Chapter 3

Causal Diagrams and the Identification of Causal Effects

The eye obeys exactly the action of the mind.
Emerson (1860)

Preface

In the previous chapter we dealt with ways of learning causal relationships from raw data. In this chapter we explore the ways of learning such relationships from a combination of data and qualitative causal assumptions that are deemed plausible in a given domain. More broadly, this chapter aims to help researchers communicate qualitative assumptions about cause-effect relationships, elucidate the ramifications of such assumptions, and derive causal inferences from a combination of assumptions, experiments, and data. Our major task will be to decide whether the assumptions given are sufficient for assessing the strength of causal effects from nonexperimental data.

Causal effects permit us to predict how systems would respond to hypothetical interventions—for example, policy decisions or actions performed in everyday activity. As we have seen in Chapter 1 (Section 1.3), such predictions are the hallmark of causal modeling, since

they are not discernible from probabilistic information alone; they rest on—and, in fact, define—causal relationships. This chapter uses causal diagrams to give formal semantics to the notion of *intervention*, and it provides explicit formulas for postintervention probabilities in terms of preintervention probabilities. The implication is that the effects of every intervention can be estimated from nonexperimental data, provided the data is supplemented with a causal diagram that is both acyclic and contains no latent variables.

If some variables are not measured then the question of identifiability arises, and this chapter develops a nonparametric framework for analyzing the identification of causal relationships in general and causal effects in particular. We will see that causal diagrams provide a powerful mathematical tool in this analysis; they can be queried, using extremely simple tests, to determine if the assumptions available are sufficient for identifying causal effects. If so, the diagrams produce mathematical expressions for causal effects in terms of observed distributions; otherwise, the diagrams can be queried to suggest additional observations or auxiliary experiments from which the desired inferences can be obtained.

Another tool that emerges from the graphical analysis of causal effects is a *calculus of interventions*—a set of inference rules by which sentences involving interventions and observations can be transformed into other such sentences, thus providing a syntactic method of deriving (or verifying) claims about interventions and the way they interact with observations. With the help of this calculus the reader will be able to (i) determine mathematically whether a given set of covariates is appropriate for control of confounding, (ii) deal with measurements that lie on the causal pathways, and (iii) trade one set of measurements for another.

Finally, we will show how the new calculus disambiguates concepts that have triggered controversy and miscommunication among philosophers, statisticians, economists, and psychologists. These include distinctions between structural and regression equations, definitions of direct and indirect effects, and relationships between structural equations and Neyman-Rubin models.

3.1 Introduction

The problems addressed in this chapter can best be illustrated through a classical example due to Cochran (see Wainer 1989). Consider an experiment in which soil fumigants (X) are used to increase oat crop yields (Y) by controlling the eelworm population (Z); the fumigants may also have direct effects (both beneficial and adverse) on yields beside the control of eelworms. We wish to assess the total effect of the fumigants on yields when this typical study is complicated by several factors. First, controlled randomized experiments are unfeasible—farmers insist on deciding for themselves which plots are to be fumigated. Second, farmers' choice of treatment depends on last year's eelworm population (Z_0), an unknown quantity that is strongly correlated with this year's population. Thus we have a classical case of confounding bias that interferes with the assessment of treatment effects regardless of sample size. Fortunately, through laboratory analysis of soil samples, we can determine the eelworm populations before and after the treatment; furthermore, because the fumigants are known to be active for a short period only, we can safely assume that they do not affect the growth of eelworms surviving the treatment. Instead, eelworms' growth depends on the population of birds (and other predators), which is correlated with last year's eelworm population and hence with the treatment itself.

The method developed in this chapter permits the investigator to translate complex considerations of this sort into a formal language and thereby facilitate the following tasks:

1. explicating the assumptions that underlie the model;
2. deciding whether the assumptions are sufficient to obtain consistent estimates of the target quantity—the total effect of the fumigants on yields;
3. providing (if the answer to item 2 is affirmative) a closed-form expression for the target quantity in terms of distributions of observed quantities; and
4. suggesting (if the answer to item 2 is negative) a set of observations and experiments that, if performed, would render a consistent estimate feasible.

The first step in this analysis is to construct a causal diagram like the one given in Figure 3.1, which represents the investigator's under-

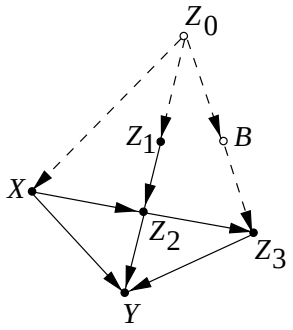

Figure 3.1: A causal diagram representing the effect of fumigants (X) on yields (Y).

standing of the major causal influences among measurable quantities in the domain. For example, the quantities Z_1, Z_2, Z_3 represent the eelworm population before treatment, after treatment, and at the end of the season, respectively. The Z_0 term represents last year's eelworm population; because it is an unknown quantity, it is denoted by a hollow circle, as is the quantity B, the population of birds and other predators. Links in the diagram are of two kinds: those that connect unmeasured quantities are designated by dashed arrows, those connecting measured quantities by solid arrows. The substantive assumptions embodied in the diagram are negative causal assertions which are conveyed through the links *missing* from the diagram. For example, the missing arrow between Z_1 and Y signifies the investigator's understanding that pretreatment eelworms can not affect oat plants directly; their entire influence on oat yields is mediated by the posttreatment conditions, Z_2 and Z_3. Our purpose is not to validate or repudiate such domain-specific assumptions but rather to test whether a given set of assumptions is sufficient for quantifying causal effects from nonexperimental data—here, estimating the total effect of fumigants on yields.

The causal diagram in Figure 3.1 is similar in many respects to the path diagrams devised by Wright (1921); both reflect the investigator's subjective and qualitative knowledge of causal influences in the domain, both employ directed acyclic graphs, and both allow for the incorpora-

tion of latent or unmeasured quantities. The major differences lie in the method of analysis. First, whereas path diagrams have been analyzed mostly in the context of linear models with Gaussian noise, causal diagrams permit arbitrary nonlinear interactions. In fact, our analysis of causal effects will be entirely nonparametric, entailing no commitment to a particular functional form for equations and distributions. Second, causal diagrams will be used not only as a passive language to communicate assumptions but also as an active computational device through which the desired quantities are derived. For example, the method to be described allows an investigator to inspect the diagram of Figure 3.1 and make the following immediate conclusions.

1. The total effect of X on Y can be estimated consistently from the observed distribution of $X,\ Z_1,\ Z_2,\ Z_3$, and Y.

2. The total effect of X on Y (assuming discrete variables throughout) is given by the formula[1]

$$P(y|\hat{x}) = \sum_{z_1}\sum_{z_2}\sum_{z_3} P(y|z_2,\ z_3,\ x)P(z_2|z_1,\ x) \times \sum_{x'} P(z_3|z_1,\ z_2,\ x')P(z_1,\ x'), \tag{3.1}$$

   where $P(y|\hat{x})$ stands for the probability of achieving a yield level of $Y = y$, given that the treatment is set to level $X = x$ by external intervention.

