them to what their colleagues in the potential-outcome camp are doing.<sup>16</sup>

The surgery semantics (pp. 98–102) and the mathematical properties entailed by it (Chapters 7–10), offer a simple and precise unification of these two antagonistic and narrowly focused schools of econometric research – a theorem in one approach entails a theorem in the other, and vice versa. Economists will do well resurrecting the basic

---

<sup>16</sup> Notably, the bibliographical list in the comprehensive review article by economist Hoover (2008) is almost disjoint from those of economists Angrist (2004) and Imbens (2004) – the cleavage is culturally deep.

ideas of Haavelmo (1943), Marschak (1950), and Strotz and Wold (1960) and re-invigorating them with the logic of graphs and counterfactuals presented in this book.

For completeness, I reiterate here explicitly (using parenthetical notation) the two fundamental connections between counterfactuals and structural equations.

1. The structural definition of counterfactuals is:

$$Y_M(x, u) = Y_{M_x}(u).$$

Read: For any model  $M$  and background information  $u$ , the counterfactual conditional “ $Y$  if  $X$  had been  $x$ ” is given by the solution for  $Y$  in submodel  $M_x$  (i.e., the mutilated version of  $M$  with the equation determining  $X$  replaced by  $X = x$ ).

2. The empirical claim of the structural equation  $y = f(x, e(u))$  is:

$$Y(x, z, u) = f(x, e(u)),$$

for any set  $Z$  not intersecting  $X$  or  $Y$ .

Read: Had  $X$  and  $Z$  been  $x$  and  $z$ , respectively,  $Y$  would be  $f(x, e(u))$ , independently of  $z$ , and independently of other equations in the model.

## 11.6 DECISIONS AND CONFOUNDING (CHAPTER 6)

### 11.6.1 Simpson’s Paradox and Decision Trees

*Nimrod Megiddo (IBM Almaden) Wrote:*

“I do not agree that ‘causality’ is the key to resolving the paradox (but this is also a matter of definition) and that tools for looking at it did not exist twenty years ago. Coming from game theory, I think the issue is not difficult for people who like to draw decision trees with ‘decision’ nodes distinguished from ‘chance’ nodes.

I drew two such trees [Figure 11.16(a) and (b)] which I think clarify the correct decision in different circumstances.”

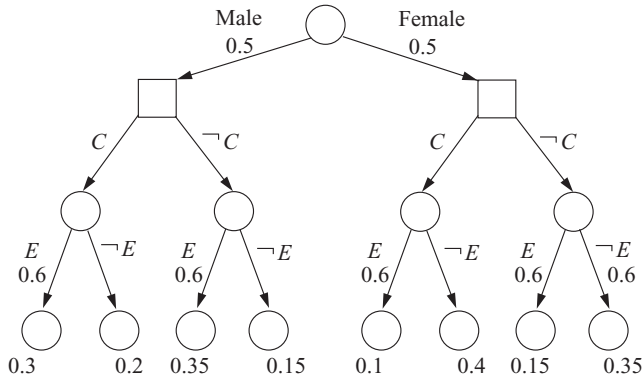
*Author’s Reply:*

The fact that you have constructed two different decision trees for the same input tables implies that the key to the construction was not in the data, but in some information you obtained from the story behind the data. What is that information?

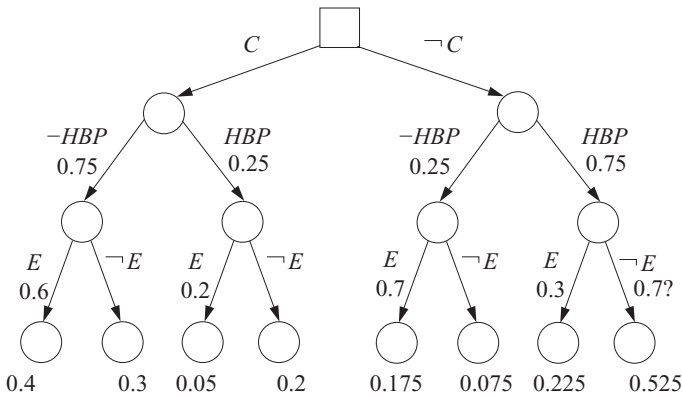
The literature of decision tree analysis has indeed been in existence for at least fifty years, but, to the best of my knowledge, it has not dealt seriously with the problem posed above: “What information do we use to guide us into setting up the correct decision tree?”

We agree that giving a robot the frequency tables *alone* would not be sufficient for the task. But what else would Mr. Robot (or a statistician) need? Changing the story from  $F = \text{“female”}$  to  $F = \text{“blood pressure”}$  seems to be enough for people, because people understand informally the distinct roles that gender and blood pressure play in the scheme of things. Can we characterize these roles formally, so that our robot would be able to construct the correct decision tree?

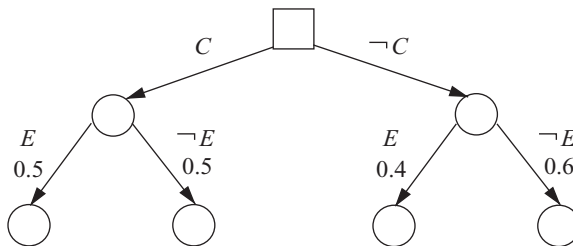
My proposal: give the robot (or a statistician or a decision-tree expert) a pair  $(T, G)$ , where  $T$  is the set of frequency tables and  $G$  is a causal graph, and, lo and behold, the



**Figure 11.16(a)** Decision tree corresponding to Figure 11.17(a). Given Male,  $\neg C$  is better than  $C$  ( $0.35 > 0.3$ ). Given Female,  $\neg C$  is also better ( $0.15 > 0.1$ ). Unconditionally, with any probability  $p$  for Male and  $1 - p$  for Female, again,  $\neg C$  is better than  $C$  ( $0.35p + 0.15(1 - p) > 0.3p = 0.1(1 - p)$ ).



**Figure 11.16(b)** Decision tree corresponding to Figure 11.17(b). The tree can be compressed as shown in Figure 11.16(c).



**Figure 11.16(c)** Compressed decision tree. Obviously,  $C$  is better than  $\neg C$  (since  $0.5 > 0.4$ ).

robot would be able to set up the correct decision tree automatically. This is what I mean in saying that the resolution of the paradox lies in causal considerations. Moreover, one can go further and argue: “If the information in  $(T, G)$  is sufficient, why not skip the construction of a decision tree altogether, and get the right answer directly from  $(T, G)$ ?” This is the gist of Chapters 3–4 in this book. Can the rich literature on decision analysis

benefit from the more economical encoding of decision problems in the syntax of  $(T, G)$ ? The introduction of influence diagrams (Howard and Matheson 1981; Pearl 2005) was a step in this direction, and, as Section 4.1.2 indicates, the second step might not be too far off. While an influence diagram is parameterized for a specific choice of decision and chance variables, a causal diagram is not specific to any such choice; it contains the information for parameterizing all influence diagrams (and decision trees) for any choice of decision variables and any choice of chance variables that one may observe at decision time.

More recently, Dawid (2002) has developed a hybrid representation, combining influence diagrams and causal diagrams, by attaching to every variable in the latter a decision node (as in Figure 3.2) representing a potential intervention, capable of switching from an observational (idle) to interventional (active) mode (as in equation (3.8)).

### 11.6.2 Is Chronological Information Sufficient for Decision Trees?

#### *Megiddo Wrote Back:*

“The term ‘causality’ introduces into the problem issues that do not have to be there, such as determinism, free will, cause and effect, etc. What does matter is a specification that, in the outcome fixing process, fixing the value of variable  $X$  occurs before fixing the value of a variable  $Y$ , and  $Y$  depends on  $X$ . You like to call this situation a causality relation. Of course in a mathematical theory you can choose any name you like, but then people are initially tempted to develop some intuition, which may be wrong due to the external meaning of the name you choose. The interpretation of this intuition outside the mathematical model often has real-life implications that may be wrong, for example, that  $X$  really causes  $Y$ . The decision tree is a simple way to demonstrate the additional chronological information, and simple directed graphs can of course encode that information more concisely. When you have to define precisely what these graphs mean, you refer to a fuller description like the trees. So, in summary, my only objection is to the use of the word ‘causality’ and I never had doubts that chronological order information was crucial to a correct decision making based on past data.”

#### *Author’s Reply*

1. Before the 21st century, there was some danger in attributing to certain mathematical relations labels such as “causal,” which were loaded with intuition, mystery, and controversy. This is no longer the case – the mystery is gone, and the real-life implication of this interpretation is not wrong;  $X$  really causes  $Y$ .

Moreover, if in addition to getting the mathematics right, one is also interested in explicating those valuable intuitions, so that we can interpret them more precisely and even teach them to robots, then there is no escape but to label those relationships with whatever names they currently enjoy in our language, namely, “causal.”

2. There is more that enters a decision tree than chronological and dependence information. For example, the chronological and dependence information that is conveyed by Figure 11.17(c) is identical to that of Figure 11.17(a) (assuming  $F$  occurs before  $C$ ), yet (c) calls for a different decision tree (and yields a different conclusion), because the dependence between  $F$  and  $Y$  is “causal” in (a) and associational in (c). Thus, causal considerations must supplement chronological

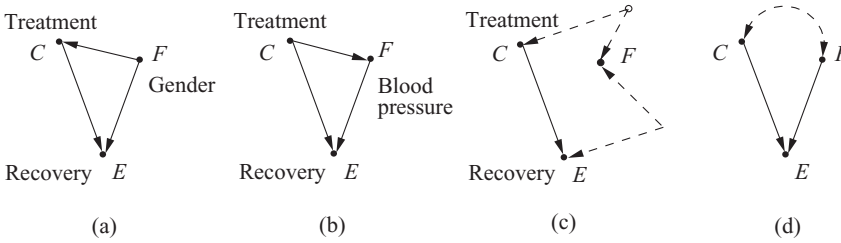


Figure 11.17 Graphs demonstrating the insufficiency of chronological information.

and dependence information if we are to construct correct decision trees and to load their branches with correct probabilities.

As a thought experiment, imagine that we wish to write a program that automatically constructs decision trees from stories like those in Figure 11.17(a)-(b)-(c). The program is given the empirical frequency tables and is allowed to ask us questions about chronological and dependence relationships among  $C$ ,  $E$ , and  $F$ , but is not allowed to use any causal vocabulary. Would the program be able to distinguish between (a) and (c)? The answer is: No. If one ignores causal considerations and attends, as you suggest, to chronological and dependence information alone, then the decision tree constructed for Figure 11.17(c) would be identical to the one drawn for  $F = \textit{gender}$  (Figure 11.16(a)), and the wrong conclusion would ensue, namely, that the drug ( $C$ ) is harmful for both  $F = \textit{true}$  and  $F = \textit{false}$  patients. This is wrong, because the correct answer is that the drug is beneficial to the population as a whole (as in the blood pressure example), hence it must be beneficial for either  $F = \textit{true}$  or  $F = \textit{false}$  patients (or both). The decision tree hides this information and yields the wrong results.

The error stems from attributing the wrong probabilities to the branches of the decision tree. For example, the leftmost branch in the tree is assigned a probability  $P(E | C, \neg F) = 0.6$ , which is wrong; the correct probability should be  $P(E | \textit{do}(C), \neg F)$ , which, in the case of Figure 11.17(c), cannot be determined from the graph and the tables. We do know, however, that  $P(E | \textit{do}(C)) = 0.5$  and  $P(E | \textit{do}(\neg C)) = 0.4$ , so either equation (6.4) or equation (6.5) must be violated.

The conventional wisdom in decision analysis is to assign the tree branches conditional probabilities of the type  $P(E | \textit{action}, Z)$ , where  $Z$  is the information available at decision time. Practitioners should be warned against following this convention blindly. The correct assignment should be, of course,  $P(E | \textit{do}(\textit{action}), Z)$  which is estimable either from experimental studies or when the causal graph renders this quantity identifiable.

3. I do not agree that in order to define precisely what causal graphs mean, we must “refer to a fuller description like the trees.” In Section 7.1 we find a formal definition of causal graphs as a collection of functions, and this definition invokes no decision trees (at least not explicitly). Thus, a causal graph has meaning of its own, independent of the many decision trees that the graph may help us construct. By analogy, it would be awkward (though not mathematically wrong) to

say that the meaning of a differential equation (say, for particle motion) lies in the set of trajectories that obey that equation; the meaning of each term in a differential equation can be defined precisely, from first principles, without having a clue about the solution to the equation.

### 11.6.3 Lindley on Causality, Decision Trees, and Bayesianism

#### *Question to Author (from Dennis Lindley):*

If your assumption, that controlling  $X$  at  $x$  is equivalent to removing the function for  $X$  and putting  $X = x$  elsewhere, is applicable, then it makes sense. What I do not understand at the moment is the relevance of this to decision trees. At a decision node, one conditions on the quantities known at the time of the decision. At a random node, one includes all relevant uncertain quantities under known conditions. Nothing more than the joint distributions (and utility considerations) are needed. For example, in the medical case, the confounding factor may either be known or not at the time the decision about treatment is made, and this determines the structure of the tree. Where causation may enter is when the data are used to assess the probabilities needed in the tree, and it is here that Novick and I used exchangeability. The Bayesian paradigm makes a sharp distinction between probability as belief and probability as frequency, calling the latter, chance. If I understand causation, it would be reasonable that our concept could conveniently be replaced by yours in this context.

#### *Author's Reply:*

Many decision analysts take the position that causality is not needed because: “Nothing more than the joint distributions (and utility considerations) are needed” (see discussion in section 11.6.2). They are right and wrong. While it is true that probability distributions are all that is needed, the distributions needed,  $P(y | do(x), see(z))$ , are of a peculiar kind; they are not derivable from the joint distribution  $P(y, x, z)$  which governs our data, unless we supplement the data with causal knowledge, such as that provided by a causal graph.

Your next sentence says it all:

Where causation may enter is when the data are used to assess the probabilities needed in the tree,...