3. A consistent estimation of the total effect of X on Y would not be feasible if Y were confounded with Z_3; however, confounding Z_2 and Y will not invalidate the formula for $P(y|\hat{x})$.

These conclusions will be obtained either by analyzing the graphical properties of the diagram or by performing a sequence of symbolic derivations (governed by the diagram) that gives rise to causal effect formulas such as (3.1).

[1] The notation $P_x(y)$ was used in Chapter 1; it is changed henceforth to $P(y|\hat{x})$ or $P(y|do(x))$ because of the inconvenience in handling subscripts. The reader need not be intimidated if, at this point, (3.1) appears unfamiliar. After reading Section 3.4, the reader should be able to derive such formulas with greater ease than solving algebraic equations. Note that x' is merely an index of summation that ranges over the values of X.

3.2 Intervention in Markovian Models

3.2.1 Graphs as Models of Interventions

In Chapter 1 (Section 1.3) we saw how causal models, unlike probabilistic models, can serve to predict the effect of interventions. This added feature requires that the joint distribution P be supplemented with a causal diagram—that is, a directed acyclic graph G that identifies the causal connections among the variables of interest. In this section we elaborate on the nature of interventions and give explicit formulas for their effects.

The connection between the causal and associational readings of DAGs is formed through the mechanism-based account of causation, which owes its roots to early works in econometrics (Frisch 1938; Haavelmo 1943; Simon 1953). In this account, assertions about causal influences, such as those specified by the links in Figure 3.1, stand for autonomous physical mechanisms among the corresponding quantities; these mechanisms are represented as functional relationships perturbed by random disturbances. Echoing this tradition, Pearl and Verma (1991) interpreted the causal reading of a DAG in terms of functional, rather than probabilistic, relationships (see (1.40) and Definition 2.2.2); in other words, each child-parent family in a DAG G represents a deterministic function

$$x_i = f_i(pa_i, \epsilon_i), \qquad i = 1, ..., n, \tag{3.2}$$

where pa_i are the parents of variable X_i in G; the ϵ_i $(1 \leq i \leq n)$ are mutually independent, arbitrarily distributed random disturbances. These disturbance terms represent independent background factors that the investigator chooses not to include in the analysis. If any of these factors is judged to be influencing two or more variables (thus violating the independence assumption), then that factor must enter the analysis as an unmeasured (or latent) variable and be represented in the graph by a hollow node, such as Z_0 and B in Figure 3.1. For example, the causal assumptions conveyed by the model in Figure 3.1 correspond to

the following set of equations:

$$\begin{aligned} Z_0 &= f_0(\epsilon_0), & B &= f_B(Z_0, \epsilon_B), \\ Z_1 &= f_1(Z_0, \epsilon_1), & X &= f_X(Z_0, \epsilon_X), \\ Z_2 &= f_2(X, Z_1, \epsilon_2), & Y &= f_Y(X, Z_2, Z_3, \epsilon_Y), \\ Z_3 &= f_3(B, Z_2, \epsilon_3). \end{aligned} \tag{3.3}$$

More generally, we may lump together all unobserved factors (including the ϵ_i) into a set U of background variables and then summarize their characteristics by a distribution function $P(u)$—or by some aspects (e.g. independencies) of $P(u)$. Thus, a full specification of a causal model would entail two components: a set of functional relationships

$$x_i = f_i(pa_i, u_i), \qquad i = 1, \ldots, n, \tag{3.4}$$

and a joint distribution function $P(u)$ on the background factors. If the diagram $G(M)$ associated with a causal model M is acyclic, then M is called *semi-Markovian*. If, in addition, the background variables are independent, M is called *Markovian*, since the resulting distribution of the observed variables would then be Markov relative to $G(M)$ (see Theorem 1.4.1). Thus, the model described in Figure 3.1 is semi-Markovian if the observed variables are $\{X, Y, Z_1, Z_2, Z_3\}$; it would turn Markovian if Z_0 and B were observed as well. In Chapter 7 we will pursue the analysis of general non-Markovian models, but in this chapter all models are assumed to be either Markovian or Markovian with unobserved variables (i.e. semi-Markovian).

Needless to state, we would seldom be in possession of $P(u)$ or even f_i. It is important nevertheless to explicate the mathematical content of a fully specified model in order to draw valid inferences from partially specified models, such as the one described in Figure 3.1.

The equational model (3.2) is the nonparametric analog of the so-called structural equations model (Wright 1921; Goldberger 1973), except that: the functional form of the equations (as well as the distribution of the disturbance terms) will remain unspecified. The equality signs in structural equations convey the asymmetrical counterfactual relation of "is determined by," and each equation represents a stable autonomous mechanism. For example, the equation for Y states that, regardless of what we currently observe about Y and regardless of any

changes that might occur in other equations, if variables $(X, Z_2, Z_3, \epsilon_Y)$ were to assume the values $(x, z_2, z_3, \epsilon_Y)$, respectively, then Y would take on the value dictated by the function f_Y.

Recalling our discussion in Section 1.4, the functional characterization of each child-parent relationship leads to the same recursive decomposition of the joint distribution that characterizes Bayesian networks:

$$P(x_1, ..., x_n) = \prod_i P(x_i \mid pa_i), \tag{3.5}$$

which, in our example of Figure 3.1, yields

$$\begin{aligned} P(z_0, x, z_1, b, z_2, z_3, y) &= P(z_0)P(x|z_0)P(z_1|z_0)P(b|z_0) \\ &\quad \times P(z_2|x, z_1)P(z_3|z_2, b)P(y|x, z_2, z_3) \end{aligned} \tag{3.6}$$

Moreover, the functional characterization provides a convenient language for specifying how the resulting distribution would change in response to external interventions. This is accomplished by encoding each intervention as an alteration on a select subset of functions while keeping the other functions intact. Once we know the identity of the mechanisms altered by the intervention and the nature of the alteration, the overall effect of the intervention can be predicted by modifying the corresponding equations in the model and using the modified model to compute a new probability function.