I take the frequency/belief distinction to be tangential to discussions of causality. Let us assume that the tables in Simpson's story were not frequency tables, but summaries of one's subjective beliefs about the occurrence of various joint events,  $(C, E, F)$ ,  $(C, E, -F)$ ... etc. My assertion remains that this summary of beliefs is not sufficient for constructing our decision tree. We need also to assess our belief in the hypothetical event “ $E$  would occur if a decision  $do(C)$  is taken,” and, as we have seen, temporal information alone is insufficient for deriving this assessment from the tabulated belief summaries, hence, we cannot construct the decision tree from this belief summary; we need an extra ingredient, which I call “causal” information and you choose to call “exchangeability” – I would not quarrel about nomenclature except to note that, if we trust human beings to provide us with reliable information, we should also respect the vocabulary in which they cast that information, and, much to the disappointment of prudent academics, that vocabulary is unmistakably causal.

***Update Question to Author:***

Your point about probability and decision trees is well taken and I am in agreement with what you say here; a point that I had not appreciated before. Thank you. Let me rephrase the argument to see whether we are in agreement. In handling a decision tree it is easy to see what probabilities are needed to solve the problem. It is not so easy to see how these might be assessed numerically. However, I do not follow how causality completely resolves the issue. Nor, of course, does exchangeability.

***Author's Reply:***

I am very glad that we have narrowed the problem down to simple and concrete issues: (1) how to assess the probabilities needed for a decision tree, (2) where those probabilities come from, (3) how those probabilities can be encoded economically, and perhaps even (4) whether those probabilities must comply with certain rules of internal coherence, especially when we construct several decision trees, differing in the choice of decision and chance nodes.

The reason I welcome this narrowing of the problem is that it would greatly facilitate my argument for explicit distinction between causal and probabilistic relationships. In general, I have found Bayesian statisticians to be the hardest breed of statisticians to convince of the necessity of this distinction. Why? Because whereas classical statisticians are constantly on watch against assumptions that cannot be substantiated by hard data, Bayesians are more permissive in this regard, and rightly so. However, by licensing human judgment as a legitimate source of information, Bayesians have become less meticulous in keeping track of the character and origin of that information. Your earlier statement describes the habitual, uncritical use of conditioning among Bayesian researchers, which occasionally lures the unwary into blind alleys (e.g., Rubin (2009)):

At a decision node, one conditions on the quantities known at the time of the decision.

At a random node, one includes all relevant uncertain quantities under known conditions. Nothing more than the joint distributions (and utility considerations) are needed.

As Newcomb's paradox teaches us (see Section 4.1), it is not exactly true that "at a decision node, one conditions on the quantities known at the time of the decision" – at least some of the "conditionings" need to be replaced with "doing." If this were not the case, then all decision trees would turn into a joke; "patients should avoid going to the doctor to reduce the probability that one is seriously ill (Skyrms 1980, p. 130); workers should never hurry to work, to reduce the probability of having overslept; students should not prepare for exams, lest this would prove them behind in their studies; and so on. In short, all remedial actions should be banished lest they increase the probability that a remedy is indeed needed." (Chapter 4, p. 108).

But even after escaping this "conditioning" trap, the Bayesian philosopher does not see any difference between assessing probabilities for branches emanating from decision nodes and those emanating from chance nodes. For a Bayesian, both assessments are probability assessments. That the former involves a mental simulation of hypothetical experiment while the latter involves the mental envisioning of passive observations is of minor significance, because Bayesians are preoccupied with the distinction between probability as belief and probability as frequency.

This preoccupation renders them less sensitive to the fact that beliefs come from a variety of sources and that it is important to distinguish beliefs about outcomes of experiments from beliefs about outcomes of passive observations (Pearl 2001a, “Bayesianism and Causality”). First, the latter are robust, the former are not; that is, if our subjective beliefs about passive observations are erroneous, the impact of the error will get washed out in due time, as the number of observations increases. This is not the case for belief about outcomes of experiments; erroneous priors will produce a permanent bias regardless of sample size. Second, beliefs about outcomes of experiments cannot be articulated in the grammar of probability calculus; therefore, in order to formally express such beliefs and coherently combine them with data, probability calculus must be enriched with causal notation (i.e., graphs,  $do(x)$ , or counterfactuals).

So far, I have found two effective ways to win the hearts of Bayesians, one involving the notion of “economy” (see my discussion in Section 11.6.2), the other the notion of “coherence.”

Given a set of  $n$  variables of interest, there is a huge number of decision trees that can conceivably be constructed from these variables, each corresponding to a different choice of temporal ordering and a different choice of decision nodes and chance nodes from those variables. The question naturally arises, how can a decision maker ensure that probability assessments for all these decision trees be reproducible? Surely we cannot assume that human explicitly store all these potential decision trees in their heads. For reproducibility, we must assume that all these assessments must be derived from some economical representation of knowledge about decisions and chance events. Causal relationships can thus be viewed as an economical representation from which decision trees are constructed. Indeed, as I wrote to Megiddo (Section 11.6.2), if we were in need of instructing a robot to construct such decision trees on demand, in accordance with the available knowledge and belief, our best approach would be to feed the robot a pair of inputs  $(G, P)$ , where  $G$  is a causal graph and  $P$  is our joint distribution over the variables of interest (subjective distribution, if we were Bayesian). With the help of this pair of objects, the robot should be able to construct consistently all the decision trees required, for any partition of the variables into decision and chance nodes, and replicate the parameters on the branches. This is one way a Bayesian could appreciate causality without offending the traditional stance that “it is nothing more than the joint distributions...”

The second approach involves “coherence.” Coherence is something of which Bayesians are very proud, because DeFinetti, Savage, and others have labored hard to construct qualitative axioms that prevent probability judgments from being totally whimsical, and that compel beliefs to conform to the calculus of probability.

We can ask the Bayesian philosopher to tell us whether judgments about joint probabilities, say  $P(x, y)$ , should in some way cohere with judgments about decision-based probabilities, say  $P(y|do(x))$ , which quantify the branch emanating from a decision node with two alternatives  $do(X = x)$  and  $do(X = x')$ . We can then ask the Bayesian whether these probabilities should bear any connection to the usual conditional probabilities,  $P(y|x)$ , namely, the probability assessed for outcome  $Y = y$  that emanates (in some other decision tree) from a chance event  $X = x$ .

It should not be too hard to convince our Bayesian that these two assessments could not be totally arbitrary, but must obey some restrictions of coherence. For example, the inequality  $P(y | do(x)) \geq P(y, x)$  should be obeyed for all events  $x$  and  $y$ .<sup>17</sup> Moreover, coherence restrictions of this kind are automatically satisfied whenever  $P(y | do(x))$  is derived from a causal network according to the rules of Chapter 3. These two arguments should be inviting for a Bayesian to start drawing mathematical benefits from causal calculus, while maintaining caution and skepticism, and, as they say in the Talmud:

*“From benefits comes understanding”*

(free translation of *“mitoch shelo lishma, ba lishma”* (Talmud, Psahim, 50b)).

Bayesians will eventually embrace causal vocabulary, I have no doubt.

#### 11.6.4 Why Isn't Confounding a Statistical Concept?

In June 2001, I received two anonymous reviews of my paper “Causal Inference in the Health Sciences” (Pearl 2001c). The questions raised by the reviewers astounded me, for they reminded me of the archaic way some statisticians still think about causality and of the immense educational effort that still lies ahead. In the interest of contributing to this effort, I am including my reply in this chapter. Related discussion on the causal–statistical distinction is presented in Section 11.1.