The simplest type of external intervention is one in which a single variable, say X_i, is forced to take on some fixed value x_i. Such an intervention, which we call "atomic," amounts to lifting X_i from the influence of the old functional mechanism $x_i = f_i(pa_i, u_i)$ and placing it under the influence of a new mechanism that sets the value x_i while keeping all other mechanisms unperturbed. Formally, this atomic intervention, which we denote by $do(X_i = x_i)$, or $do(x_i)$ for short,[2] amounts

[2] An equivalent notation, using $set(x)$ instead of $do(x)$, was used in Pearl (1995a). The $do(x)$ notation was first used in Goldszmidt and Pearl (1992) and is gaining in popular support. The expression $P(y|do(x))$ is equivalent in intent to $P(Y_x = y)$ in the potential-outcome model introduced by Neyman (1923) and Rubin (1974) and to the expression $P[(X = x) \; \square\!\!\rightarrow \; (Y = y)]$ in the counterfactual theory of Lewis (1973b). The semantical differences among these notions are discussed in Section 3.6.3 and in Chapter 7.

to removing the equation $x_i = f_i(pa_i, u_i)$ from the model and substituting $X_i = x_i$ in the remaining equations. The new model thus created represents the system's behavior under the intervention $do(X_i = x_i)$ and, when solved for the distribution of X_j, yields the causal effect of X_i on X_j, which is denoted $P(x_j|\hat{x}_i)$. More generally, when an intervention forces a subset X of variables to attain fixed values x, then a subset of equations is to be pruned from the model given in (3.4), one for each member of X, thus defining a new distribution over the remaining variables that completely characterizes the effect of the intervention.[3]

Definition 3.2.1 (Causal Effect)
Given two disjoint sets of variables, X and Y, the causal effect *of X on Y, denoted either as $P(y|\hat{x})$ or as $P(y|do(x))$, is a function from X to the space of probability distributions on Y. For each realization x of X, $P(y|\hat{x})$ gives the probability of $Y = y$ induced by deleting from the model of* (3.4) *all equations corresponding to variables in X and substituting $X = x$ in the remaining equations.*

Clearly, the graph corresponding to the reduced set of equations is a subgraph of G from which all arrows entering X have been pruned (Spirtes et al. 1993). The difference $E(Y|do(x')) - E(Y|do(x''))$ is sometimes taken as the definition of "causal effect" (Rosenbaum and Rubin 1983), where x' and x'' are two distinct realizations of X. This difference can always be computed from the general function $P(y|do(x))$, which is defined for every level x of X and provides a more refined characterization of the effect of interventions.

3.2.2 Interventions as Variables

An alternative (but sometimes more appealing) account of intervention treats the force responsible for the intervention as a variable within the

[3]The basic view of interventions as equation modifiers originates with Marschak (1950) and Simon (1953). An explicit translation of interventions to "wiping out" equations from the model was first proposed by Strotz and Wold (1960) and later used in Fisher (1970) and Sobel (1990). Graphical ramifications of this translation were explicated first in Spirtes et al. (1993) and later in Pearl (1993b).

system (Pearl 1993b). This is facilitated by representing the function f_i itself as a value of a variable, F_i and then writing (3.2) as

$$x_i = I(pa_i, f_i, u_i), \tag{3.7}$$

where I is a three-argument function satisfying

$$I(a, b, c) = f_i(a, c) \text{ whenever } b = f_i.$$

This amounts to conceptualizing the intervention as an external force F_i that alters the function f_i between X_i and its parents. Graphically, we can represent F_i as an added parent node of X_i, and the effect of such an intervention can be analyzed by standard conditionalization—that is, by conditioning our probability on the event that variable F_i attains the value f_i.

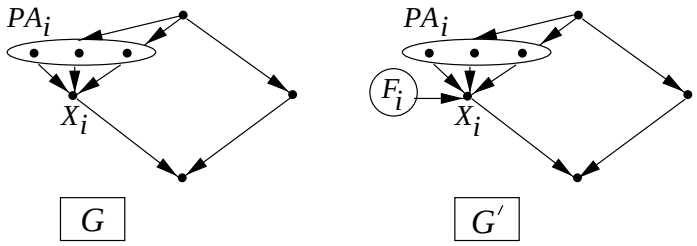

Figure 3.2: Representing external intervention F_i by an augmented network $G' = G \cup \{F_i \to X_i\}$.

The effect of an atomic intervention $do(X_i = x_i')$ is encoded by adding to G a link $F_i \longrightarrow X_i$ (see Figure 3.2), where F_i is a new variable taking values in $\{do(x_i'),\ \text{idle}\}$, x_i' ranges over the domain of X_i, and "idle" represents no intervention. Thus, the new parent set of X_i in the augmented network is $PA_i' = PA_i \cup \{F_i\}$, and it is related to X_i by the conditional probability

$$P(x_i \mid pa_i') = \begin{cases} P(x_i \mid pa_i) & \text{if } F_i = \text{idle}, \\ 0 & \text{if } F_i = do(x_i') \text{ and } x_i \neq x_i', \\ 1 & \text{if } F_i = do(x_i') \text{ and } x_i = x_i'. \end{cases} \tag{3.8}$$

The effect of the intervention $do(x_i')$ is to transform the original probability function $P(x_1, ..., x_n)$ into a new probability function

$P(x_1, ..., x_n|\hat{x}_i')$, given by

$$P(x_1, ..., x_n|\hat{x}_i') = P'(x_1, ..., x_n \mid F_i = do(x_i')), \tag{3.9}$$

where P' is the distribution specified by the augmented network $G' = G \cup \{F_i \rightarrow X_i\}$ and (3.8), with an arbitrary prior distribution on F_i. In general, by adding a hypothetical intervention link $F_i \rightarrow X_i$ to each node in G, we can construct an augmented probability function $P'(x_1, ..., x_n; F_1, ..., F_n)$ that contains information about richer types of interventions. Multiple interventions would be represented by conditioning P' on a subset of the F_i (taking values in their respective $do(x_i')$ domains), and the preintervention probability function P would be viewed as the posterior distribution induced by conditioning each F_i in P' on the value "idle."

One advantage of the augmented network representation is that it is applicable to *any* change in the functional relationship f_i and not merely to the replacement of f_i by a constant. It also displays clearly the ramifications of spontaneous changes in f_i, unmediated by external control. Figure 3.2 predicts, for example, that only descendants of X_i would be effected by changes in f_i and hence the marginal probability $P(z)$ will remain unaltered for every set Z of nondescendants of X_i. Likewise, Figure 3.2 dictates that the conditional probability $P(y|x_i)$ remains invariant to changes in f_i for any set Y of descendants of X_i, provided X_i d-separates F_i from Y. Kevin Hoover (1990, 1999) used such invariant features to determine the direction of causal influences among economic variables (e.g., employment and money supply) by observing the changes induced by sudden modifications in the processes that govern these variables (e.g., tax reform, labor dispute). Indeed, whenever we obtain reliable information (e.g., from historical or institutional knowledge) that an abrupt local change has taken place in a specific mechanism f_i that constrains a given family $(X_i,\ PA_i)$ of variables, we can use the observed changes in the marginal and conditional probabilities surrounding those variables to determine whether X_i is indeed the child (or dependent variable) of that family, thus determining the direction of causal influences in the domain. The statistical features that remain invariant under such changes, as well as the causal assumptions underlying this invariance, are displayed in the augmented network G'.

3.2.3 Computing the Effect of Interventions

Regardless of whether we represent interventions as a modification of an existing model (Definition 3.2.1) or as a conditionalization in an augmented model (equation (3.9)), the result is a well-defined transformation between the preintervention and postintervention distributions. In the case of an atomic intervention $do(X_i = x_i')$, this transformation can be expressed in a simple *truncated-factorization* formula that follows immediately from (3.2) and Definition 3.2.1:[4]

$$P(x_1, ..., x_n|\hat{x}_i') = \begin{cases} \prod_{j \neq i} P(x_j|pa_j) & \text{if } x_i = x_i', \\ 0 & \text{if } x_i \neq x_i'. \end{cases} \tag{3.10}$$

Equation (3.10) reflects the removal of the term $P(x_i \mid pa_i)$ from the product of (3.5), since pa_i no longer influence X_i. For example, the intervention $do(X = x')$ will transform the pre-intervention distribution given in (3.6) into the product

$$\begin{aligned} P(z_0, z_1, b, z_2, z_3, y|\hat{x}') &= P(z_0)P(z_1|z_0)P(b|z_0) \\ &\quad \times P(z_2|x', z_1)P(z_3|z_2, b)P(y|x', z_2, z_3). \end{aligned}$$

Graphically, the removal of the term $P(x_i|pa_i)$ is equivalent to removing the links between PA_i and X_i while keeping the rest of the network intact. Clearly, the transformation defined in (3.10) satisfies the condition of Definition 1.3.1 as well as the properties of (1.38)–(1.39).

Multiplying and dividing (3.10) by $P(x_i'|pa_i)$, the relationship to the preintervention distribution becomes more transparent:

$$P(x_1, \ldots, x_n|\hat{x}_i') = \begin{cases} \dfrac{P(x_1, \ldots, x_n)}{P(x_i' \mid pa_i)} & \text{if } x_i = x_i', \\ 0 & \text{if } x_i \neq x_i'. \end{cases} \tag{3.11}$$

If we regard a joint distribution as an assignment of mass to a collection of abstract points $(x_1, \ldots, x_n)$, each representing a possible state

[4] Equation (3.10) can also be obtained from the G-computation formula of Robins (1986, p. 1423; see Section 3.6.4) and the manipulation theorem of Spirtes et al. (1993) (according to this source, said formula was "independently conjectured by Fienberg in a seminar in 1991"). Additional properties of the transformation defined in (3.10) and (3.11) are given in Goldszmidt and Pearl (1992) and Pearl (1993b).

of the world, then the transformation described in (3.11) reveals some interesting properties of the change in mass distribution that take place as a result of an intervention $do(X_i = x_i')$ (Goldszmidt and Pearl 1992). Each point $(x_1, \ldots, x_n)$ is seen to increase its mass by a factor equal to the inverse of the conditional probability $P(x_i'|pa_i)$ corresponding to that point. Points for which this conditional probability is low would boost their mass value substantially, while those possessing a pa_i value that anticipates a natural (noninterventional) realization of x_i' (i.e., $P(x_i'|pa_i) \approx 1$) will keep their mass unaltered. In standard Bayes conditionalization, each excluded point $(x_i \neq x_i')$ transfers its mass to the entire set of preserved points through a renormalization constant. However, (3.11) describes a different transformation: each excluded point $(x_i \neq x_i')$ transfers its mass to a select set of points that share the same value of pa_i. This can be seen from the constancy of both the total mass assigned to each stratum pa_i and the relative masses of points within each such stratum:

$$\begin{aligned} P(pa_i|do(x_i')) &= P(pa_i); \\ \frac{P(s_i, pa_i|do(x_i'))}{P(s_i', pa_i|do(x_i'))} &= \frac{P(s_i, pa_i)}{P(s_i', pa_i)}. \end{aligned}$$

Here S_i denotes the set of all variables excluding $\{PA_i \cup X_i\}$. This select set of mass-receiving points can be regarded as "closest" to the point excluded by virtue of sharing the same history, as summarized by pa_i (see Sections 4.1.3 and 7.4.3).

Another interesting form of (3.11) obtains when we interpret the division by $P(x_i'|pa_i)$ as conditionalization on x_i' and pa_i:

$$P(x_1, \ldots, x_n|\hat{x}_i') = \begin{cases} P(x_1, \ldots, x_n|x_i', pa_i)P(pa_i) & \text{if } x_i = x_i', \\ 0 & \text{if } x_i \neq x_i'. \end{cases} \quad (3.12)$$

This formula becomes familiar when used to compute the effect of an intervention $do(X_i = x_i')$ on a set of variables Y disjoint of $(X_i \cup PA_i)$. Summing (3.12) over all variables except $Y \cup X_i$ yields the following theorem.

Theorem 3.2.2 (Adjustment for Direct Causes)
Let PA_i denote the set of direct causes of variable X_i, and let Y be any

set of variables disjoint of $\{X_i \cup PA_i\}$. *The effect of the intervention* $do(X_i = x_i')$ *on* Y *is given by*

$$P(y|\hat{x}_i') = \sum_{pa_i} P(y|x_i', pa_i)P(pa_i), \tag{3.13}$$

where $P(y|x_i', pa_i)$ *and* $P(pa_i)$ *represent preintervention probabilities.*

Equation (3.13) calls for conditioning $P(y|x_i')$ on the parents of X_i and then averaging the result, weighted by the prior probability of $PA_i = pa_i$. The operation defined by this conditioning and averaging is known as "adjusting for PA_i."

Variations of this adjustment have been advanced by many philosophers as probabilistic definitions of causality and causal effect (see Section 7.5). Good (1961), for example, calls for conditioning on "the state of the universe just before" the occurrence of the cause. Suppes (1970) calls for conditioning on the entire past, up to the occurrence of the cause. Skyrms (1980, p. 133) calls for conditioning on "maximally specific specifications of the factors outside of our influence at the time of the decision which are causally relevant to the outcome of our actions ...". The aim of conditioning in these proposals is, of course, to eliminate spurious correlations between the cause (in our case, $X_i = x_i'$) and the effect ($Y = y$); clearly, the set of parents PA_i can accomplish this aim with great economy. In the structural account that we pursue in this book, causal effects are defined in a radically different way. The conditioning operator is not introduced into (3.13) as a remedial "adjustment" aimed at eradicating spurious correlations. Rather, it emerges formally from the deeper principle represented in (3.10)—that of preserving all the invariant information that the preintervention distribution can provide.

The transformation of (3.10) can easily be extended to more elaborate interventions in which several variables are manipulated simultaneously. For example, if we consider the compound intervention $do(S = s)$ where S is a subset of variables, then (echoing (1.37)) we should delete from the product of (3.5) all factors $P(x_i|pa_i)$ corresponding to variables in S and obtain the more general truncated factorization

$$P(x_1, \ldots, x_n|\hat{s}) = \begin{cases} \prod_{i|X_i \notin S} P(x_i|pa_i) & \text{for } x_1, \ldots, x_n \text{ consistent with } s, \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

Likewise, we need not limit ourselves to simple interventions that set variables to constants. Instead, we may consider a more general modification of the causal model whereby some mechanisms are *replaced*. For example, if we replace the mechanism that determines the value of X_i by another equation, one that involves perhaps a new set PA_i^* of variables, then the resultant distribution would obtain by replacing the factor $P(x_i|pa_i)$ with the conditional probability $P^*(x_i|pa_i^*)$ induced by the new equation. The modified joint distribution would then be given by $P^*(x_1, \ldots, x_n) = P(x_1, \ldots, x_n)P^*(x_i|pa_i^*)/P(x_i|pa_i)$.