##### *Excerpts from Reviewers' Comments:*

###### **Reviewer 1.**

“The contrast between statistical and causal concepts is overdrawn. Randomization, instrumental variables, and so forth have clear statistical definitions. ... [the paper urges] ‘that any systematic approach to causal analysis must require new mathematical notation.’ This is false: there is a long tradition of informal – but systematic and successful – causal inference in the medical sciences.”

###### **Reviewer 2.**

“The paper makes many sweeping comments which rely on distinguishing ‘statistical’ and ‘causal’ concepts ... Also, included in the list of causal (and therefore, according to the paper, non-statistical) concepts is, for example, confounding, which is solidly founded in standard, frequentist statistics. Statisticians are inclined to say things like ‘ $U$  is a potential confounder for examining the effect of treatment  $X$  on outcome  $Y$  when both  $U$  and  $X$  and  $U$  and  $Y$  are not independent. So why isn't confounding a statistical concept?’ ... If the author wants me to believe this, he's going to have to show at least one example of how the usual analyses fail.”

<sup>17</sup> This inequality follows from (3.52) or (9.33). A complete characterization of coherence constraints is given in Tian, Kang, and Pearl (2006). As an example, for any three variables  $X, Y, Z$ , coherence dictates:  $P(y | do(x, z)) - P(y, x | do(z)) - P(y, z | do(x)) + P(x, y, z) \geq 0$ . If the structure of a causal graph is known, the conditions of Definition 1.3.1 constitute a complete characterization of all coherence requirements.

*Author's Response:*

Reviewer #1 seems to advocate an informal approach to causation (whatever that means; it reminds me of informal statistics before Bernoulli), and, as such, he/she comes to this forum with set ideas against the basic aims and principles of my paper. Let history decide between us.

The target of this paper are readers of Reviewer #2's persuasion, who attempt to reconcile the claims of this paper (occasionally sweeping, I admit) with traditional statistical wisdom. To this reviewer I have the following comments.

You question the usefulness of my proposed demarcation line between statistical and causal concepts. Let me try to demonstrate this usefulness by considering the example that you bring up: confounding. You write that "confounding is solidly founded in standard, frequentist statistics." and that statisticians are inclined to say things like " $U$  is a potential confounder for examining the effect of treatment  $X$  on outcome  $Y$  when both  $U$  and  $X$  and  $U$  and  $Y$  are not independent. So why isn't confounding a statistical concept?"

Chapter 6 of my book goes to great lengths explaining why this definition fails on both sufficiency and necessity tests, and why all variants of this definition must fail by first principles. I will bring just a couple of examples to demonstrate the point; additional discussion is provided in Section 11.3.3. Consider a variable  $U$  that is affected by both  $X$  and  $Y$  – say, one that turns 1 whenever both  $X$  and  $Y$  reach high levels.  $U$  satisfies your criterion, and yet  $U$  is not a confounder for examining the effect of treatment  $X$  on outcome – in fact,  $U$  can safely be ignored in our analysis. (The same goes for any variable  $U$  whose effect on  $Y$  is mediated by  $X$ , like  $Z$  in Figure 2 of my paper.) As a second example, consider a variable  $U$  that resides "on the causal pathway" from  $X$  to  $Y$ . This variable too satisfies your criterion yet it is not a confounder – generations of statisticians have been warned (see Cox 1958) not to adjust for such variables. One might argue that your definition is merely a necessary, but not sufficient condition for confounding. But this too fails. Chapter 6 (pp. 185–6) describes examples where there is no variable that satisfies your definition, and still the effect of treatment  $X$  on outcome  $Y$  is confounded.

One can also construct an example (Figure 6.5) where  $U$  is a confounder (i.e., must be adjusted to remove effect bias), and still  $U$  is not associated with either  $X$  or  $Y$ .

I am not the first to discover discrepancies between confounding and its various statistical "definitions." Miettinen and Cook and Robins and Greenland have been arguing this point in epidemiology since the mid-1980s – to no avail. Investigators continue to equate collapsibility with no-confounding and continue to adjust for the wrong variables (Weinberg 1993). Moreover, the popular conception that any important concept (e.g., randomization, confounding, instrumental variables) *must* have a statistical definition is so deeply entrenched in the health sciences that even today, 15 months past the publication of my book, people with the highest qualifications and purest of intentions continue to ask: "So why isn't confounding a statistical concept?"

I believe that any attempt to correct this tradition would necessarily sound sweeping, and nothing but a sweeping campaign can ever eradicate these misconceptions about confounding, statistics, and causation. Statistical education is firmly in the hands of people of Reviewer 1's persuasion. And people with your quest for understanding are rarely given a public forum to ask, "So why isn't confounding a statistical concept?"

The same argument applies to the concepts of “randomization” and “instrumental variables.” (Ironically, Reviewer #1 states authoritatively that these concepts “have clear statistical definitions.” I would hand him a bivariate distribution  $f(x, y)$  and ask to tell us if  $X$  is randomized.) Any definition of these concepts must invoke causal vocabulary, undefined in terms of distributions – there is simply no escape.

And this brings me to reemphasize the usefulness of the statistical–causal demarcation line, as defined in Section 1.5. Those who recognize that concepts such as randomization, confounding, and instrumental variables must rest on causal information would be on guard to isolate and explicate the causal assumptions underlying studies invoking these concepts. In contrast, those who ignore the causal–statistical distinction (e.g., Reviewer #1) will seek no such explication, and will continue to turn a blind eye to “how the usual analysis fails.”

## 11.7 THE CALCULUS OF COUNTERFACTUALS

### 11.7.1 Counterfactuals in Linear Systems

We know that, in general, counterfactual queries of the form  $P(Y_x = y | e)$  may or may not be empirically identifiable, even in experimental studies. For example, the probability of causation,  $P(Y_x = y | x', y')$ , is in general not identifiable from either observational or experimental studies (p. 290, Corollary 9.2.12). The question we address in this section is whether the assumption of linearity renders counterfactual assertions more empirically grounded. The answer is positive:

**Claim A.** Any counterfactual query of the form  $E(Y_x | e)$  is empirically identifiable in linear causal models, with  $e$  an arbitrary evidence.

**Claim B.** Whenever the causal effect  $T$  is identified,  $E(Y_x | e)$  is identified as well.

**Claim C.**  $E(Y_x | e)$  is given by

$$E(Y_x | e) = E(Y | e) + T[x - E(X | e)], \tag{11.28}$$

where  $T$  is the total effect coefficient of  $X$  on  $Y$ , i.e.,

$$T = dE[Y_x]/dx = E(Y | do(x + 1)) - E(Y | do(x)). \tag{11.29}$$

Claim A is not surprising. It has been established in generality by Balke and Pearl (1994b) where  $E(Y_x | e)$  is given explicit expression in terms of the covariance matrix and the structural coefficients; the latter are empirically identifiable.

Claim B renders  $E(Y_x | e)$  observationally identifiable as well, in all models where the causal effect  $E(Y_x)$  is identifiable.