An Example: Process Control

To illustrate these operations, let us consider an example involving process control; analogous applications in the areas of health management, economic policy making, product marketing, or robot motion planning should follow in a straightforward way. Let the variable Z_k stand for the state of a production process at time t_k, and let X_k stand for a set of variables (at time t_k) that is used to control that process (see Figure 3.3). For example, Z_k could stand for such measurements as

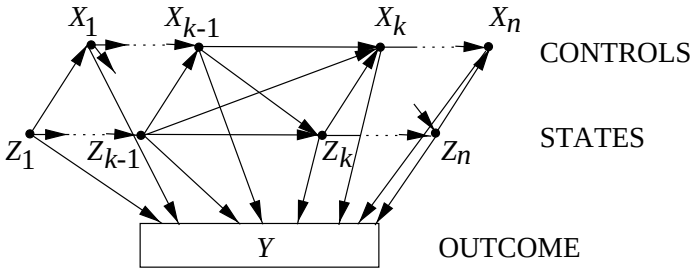

Figure 3.3: Dynamic causal diagram illustrating typical dependencies among the control variables $X_1, \ldots, X_n$, the state variables $Z_1, \ldots, Z_n$, and the outcome variable Y of a sequential process.

temperature and pressure at various location in the plant, and X_k could stand for the rate at which various chemicals are permitted to flow in strategic conduits. Assume that data are gathered while the process is controlled by a strategy S in which each X_k is determined by (i) monitoring three previous variables (X_{k-1}, Z_k, and Z_{k-1}), and (ii) choosing

$X_k = x_k$ with probability $P(x_k|x_{k-1}, z_k, z_{k-1})$. The performance of S is monitored and summarized in the form of a joint probability function $P(y, z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n)$, where Y is an outcome variable (e.g., the quality of the final product). Finally, let us assume (for simplicity) that the state Z_k of the process depends only on the previous state Z_{k-1} and on the previous control X_{k-1}. We wish to evaluate the merit of replacing S with a new strategy, S^*, in which X_k is chosen according to a new conditional probability $P^*(x_k|x_{k-1}, z_k, z_{k-1})$.

Based on our previous analysis (equation (3.14)), the performance $P^*(y)$ of the new strategy S^* will be governed by the distribution

$$\begin{aligned} & P^*(y, z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \qquad (3.15) \\ &= P^*(y|z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \\ &\quad \times \textstyle\prod_k P^*(z_k|z_{k-1}, x_{k-1}) \prod_k P^*(x_k|x_{k-1}, z_k, z_{k-1}). \end{aligned}$$

Because the first two terms remain invariant and the third one is known, we have

$$\begin{aligned} P^*(y) &= \sum_{z_1,\ldots,z_n,x_1,\ldots,x_n} P^*(y, z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \\ &= \sum_{z_1,\ldots,z_n,x_1,\ldots,x_n} P(y|z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \\ &\quad \times \prod_k P(z_k|z_{k-1}, x_{k-1}) \prod_k P^*(x_k|x_{k-1}, z_k, z_{k-1}). \qquad (3.16) \end{aligned}$$

In the special case where S^* is deterministic and time-invariant, X_k becomes a function of X_{k-1}, Z_k, and Z_{k-1}:

$$x_k = g(x_{k-1}, z_k, z_{k-1}).$$

Then the summation over $x_1, \ldots, x_n$ can be performed, yielding

$$\begin{aligned} P^*(y) &= \sum_{z_1,\ldots,z_n} P(y|z_1, z_2, \ldots, z_n, g_1, g_2, \ldots, g_n) \\ &\quad \times \prod_k P(z_k|z_{k-1}, g_{k-1}), \qquad (3.17) \end{aligned}$$

where g_k is defined recursively as

$$g_1 = g(z_1) \text{ and } g_k = g(g_{k-1}, z_k, z_{k-1}).$$

In the special case of a strategy X^* composed of elementary actions $do(X_k = x_k)$, the function g degenerates into a constant, x_k, and we obtain

$$\begin{aligned} P^*(y) &= P(y|\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n) \\ &= \sum_{z_1,.,z_n} P(y|z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n) \prod_k P(z_k|z_{k-1}, x_{k-1}) \end{aligned} \tag{3.18}$$

which can also be obtained from (3.14).

The planning problem illustrated by this example is typical of Markov decision processes (MDPs) (Howard 1960; Dean and Wellman 1991; Bertsekas and Tsitsiklis 1996), where the target of analysis is finding the best next action $do(X_k = x_k)$, given the current state Z_k and past actions. In MDPs, we are normally given the transition functions $P(z_{k+1}|z_k, \hat{x}_k)$ and the cost function to be minimized. In the problem we have just analyzed, neither function is given; instead, they must be learned from data gathered under past (presumably suboptimal) strategies. Fortunately, because all variables in the model were measured, both functions were identifiable and could be estimated directly from the corresponding conditional probabilities as follows:

$$P(z_{k+1}|z_k, \hat{x}_k) = P(z_{k+1}|z_k, x_k);$$
$$P(y|z_1, z_2, \ldots, z_n, \hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n) = P(y|z_1, z_2, \ldots, z_n, x_1, x_2, \ldots, x_n).$$

In Chapter 4 (Section 4.4) we will deal with partially observable Markov decision processes (POMDPs), where some states Z_k are unobserved; learning the transition and cost functions in those problems will require a more intricate method of identification.

It is worth noting that, in this example, to predict the effect of a new strategy it is necessary first to measure variables (Z_k) that are affected by some control variables (X_{k-1}). Such measurements are generally shunned in the classical literature on experimental design (Cox 1958, p. 48) because they lie on the causal pathways between treatment and outcome and thus tend to confound the desired effect estimate. However, our analysis shows that, when properly processed, such measurements may be indispensable in predicting the effect of certain control programs. This will be especially true in semi-Markovian models (i.e.,

DAGs involving unmeasured variables), which are analyzed in Section 3.3.2.

Summary

The immediate implication of the analysis provided in this section is that—given a causal diagram in which all direct causes (i.e. parents) of intervened variables are observable—one can infer postintervention distributions from preintervention distributions; hence, under such assumptions we can estimate the effects of interventions from passive (i.e. nonexperimental) observations, using the truncated factorization of (3.14). Yet the more challenging problem is to derive causal effects in situations like Figure 3.1, where some members of PA_i are unobservable and so prevent estimation of $P(x_i'|pa_i)$. In Sections 3.3 and 3.4 we provide simple graphical tests for deciding when $P(x_j|\hat{x}_i)$ is estimable in such models. But first we need to define more formally what it means for a causal quantity Q to be estimable from passive observations, a question that falls under the technical term *identification*.

3.2.4 Identification of Causal Quantities

Causal quantities, unlike statistical parameters, are defined relative to a causal model M and not relative to a joint distribution $P_M(v)$ over the set V of observed variables. Since nonexperimental data provides information about $P_M(v)$ alone, and since several models can generate the same distribution, the danger exists that the desired quantity will not be discernible unambiguously from the data—even when infinitely many samples are taken. Identifiability ensures that the added assumptions we make about M (e.g., the causal graph or the zero coefficients in structural equations) will supply the missing information without explicating M in full detail.

Definition 3.2.3 (Identifiability)
Let $Q(M)$ be any computable quantity of a model M. We say that Q is identifiable *in a class $\boldsymbol{M}$ of models if, for any pairs of models M_1 and M_2 from $\boldsymbol{M}$, $Q(M_1) = Q(M_2)$ whenever $P_{M_1}(v) = P_{M_2}(v)$. If our observations are limited, and permit only a partial set F_M of features*

(*of* $P_M(v)$) *to be estimated, we define* Q *to be* identifiable from F_M *if* $Q(M_1) = Q(M_2)$ *whenever* $F_{M_1} = F_{M_2}$.

Identifiability is essential for integrating statistical data (summarized by $P(v)$) with incomplete causal knowledge of $\{f_i\}$, as it enables us to estimate quantities Q consistently from large samples of P without specifying the details of M; the general characteristics of the class $\boldsymbol{M}$ suffice. For the purpose of our analysis, the quantity Q of interest is the causal effect $P_M(y|\hat{x})$, which is certainly computable from a given model M (using Definition 3.2.1) but which we often need to compute from an incomplete specification of M—in the form of general characteristics portrayed in the graph G associated with M. We will therefore consider a class $\boldsymbol{M}$ of models that have the following characteristics in common:

(i) they share the same parent-child families (i.e., the same causal graph G); and

(ii) they induce positive distributions on the observed variables (i.e., $P(v) > 0$).

Relative to such classes, we now have the following.

Definition 3.2.4 (Causal-Effect Identifiability)
The causal effect *of* X *on* Y *is said to be* identifiable *from a graph* G *if the quantity* $P(y|\hat{x})$ *can be computed uniquely from any positive probability of the observed variables—that is, if* $P_{M_1}(y|\hat{x}) = P_{M_2}(y|\hat{x})$ *for every pair of models* M_1 *and* M_2 *with* $P_{M_1}(v) = P_{M_2}(v) > 0$ *and* $G(M_1) = G(M_2) = G$.

The identifiability of $P(y|\hat{x})$ ensures that it is possible to infer the effect of action $do(X = x)$ on Y from two sources of information:

(i) passive observations, as summarized by the probability function $P(v)$; and

(ii) the causal graph G, which specifies (qualitatively) which variables make up the stable mechanisms in the domain or, alternatively, which variables participate in the determination of each variable in the domain.

Restricting identifiability to positive distributions assures us that the condition $X = x$ is represented in the data in the appropriate context, thus avoiding a zero denominator in (3.10). It would be impossible to infer the effect of action $do(X = x)$ from data in which X never attains the value x in the context wherein the action is applied. Extensions to some nonpositive distributions are feasible but will not be treated here. Note that, to prove nonidentifiability, it is sufficient to present two sets of structural equations that induce identical distributions over observed variables but have different causal effects.

Using the concept of identifiability, we can now summarize the results of Section 3.2.3 in the following theorem.

Theorem 3.2.5 *Given a causal diagram G of any Markovian model in which a subset V of variables are measured, the causal effect $P(y|\hat{x})$ is identifiable whenever $\{X \cup Y \cup PA_X\} \subseteq V$, that is, whenever X, Y, and all parents of variables in X are measured. The expression of $P(y|\hat{x})$ is then obtained by adjusting for PA_x, as in* (3.13).

A special case of Theorem 3.2.5 holds when *all* variables are assumed to be observed.

Corollary 3.2.6 *Given the causal diagram G of any Markovian model in which all variables are measured, the causal effect $P(y|\hat{x})$ is identifiable for every two subsets of variables X and Y and is obtained from the truncated factorization of* (3.14).

We now turn our attention to identification problems in semi-Markovian models.

3.3 Controlling Confounding Bias

Whenever we undertake to evaluate the effect of one factor (X) on another (Y), the question arises as to whether we should adjust our measurements for possible variations in some other factors (Z), otherwise known as "covariates," "concomitants," or "confounders" (Cox 1958, p. 48). Adjustment amounts to partitioning the population into groups that are homogeneous relative to Z, assessing the effect of X

on Y in each homogeneous group, and then averaging the results (as in (3.13)). The illusive nature of such adjustment was recognized as early as 1899, when Karl Pearson discovered what is now called *Simpson's paradox* (see Section 6.1): Any statistical relationship between two variables may be reversed by including additional factors in the analysis. For example, we may find that students who smoke obtain higher grades than those who do not smoke but, adjusting for age, smokers obtain lower grades in every age group and, further adjusting for family income, smokers again obtain higher grades than nonsmokers in every income-age group, and so on.

Despite a century of analysis, Simpson's reversal continues to "trap the unwary" (Dawid 1979), and the practical question that it poses—whether an adjustment for a given covariate is appropriate—has resisted mathematical treatment. Epidemiologists, for example, are still debating the meaning of "confounding" (Grayson 1987; Shapiro 1997) and often adjust for wrong sets of covariates (Weinberg 1993; see also Chapter 6). The potential-outcome analyses of Rosenbaum and Rubin (1983) and Pratt and Schlaifer (1988) have led to a concept named "ignorability," which recasts the covariate selection problem in counterfactual vocabulary but falls short of providing a workable solution. Ignorability reads: "Z is an admissible set of covariates if, given Z, the value that Y would obtain had X been x is independent of X." Since counterfactuals are not observable, and since judgments about conditional independence of counterfactuals are not readily assertable from ordinary understanding of causal processes, the question has remained open: What criterion should one use to decide which variables are appropriate for adjustment?

Section 3.3.1 presents a general and formal solution of the adjustment problem using the language of causal graphs. In Section 3.3.2 we extend this result to nonstandard covariates that are affected by X and hence require several steps of adjustment. (Finally, Section 3.3.3 illustrates the use of these criteria in an example.

3.3.1 The Back-Door Criterion

Assume we are given a causal diagram G, together with nonexperimental data on a subset V of observed variables in G, and suppose we wish

to estimate what effect the interventions $do(X = x)$ would have on a set of response variables Y, where X and Y are two subsets of V. In other words, we seek to estimate $P(y|\hat{x})$ from a sample estimate of $P(v)$.