Claim C offers an intuitively compelling interpretation of  $E(Y_x | e)$  that reads as follows: Given evidence  $e$ , to calculate  $E(Y_x | e)$  (i.e., the expectation of  $Y$  under the hypothetical assumption that  $X$  were  $x$ , rather than its current value), first calculate the best estimate of  $Y$  conditioned on the evidence  $e$ ,  $E(Y | e)$ , then add to it whatever change is expected in  $Y$  when  $X$  undergoes a forced transition from its current best estimate,  $E(X | e)$ , to its hypothetical value  $X = x$ . That last addition is none other than the effect coefficient  $T$  times the expected change in  $X$ , i.e.,  $T[x - E(X | e)]$ .

Note: Equation (11.28) can also be written in  $do(x)$  notation as

$$E(Y_x|e) = E(Y|e) + E(Y|do(x)) - E[Y|do(X = E(X|e))]. \quad (11.30)$$

**Proof:**

(With help from Ilya Shpitser)

Assume, without loss of generality, that we are dealing with a zero-mean model. Since the model is linear, we can write the relation between  $X$  and  $Y$  as:

$$Y = TX + I + U, \quad (11.31)$$

where  $T$  is the total effect of  $X$  on  $Y$ , given in (11.29);  $I$  represents terms containing other variables in the model, nondescendants of  $X$ ; and  $U$  represents exogenous variables.

It is always possible to bring the function determining  $Y$  into the form (11.31) by recursively substituting the functions for each r.h.s. variable that has  $X$  as an ancestor, and grouping all the  $X$  terms together to form  $TX$ . Clearly,  $T$  is the Wright-rule sum of the path costs originating from  $X$  and ending in  $Y$  (Wright 1921).

From (11.31) we can write:

$$Y_x = Tx + I + U, \quad (11.32)$$

since  $I$  and  $U$  are not affected by the hypothetical change from  $X = x$ , and, moreover,

$$E(Y_x|e) = Tx + E(I + U|e), \quad (11.33)$$

since  $x$  is a constant.

The last term in (11.33) can be evaluated by taking expectations on both sides of (11.31), giving:

$$E(I + U|e) = E(Y|e) - TE(X|e), \quad (11.34)$$

which, substituted into (11.33), yields

$$E(Y_x|e) = Tx + E(Y|e) - E(X|e) \quad (11.35)$$

and proves our target formula (11.28).  $\square$

Three special cases of  $e$  are worth noting:

**Example 11.7.1**  $e : X = x', Y = y'$  (the linear equivalent of the probability of causation, Chapter 9). From (11.30) we obtain directly

$$E(Y_x | Y = y', X = x') = y' + T(x - x').$$

This is intuitively compelling. The hypothetical expectation of  $Y$  is simply the observed value,  $y'$ , plus the anticipated change in  $Y$  due to the change  $x - x'$  in  $X$ .

**Example 11.7.2**  $e : X = x'$  (the effect of treatment on the treated, Chapter 8.2.5).

$$\begin{aligned} E(Y_x | X = x') &= E(Y | x') + T(x - x') \\ &= rx' + T(x - x') \\ &= rx' + E(Y | do(x)) - E(Y | do(x')), \end{aligned} \quad (11.36)$$

where  $r$  is the regression coefficient of  $Y$  on  $X$ .

**Example 11.7.3**  $e : Y = y'$  (e.g., the expected income  $Y$  of those who currently earn  $Y = y'$  if we were to mandate  $x$  hours of training each month).

$$\begin{aligned} E(Y_x | Y = y') &= y' + T[x - E(X | y')] \\ &= y' + E(Y | do(x)) - E[Y | do(X = r'y')], \end{aligned} \tag{11.37}$$

where  $r'$  is the regression coefficient of  $X$  on  $Y$ .

**Example 11.7.4** Consider the nonrecursive supply-demand model of p. 215, equations (7.9)–(7.10):

$$\begin{aligned} q &= b_1p + d_1i + u_1 \\ p &= b_2q + d_2w + u_2. \end{aligned} \tag{11.38}$$

Our counterfactual problem (p. 216) reads: Given that the current price is  $P = p_0$ , what would be the expected value of the demand  $Q$  if we were to control the price at  $P = p_1$ ?

Making the correspondence  $P = X, Q = Y, e = \{P = p_0, i, w\}$ , we see that this problem is identical to Example 11.7.2 above (effect of treatment on the treated), subject to conditioning on  $i$  and  $w$ . Hence, since  $T = b_1$ , we can immediately write

$$\begin{aligned} E(Q_{p_1} | p_0, i, w) &= E(Y | p_0, i, w) + b_1(p_1 - p_0) \\ &= r_p p_0 + r_i i + r_w w + b_1(p_1 - p_0), \end{aligned} \tag{11.39}$$

where  $r_p, r_i,$  and  $r_w$  are the coefficients of  $P, i$  and  $w$ , respectively, in the regression of  $Q$  on  $P, i,$  and  $w$ .

Equation (11.39) replaces equation (7.17) on page 217. Note that the parameters of the price equation,  $p = b_2q + d_2w + u_2$ , enter (11.39) only via the regression coefficients. Thus, they need not be calculated explicitly in cases where they are estimated directly by least square.

**Remark:** Example 11.7.1 is not really surprising; we know that the probability of causation is empirically identifiable under the assumption of monotonicity (p. 293). But examples 11.7.2 and 11.7.3 trigger the following conjecture:

**Conjecture:**

Any counterfactual query of the form  $P(Y_x | e)$  is empirically identifiable when  $Y$  is monotonic relative to  $X$ .

It is good to end on a challenging note.

**11.7.2 The Meaning of Counterfactuals**

**Question to Author:**

I have a hard time understanding what counterfactuals are actually useful for. To me, they seem to be answering the wrong question. In your book, you give at least a couple of different reasons for when one would need the answer to a counterfactual question, so let me tackle these separately:

1. Legal questions of responsibility. From your text, I infer that the American legal system says that a defendant is guilty if he or she caused the plaintiff's

misfortune. But in my mind, the law is clearly flawed. Responsibility should rest with the predicted outcome of the defendant's action, not with what actually happened. A doctor should not be held responsible if he administers, for a serious disease, a drug which cures 99.99999% of the population but kills 0.00001%, even if he was unlucky and his patient died. If the law is based on the counterfactual notion of responsibility then the law is seriously flawed, in my mind.

2. The use of context in decision making. On p. 217, you write "At this point, it is worth emphasizing that the problem of computing counterfactual expectations is not an academic exercise; it represents in fact the typical case in almost every decision-making situation." I agree that context is important in decision making, but do not agree that we need to answer counterfactual questions.

In decision making, the thing we want to estimate is  $P(\text{future} \mid \text{do}(\text{action}), \text{see}(\text{context}))$ . This is of course a regular *do*-probability, not a counterfactual query. So why do we need to compute counterfactuals?

3. In the latter part of your book, you use counterfactuals to define concepts such as 'the cause of *X*' or 'necessary and sufficient cause of *Y*'. Again, I can understand that it is tempting to mathematically define such concepts since they are in use in everyday language, but I do not think that this is generally very helpful. Why do we need to know 'the cause' of a particular event? Yes, we are interested in knowing 'causes' of events in the sense that they allows us to predict the future, but this is again a case of point (2) above.

To put it in the most simplified form, my argument is the following: Regardless of whether we represent individuals, businesses, organizations, or government, we are constantly faced with decisions of how to act (and these are the only decisions we have!). What we want to know is, what will likely happen if we act in particular ways. So what we want to know is  $P(\text{future} \mid \text{do}(\text{action}), \text{see}(\text{context}))$ . We do not want or need the answers to counterfactuals.

Where does my reasoning go wrong?

### ***Author's Reply:***

1. Your first question doubts the wisdom of using single-event probabilities, rather than population probabilities, in deciding legal responsibility. Suppose there is a small percentage of patients who are allergic to a given drug, and the manufacturer nevertheless distributes the drug with no warning about possible allergic reactions. Wouldn't we agree that when an allergic patient dies he is entitled to compensation? Normally, drug makers take insurance for those exceptional cases, rather than submit the entire population to expensive tests prior to taking the drug – it pays economically. The physician, of course, is exonerated from guilt, for he/she just followed accepted practice. But the law makes sure that someone pays for the injury if one can prove that, counterfactually, the specific death in question would not have occurred had the patient not taken the drug.
2. Your second question deals with decisions that are conditioned on the results of observations. Or, as you put it: "In decision making, the thing we want to estimate is  $P(\text{future} \mid \text{do}(\text{action}), \text{see}(\text{context}))$ ."

The problem is that, to make this prediction properly, we need a sequential, time-indexed model, where “future” variables are clearly distinct from the “context” variables. Often, however, we are given only a static model, in which the “context” variable are shown as *consequences* of the “action,” and then using the expression  $P(y | do(x), z)$  would be inappropriate; it represents the probability of  $Y = y$  given that we do  $X = x$  and *later* observe  $Z = z$ , while what is needed is the probability of  $Y = y$  given that we first observe  $Z = z$  and then do  $X = x$ . Counterfactuals give us a way of expressing the desired probability, by writing

$$P = P(y_x | z),$$

which stands for the probability that  $Y = y$  would occur had  $X$  been  $x$ , given that  $Z = z$  is currently observed. Note that  $P = P(y_x | z) = P(y | do(x), z)$  if  $Z$  is a non-descendant of  $X$  in the static model.

An example will be instructive. Suppose an engineer draws a circuit diagram  $M$  containing a chain of gates  $X \rightarrow Y \rightarrow Z$ . At time  $t_1$  we observe  $Z(t_1)$ , and we want to know the causal effect of  $X(t_2)$  on  $Y(t_3)$  conditioned on  $Z(t_1)$ . We can do this exercise through *do*-calculus, with all the necessary time indices, if we replicate the model  $M$  and assign each time slot  $t_i$ , a model  $M_i$ , showing the relationships among  $X(t_i), Y(t_i), Z(t_i)$  as well as to previous variables, then compute  $P(y(t_3) | do(x(t_2)), z(t_1))$ . But we can do it much better in the static model, using the counterfactuals  $P(Y_x = y | z)$ . The static model saves us miles and miles of drawing the sequential model equivalent, and counterfactuals enable us to take advantage of this savings. It is an admirable invention. One can argue, of course, that if counterfactual claims are merely “conversational shorthand” (p. 118) for scientific predictions in the sequential model equivalent, then they are not needed at all. But this would stand contrary to the spirit of scientific methods, where symbolic economy plays a crucial role. In principle, multiplication is not needed in algebra – we can live with addition alone and add a number to itself as many times as needed. But science would not have advanced very far without multiplication. Same with counterfactuals.

3. In the final analysis, the reason we want to know “causes of events” is indeed to allow us to better predict and better act in the future. But we do not always know in advance under what future circumstances our knowledge will be summoned to help in making decisions. The reason we need to know the cause of a specific accident may be multifaceted: to warn the public against the use of similar vehicles, to improve maintenance procedures of certain equipment, to calculate the cost of taking remedial action, to cause similar accidents in enemy territories, and many more. Each of these applications may demand different information from our model, so “causes of events” may be a useful way of registering our experience so as to amortize it over the wide spectrum of possible applications.

### 11.7.3 *d*-Separation of Counterfactuals

#### *Question to Author:*

I am trying to generalize the twin network method of Figure 7.3 to cases where counterfactuals involve more than two possible worlds. Consider the causal model  $X \rightarrow Z \rightarrow Y$ ,

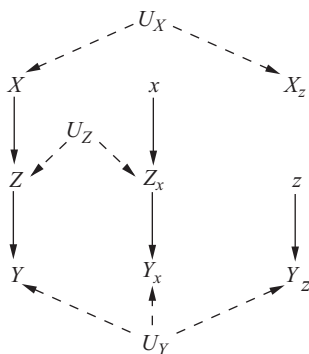


Figure 11.18 Triple network drawn to test equation (11.40).

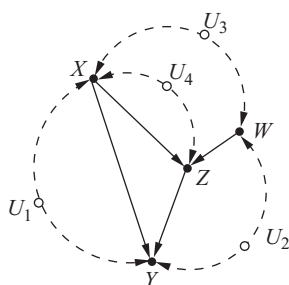


Figure 11.19 Graph in which  $Y_{xz} \perp\!\!\!\perp Z_{x^*} \mid W$  holds by virtue of the equality  $W_{x^*} = W$ .

and assume we wish to test whether the assertion

$$Y_x \perp\!\!\!\perp X \mid Y_z, Z_x, Y \tag{11.40}$$

is true in the model. I would naively construct the following “triple network”:

The left part corresponds to a world in which no intervention is imposed; the middle part to a world in which  $do(X = x)$  is imposed; and in the right part,  $do(Z = z)$  is imposed. In this network (11.40) does not follow from  $d$ -separation, since the path from  $Y_x$  to  $X$  is open by conditioning on  $Y$ . Is there anything wrong with this way of generalizing the twin network method?

**Author’s Reply (with Ilya Shpitser):**

Your generalization of the twin network to more than two worlds is correct, and so is your conclusion;  $Y_x$  is not independent of  $X$  given  $Y_z, Z_x, Y$ . In fact, a recent paper (Shpitser and Pearl 2007) formulates this graphical representation of counterfactuals in multiple worlds under the rubric “counterfactual graphs.”

**Follow-up Question to Author:**

Your answer helped me understand the twin network method, as well as the one used in the paper “Direct and Indirect Effects” (Pearl 2001c), but it raises a new question: why does  $Y_{xz} \perp\!\!\!\perp Z_{x^*} \mid W$  hold in the graph of Figure 11.19, as claimed in (Pearl 2001)? If one draws a “triple network,” it is obvious that the path between  $Y_{xz}$  and  $Z_{x^*}$  is not blocked by  $W$ ; the latter is situated in a totally different world.

**Author's Second Reply:**

The independence  $Y_{xz} \perp\!\!\!\perp Z_{x^*} \mid W$  actually holds in the graph shown in Figure 11.19.