We show that there exists a simple graphical test, named the "back-door criterion" in Pearl (1993b), that can be applied directly to the causal diagram in order to test if a set $Z \subseteq V$ of variables is sufficient for identifying $P(y|\hat{x})$.[5]

Definition 3.3.1 (Back-Door)
A set of variables Z satisfies the back-door *criterion relative to an ordered pair of variables (X_i, X_j) in a DAG G if:*

(i) *no node in Z is a descendant of X_i; and*

(ii) *Z blocks every path between X_i and X_j that contains an arrow into X_i.*

Similarly, if X and Y are two disjoint subsets of nodes in G, then Z is said to satisfy the back-door criterion relative to (X, Y) if it satisfies the criterion relative to any pair (X_i, X_j) such that $X_i \in X$ and $X_j \in Y$.

The name "back-door" echoes condition (ii), which requires that only paths with arrows pointing at X_i be blocked; these paths can be viewed as entering X_i through the back door. In Figure 3.4, for example, the sets $Z_1 = \{X_3, X_4\}$ and $Z_2 = \{X_4, X_5\}$ meet the back-door criterion, but $Z_3 = \{X_4\}$ does not because X_4 does not block the path $(X_i, X_3, X_1, X_4, X_2, X_5, X_j)$.

Theorem 3.3.2 (Back-Door Adjustment)
If a set of variables Z satisfies the back-door criterion relative to (X, Y), then the causal effect of X on Y is identifiable and is given by the formula

$$P(y|\hat{x}) = \sum_z P(y|x, z)P(z). \tag{3.19}$$

[5]This criterion may also be obtained from Theorem 7.1 of Spirtes et al. (1993). An alternative criterion, using a single d-separation test, is established in Section 3.4 (see (3.37)).

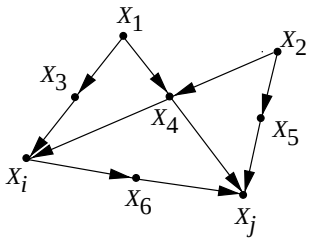

Figure 3.4: A diagram representing the back-door criterion; adjusting for variables $\{X_3, X_4\}$ (or $\{X_4, X_5\}$) yields a consistent estimate of $P(x_j|\hat{x}_i)$.

The summation in (3.19) represents the standard formula obtained under adjustment for Z; variables X for which the equality in (3.19) is valid were named "conditionally ignorable given Z" in Rosenbaum and Rubin (1983). Reducing ignorability conditions to the graphical criterion of Definition 3.3.1 replaces judgments about counterfactual dependencies with judgments about the structure of causal processes, as represented in the diagram. The graphical criterion can be tested by systematic procedures that are applicable to diagrams of any size and shape. The criterion also enables the analyst to search for an optimal set of covariate—namely, a set Z that minimizes measurement cost or sampling variability (Tian et al. 1998). The use of a similar graphical criterion for identifying path coefficients in linear structural equations is demonstrated in Chapter 5. Applications to epidemiological research are given in Greenland et al. (1999a), where the set Z is called "sufficient set" for control of confounding.

Proof of Theorem 3.3.2

The proof originally offered in Pearl (1993b) was based on the observation that, when Z blocks all back-door paths from X to Y, setting $(X = x)$ or conditioning on $X = x$ has the same effect on Y. This can best be seen from the augmented diagram G' of Figure 3.2, to which the intervention arcs $F_X \rightarrow X$ were added. If all back-door paths from X to Y are blocked, then all paths from F_X to Y must go through the children of X, and those would be blocked if we condition on X. The

implication is that Y is independent of F_X given X,

$$P(y|x, F_X = do(x)) = P(y|x, F_X = \text{idle}) = P(y|x), \tag{3.20}$$

which means that the observation $X = x$ cannot be distinguished from the intervention $F_X = do(x)$.

Formally, we can prove this observation by writing $P(y|\hat{x})$ in terms of the augmented probability function P' in accordance with (3.9) and conditioning on Z to obtain

$$\begin{aligned} P(y|\hat{x}) = P'(y|F_x) &= \sum_z P'(y|z, F_x)P'(z|F_x) \\ &= \sum_z P'(y|z, x, F_x)P'(z|F_x). \end{aligned} \tag{3.21}$$

The addition of x to the last expression is licensed by the implication $F_x \Rightarrow X = x$. To eliminate F_x from the two terms on the right-hand side of (3.21), we invoke the two conditions of Definition 3.3.1. Since F_x consists of root nodes with children restricted to X, it must be independent of all nondescendants of X, including Z. Thus, condition (i) yields

$$P'(z|F_x) = P'(z) = P(z).$$

Invoking now the back-door condition (ii), together with (3.20), permits us to eliminate F_x from (3.21), thus proving (3.19). □

3.3.2 The Front-Door Criterion

Condition (i) of Definition 3.3.1 reflects the prevailing practice that "the concomitant observations should be quite unaffected by the treatment" (Cox 1958, p. 48). This section demonstrates how concomitants that *are* affected by the treatment can be used to facilitate causal inference. The emerging criterion, named the front-door criterion in Pearl (1995a), will constitute the second building block of the general test for identifying causal effects (Section 3.4).

Consider the diagram in Figure 3.5, which represents the model of Figure 3.4 when the variables $X_1, \ldots, X_5$ are unobserved and $\{X_i, X_6, X_j\}$ are relabeled $\{X, Z, Y\}$, respectively. Although Z does not satisfy any of the back-door conditions, measurements of Z can

nevertheless enable consistent estimation of $P(y|\hat{x})$. This will be shown by reducing the expression for $P(y|\hat{x})$ to formulas that are computable from the observed distribution function $P(x, y, z)$.

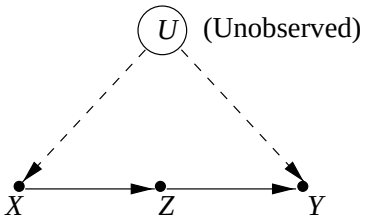

Figure 3.5: A diagram representing the front-door criterion. A two-step adjustment for Z yields a consistent estimate of $P(y|\hat{x})$.

The joint distribution associated with Figure 3.5 can be decomposed (equation (3.5)) into

$$P(x, y, z, u) = P(u)P(x|u)P(z|x)P(y|z, u). \tag{3.22}$$

From (3.10), the intervention $do(x)$ removes the factor $P(x|u)$ and induces the postintervention distribution

$$P(y, z, u|\hat{x}) = P(y|z, u)P(z|x)P(u). \tag{3.23}$$

Summing over z and u then gives

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_u P(y|z, u)P(u). \tag{3.24}$$

In order to eliminate u from the r.h.s. of (3.24), we use the two conditional independence assumptions encoded in the graph of Figure 3.5:

$$P(u|z, x) = P(u|x), \tag{3.25}$$

$$P(y|x, z, u) = P(y|z, u). \tag{3.26}$$

This yields the equalities

$$\begin{aligned} \sum_u P(y|z, u)P(u) &= \sum_x \sum_u P(y|z, u)P(u|x)P(x) \\ &= \sum_x \sum_u P(y|x, z, u)P(u|x, z)P(x) \\ &= \sum_x P(y|x, z)P(x) \end{aligned} \tag{3.27}$$

and allows the reduction of (3.24) to a form involving only observed quantities:

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x'). \tag{3.28}$$

All factors on the r.h.s. of (3.28) are consistently estimable from nonexperimental data, so it follows that $P(y|\hat{x})$ is estimable as well. Thus, we are in possession of an identifiable nonparametric estimand for the causal effect of X on Y whenever we can find a mediating variable Z that meets the conditions of (3.25) and (3.26).