This is because  $Y_{xz}$  is separated from  $Z_{x^*}$  by the variable  $W_{x^*}$ , in the “triple network” that you mentioned. The license to replace  $W$  with  $W_{x^*}$  is obtained from rule 3 of *do*-calculus, which implies  $W_{x^*} = W$ , since  $X$  is a nondescendant of  $X$ . This points to an important refinement needed in the twin network generalization: causal axioms may entail certain equality constraints among seemingly distinct counterfactual variables, and these hidden equalities need to be considered when we apply *d*-separation to counterfactual networks. A systematic way of encoding and managing these equalities is presented in Shpitser and Pearl (2007).

**11.8 INSTRUMENTAL VARIABLES AND NONCOMPLIANCE****11.8.1 Tight Bounds under Noncompliance (Question to Author)**

I am referring to the way you improved Manski's bounds on treatment effects when we have imperfect compliance. Which information does your approach exploit that the one by Manski does not? What is the intuition behind it?

**Author's Reply:**

We used the same information and same assumptions as Manski, and we derived the tight bounds using linear-programming analysis in the 16-dimensional space defined by canonical partition of  $U$  (Balke and Pearl 1994a, 1995a). Manski, evidently, either did not aim at getting tight bounds, or was not aware of the power of partitioning  $U$  into its equivalence classes. Recall, this partition was unknown to economists before Frangakis and Rubin (2002) popularized it, under the rubric “principal stratification.”

Manski's bounds, as I state on page 269, are tight under certain conditions, e.g., no contrarians. This means that one can get narrower bounds *only* when there are contrarians in the population, as in the examples discussed in Pearl (1995b). It is shown there how data representing the presence of contrarians can provide enough information to make the bounds collapse to a point. That article also gives an intuitive explanation of how this can happen.

It is important to mention at this point that the canonical partition conception, coupled with the linear programming method developed in Balke and Pearl (1994a, 1995a,b), has turned into a powerful analytical tool in a variety of applications. Tian and Pearl (2000) applied it to bound probabilities of causation; Kaufman et al. (2005) and Cai et al. (2008) used it to bound direct effects in the presence of confounded mediation, and, similarly, Imai et al. (2008) used it to bound natural direct and indirect effects. The closed-form expressions derived by this method enable researchers to assess what features of the distribution are critical for narrowing the widths of the bounds.

Rubin (2004), in an independent exploration, attempted to apply canonical partitions to the analysis of direct and indirect effects within the traditional potential-outcome framework but, lacking the graphical and structural perspectives, was led to conclude that such effects are “ill-defined” and “more deceptive than helpful.” I believe readers of this book, guided by the structural roots of potential-outcome analysis, will reach more positive conclusions (see Sections 4.5 and 11.4.2).

## 11.9 MORE ON PROBABILITIES OF CAUSATION

Looking back, eight years past the first publication of *Causality*, I consider the results obtained in Chapter 9 to be a triumph of counterfactual analysis, dispelling all fears and trepidations that some researchers express concerning the empirical content of counterfactuals (Dawid 2000; Pearl 2000). It demonstrates that a quantity PN which at first glance appears to be hypothetical, ill-defined, untestable, and hence unworthy of scientific analysis is nevertheless definable, testable, and, in certain cases, even identifiable. Moreover, the fact that, under certain combination of data, and making no assumptions whatsoever, an important legal claim such as “the plaintiff would be alive had he not taken the drug” can be ascertained *with probability one* is truly astonishing.

### 11.9.1 Is “Guilty with Probability One” Ever Possible?

I have presented the example of Section 9.3.4 in dozens of talks at conferences and universities, and, invariably, the result  $PN \geq 1$  of equation (9.53) meets with universal disbelief; how can we determine, from frequency data, that the defendant is guilty with probability one, i.e., that Mr. A would *definitely* be alive today had he not taken the drug. Professor Stephen Fienberg attended two of these talks, and twice shook his head with: “It can’t be true.” To me, this reaction constitutes a solid proof that counterfactual analysis can yield nontrivial results, and hence that it is real, meaningful, and useful; metaphysical analysis would not evoke such resistance.

What causes people to disbelieve this result are three puzzling aspects of the problem:

1. that a hypothetical, generally untestable quantity can be ascertained with probability one under certain conditions;
2. that frequency tables which, individually, do not reveal a substantial effect of the drug imply a perfect susceptibility to the drug when taken together; and
3. that a property of an untested individual can be assigned a probability one, on the basis of data taken from a sampled population.

The first puzzle is not really surprising for students of science who take seriously the benefits of logic and mathematics. Once we give a quantity formal semantics we essentially define its relation to the data, and it not inconceivable that data obtained under certain conditions would sufficiently constrain that quantity, to a point where it can be determined exactly.

This benefit of formal counterfactual analysis can in fact be demonstrated in a much simpler example. Consider the effect of treatment on the treated,  $P(y_{x'} | x)$  (Section 8.2.5). In words,  $P(y_{x'} | x)$  stands for the chances that a treated patient,  $X = x$ , would survive for a time period  $Y = y$  had he/she not been treated ( $X = x'$ ). This counterfactual quantity too seems to defy empirical measurement, because we can never rerun history and deny treatment for those who received it. Yet, for binary  $X$ , we can write

$$P(y_{x'}) = P(y_{x'} | x')P(x') + P(y_{x'} | x)P(x)$$

and derive

$$P(y_{x'} | x) = [P(y_{x'}) - P(y, x')]/P(x).$$

In other words,  $P(y_{x'} | x)$  is reducible to empirically estimable quantities;  $P(y_{x'}) = P(y | do(x'))$  is estimable in experimental studies and the other quantities in observational studies. Moreover, if data support the equality  $P(y_{x'}) = P(y, x')$ , we can safely conclude that a treated patient would have zero chance of survival had the treatment not been taken. Those who mistrust counterfactual analysis a priori, as a calculus dealing with undefined quantities, would never enjoy the discovery that some of those quantities are empirically definable. Logic, when gracious, can rerun history for us.

The second puzzle was given intuitive explanation in the paragraph following equation (9.54).

The third puzzle is the one that gives most people a shock of disbelief. For a statistician, in particular, it is a rare case to be able to say anything certain about a specific individual who was not tested directly. This emanates from two factors. First, statisticians normally deal with finite samples, the variability of which rules out certainty in any claim, not merely about an individual but also about any property of the underlying distribution. This factor, however, should not enter into our discussion, for we have been assuming infinite samples throughout. (Readers should imagine that the numbers in Table 9.2 stand for millions.)

The second factor emanates from the fact that, even when we know a distribution precisely, we cannot assign a definite probabilistic estimate to a property of a specific individual drawn from that distribution. The reason is, so the argument goes, that we never know, let alone measure, all the anatomical and psychological variables that determine an individual's behavior, and, even if we knew, we would not be able to represent them in the crude categories provided by the distribution at hand. Thus, because of this inherent crudeness, the sentence "Mr. A would be dead" can never be assigned a probability of one (or, in fact, any definite probability).

This argument, advanced by Freedman and Stark (1999), is incompatible with the way probability statements are used in ordinary discourse, for it implies that every probability statement about an individual must be a statement about a restricted subpopulation that shares *all* the individual's characteristics. Taken to the extreme, such a restrictive interpretation would insist on characterizing the plaintiff in minute detail, and would reduce PN to zero or one when all relevant details were accounted for. It is inconceivable that this interpretation underlies the intent of judicial standards. By using the wording "more probable than not," lawmakers have instructed us to ignore specific features for which data is not available, and to base our determination on the most specific features for which reliable data is available. In our example, two properties of Mr. A were noted: (1) that he died and (2) that he chose to use the drug; these were properly taken into account in bounding PN. If additional properties of Mr. A become known, and deemed relevant (e.g., that he had red hair, or was left-handed), these too could, in principle, be accounted for by restricting the analysis to data representing the appropriate subpopulations. However, in the absence of such data, and knowing in advance that we will never be able to match *all* the idiosyncratic properties of Mr. A, the lawmakers' specification must be interpreted relative to the properties at hand.

### 11.9.2 Tightening the Bounds on Probabilities of Causation

Systematic work by Jin Tian (Tian and Pearl 2000) has improved on the results of Chapter 9 in several ways. First, Tian showed that for most of these results, the assumption of strong exogeneity, equation (9.10), can be replaced by weak exogeneity:

$$Y_x \perp\!\!\!\perp X \quad \text{and} \quad Y_{x'} \perp\!\!\!\perp X.$$

Second, the estimands obtained under the assumption of monotonicity (Definition 9.2.13) constitute *lower bounds* when monotonicity is not assumed. Finally, the bounds derived in Chapter 9 are *sharp*, that is, they cannot be improved without strengthening the assumptions.

Of particular interest are the bounds obtained when data from both experimental and nonexperimental studies are available, and no other assumptions are made. These read:

$$\max \left\{ \begin{array}{c} 0 \\ P(y_x) - P(y_{x'}) \\ P(y) - P(y_{x'}) \\ P(y_x) - P(y) \end{array} \right\} \leq PNS \leq \min \left\{ \begin{array}{c} P(y_x) \\ P(y_{x'}) \\ P(x, y) + P(x', y') \\ P(y_x) - P(y_{x'}) + P(x, y') + P(x', y) \end{array} \right\} \quad (11.41)$$

$$\max \left\{ \begin{array}{c} 0 \\ \frac{P(y) - P(y_{x'})}{P(x, y)} \end{array} \right\} \leq PN \leq \min \left\{ \begin{array}{c} 1 \\ \frac{P(y_{x'}) - P(x', y')}{P(x, y)} \end{array} \right\} \quad (11.42)$$

$$\max \left\{ \begin{array}{c} 0 \\ \frac{P(y_x) - P(y)}{P(x', y')} \end{array} \right\} \leq PS \leq \min \left\{ \begin{array}{c} 1 \\ \frac{P(y_x) - P(x, y)}{P(x', y')} \end{array} \right\} \quad (11.43)$$

It is worth noting that, in drug-related litigation, it is not uncommon to obtain data from both experimental and observational studies. The former is usually available at the manufacturer or the agency that approved the drug for distribution (e.g., FDA), while the latter is easy to obtain by random surveys of the population. In such cases, the standard lower bound used by epidemiologists to establish legal responsibility, the Excess Risk Ratio (equation (9.22)), can be substantially improved by using the lower bound of equation (11.42). Likewise, the upper bound of equation (11.42) can be used to exonerate the drug maker from legal responsibility. Cai and Kuroki (2006) analyzed the statistical properties of PNS, PN, and PS.

Also noteworthy is the fact that in tasks of abduction, i.e., reasoning to the best explanation, PN is the most reasonable measure to use in deciding among competing explanations of the event  $Y = y$ . In such applications, the bounds given in (11.42) can be computed from a causal Bayesian network model, where  $P(y_x)$  and  $P(y_{x'})$  are computable through equation (9.48).

### Acknowledgments

I thank all readers who sent in questions and contributed in this way to this chapter. These include: David Bessler, Nimrod Megiddo, David Kenny, Keith A. Markus, Jos

Lehmann, Dennis Lindley, Jacques A. Hagenaars, Jonathan Wilson, Stan Mulaik, Bill Shipley, Nozer D. Singpurwalla, Les Hayduk, Erich Battistin, Sampsa Hautaniemi, Melanie Wall, Susan Scott, Patrik Hoyer, Joseph Halpern, Phil Dawid, Sander Greenland, Arvid Sjolander, Eliezer S. Yudkowsky, UCLA students in CS262Z (Seminar in Causality, Spring 2006), and the UCLA Epidemiology Class – EPIDEM 200C.

I similarly thank all reviewers of the first edition of the book and the editors who helped bring these reviews to the attention of their readers. These include: *Choice* (Byerly 2000), *Structural Equation Modeling* (Shipley 2000a), *Chance* (McDonald 2001), *Technometrics* (Zelterman 2001), *Mathematical Reviews* (Lawry 2001), *Politische Vierteljahrsschrift* (Didelez and Pigeot 2001), *Technological Forecasting & Social Change* (Payson 2001), *British Journal for the Philosophy of Science* (Gillies 2001), *Human Biology* (Chakraborty 2001), *The Philosophical Review* (Hitchcock 2001), *Intelligence* (O'Rourke 2001), *Journal of Marketing Research* (Rigdon 2002), *Tijdschrift Voor* (Decock 2002), *Psychometrika* (McDonald 2002b), *International Statistical Review* (Lindley 2002), *Journal of Economic Methodology* (Leroy 2002), *Statistics in Medicine* (Didelez 2002), *Journal of Economic Literature* (Swanson 2002), *Journal of Mathematical Psychology* (Butler 2002), *IIE Transactions* (Gursoy 2002), *Royal Economic Society* (Hoover 2003), *Econometric Theory* (Neuberg 2003), *Economica* (Abbring 2003), *Economics and Philosophy* (Woodward 2003), *Sociological Methods and Research* (Morgan 2004), *Review of Social Economy* (Boumans 2004), *Journal of the American Statistical Association* (Hadlock 2005), and *Artificial Intelligence* (Kyburg 2005).

Thanks also go to the contributors to UCLA's *Causality* blog (<http://www.mii.ucla.edu/causality/>) and to William Hsu, the blog's curator.

A special word of appreciation and admiration goes to Dennis Lindley, who went to the trouble of studying my ideas from first principles, and allowed me to conclude that readers from the statistics-based sciences would benefit from this book. I am fortunate that our paths have crossed and that I was able to witness the intellect, curiosity, and integrity of a true gentleman.

This chapter could not have been written without the support and insight of Jin Tian, Avin Chen, Carlo Brito, Blai Bonet, Mark Hopkins, Ilya Shpitser, Azaria Paz, Manabu Kuroki, Zhihong Cai, Kaoru Mulvihill, and all members of the Cognitive System Laboratory at UCLA who, in the past six years, continued to explore the green pastures of causation while I was summoned to mend a world that took the life of my son Daniel (murdered by extremists in Karachi, Pakistan, 2002). These years have convinced me, beyond a shadow of doubt, that the forces of reason and enlightenment will win over fanaticism and inhumanity.

Finally, I dedicate this edition to my wife, Ruth, for her strength, love, and guidance throughout our ordeal, to my daughters, Tamara and Michelle, and my grandchildren, Leora, Torri, Adam, Ari, and Evan for being with me through this journey and making it meaningful and purposeful.