Equation (3.28) can be interpreted as a two-step application of the back-door formula. In the first step, we find the causal effect of X on Z; since there is no back-door path from X to Z, we simply have

$$P(z|\hat{x}) = P(z|x).$$

Next, we compute the causal effect of Z on Y, which we can no longer equate with the conditional probability $P(y|z)$ because there is a back-door path $Z \leftarrow X \leftarrow U \rightarrow Y$ from Z to Y. However, since X blocks (d-separates) this path, X can play the role of a concomitant in the back-door criterion, which allows us to compute the causal effect of Z on Y in accordance with (3.19), giving $P(y|\hat{z}) = \sum_{x'} P(y|x', z)P(x')$. Finally, we combine the two causal effects via

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

which reduces to (3.28).

We summarize this result by a theorem after formally defining the assumptions.

Definition 3.3.3 (Front-Door)
A set of variables Z is said to satisfy the front-door *criterion relative to an ordered pair of variables (X, Y) if:*

(i) *Z intercepts all directed paths from X to Y;*

(ii) *there is no back-door path from X to Z; and*

(iii) *all back-door paths from Z to Y are blocked by X.*

Theorem 3.3.4 (Front-Door Adjustment)
If Z satisfies the front-door criterion relative to (X, Y) and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x'). \tag{3.29}$$

The conditions stated in Definition 3.3.3 are overly restrictive; some of the back-door paths excluded by conditions (ii) and (iii) can actually be allowed provided they are blocked by some concomitants. For example, the variable Z_2 in Figure 3.1 satisfies a front-door-like criterion relative to (X, Z_3) by virtue of Z_1 blocking all back-door paths from X to Z_2 as well as those from Z_2 to Z_3. To allow the analysis of such intricate structures, including nested combinations of back-door and front-door conditions, a more powerful symbolic machinery will be introduced in Section 3.4, one that will sidestep algebraic manipulations such as those used in the derivation of (3.28). But first let us look at an example illustrating possible applications of the front-door condition.

3.3.3 Example: Smoking and the Genotype Theory

Consider the century-old debate on the relation between smoking (X) and lung cancer (Y) (Spirtes et al. 1993, pp. 291–302). According to many, the tobacco industry has managed to forestall antismoking legislation by arguing that the observed correlation between smoking and lung cancer could be explained by some sort of carcinogenic genotype (U) that involves inborn craving for nicotine.

The amount of tar (Z) deposited in a person's lungs is a variable that promises to meet the conditions listed in Definition 3.3.3, thus fitting the structure of Figure 3.5. To meet condition (i), we must assume that smoking cigarettes has no effect on the production of lung cancer except as mediated through tar deposits. To meet conditions (ii) and (iii), we must assume that, even if a genotype is aggravating the production of lung cancer, it nevertheless has no effect on the amount of tar in the lungs except indirectly (through cigarette smoking). Likewise, we must assume that no other factor that affects tar deposit has

	Group Type	$P(x,z)$ Group Size (% of Population)	$P(Y=1\|x,z)$ % of Cancer Cases in Group
$X=0,\ Z=0$	Nonsmokers, No tar	47.5	10
$X=1,\ Z=0$	Smokers, No tar	2.5	90
$X=0,\ Z=1$	Nonsmokers, Tar	2.5	5
$X=1,\ Z=1$	Smokers, Tar	47.5	85

Table 3.1:

any influence on smoking. Finally, condition $P(x,z) > 0$ of Theorem 3.3.4 requires that high levels of tar in the lungs be the result not only of cigarette smoking but also of other factors (e.g., exposure to environmental pollutants) and that tar may be absent in some smokers (owing perhaps to an extremely efficient tar-rejecting mechanism). Satisfaction of this last condition can be tested in the data.

To demonstrate how we can assess the degree to which cigarette smoking increases (or decreases) lung-cancer risk, we will assume a hypothetical study in which the three variables X, Y, Z were measured simultaneously on a large, randomly selected sample of the population. To simplify the exposition, we will further assume that all three variables are binary, taking on true (1) or false (0) values. A hypothetical data set from a study on the relations among tar, cancer, and cigarette smoking is presented in Table 3.1.

It shows that 95% of smokers and 5% of nonsmokers have developed high levels of tar in their lungs. Moreover, 81% of subjects with tar deposits have developed lung cancer, compared to only 14% among those with no tar deposits. Finally, within each of these two groups (tar and no-tar), smokers show a much higher percentage of cancer than nonsmokers.

These results seem to prove that smoking is a major contributor to lung cancer. However, the tobacco industry might argue that the table tells a different story—that smoking actually decreases one's risk

of lung cancer. Their argument goes as follows. If you decide to smoke, then your chances of building up tar deposits are 95%, compared to 5% if you decide not to smoke. In order to evaluate the effect of tar deposits, we look separately at two groups, smokers and nonsmokers. The table shows that tar deposits have a protective effect in both groups: in smokers, tar deposits lower cancer rates from 90% to 85%; in nonsmokers, they lower cancer rates from 10% to 5%. Thus, regardless of whether I have a natural craving for nicotine, I should be seeking the protective effect of tar deposits in my lungs, and smoking offers a very effective means of acquiring those deposits.

To settle the dispute between the two interpretations, we now apply the front-door formula (equation (3.29)) to the data in Table 3.1. We wish to calculate the probability that a randomly selected person will develop cancer under each of the following two actions: smoking (setting $X = 1$) or not smoking (setting $X = 0$).

Substituting the appropriate values of $P(z|x)$, $P(y|x,z)$, and $P(x)$, we have

$$\begin{aligned} P(Y=1|do(X=1)) &= .05(.10\times.50+.90\times.50) \\ &\quad +.95(.05\times.50+.85\times.50) \\ &= .05\times.50+.95\times.45=.4525, \\ P(Y=1|do(X=0)) &= .95(.10\times.50+.90\times.50) \\ &\quad +.05(.05\times.50+.85\times.50) \\ &= .95\times.50+.05\times.45=.4975. \end{aligned} \quad (3.30)$$

Thus, contrary to expectation, the data prove smoking to be somewhat beneficial to one's health.

The data in Table 3.1 are obviously unrealistic and were deliberately crafted so as to support the genotype theory. However, the purpose of this exercise was to demonstrate how reasonable qualitative assumptions about the workings of mechanisms, coupled with nonexperimental data, can produce precise quantitative assessments of causal effects. In reality, we would expect observational studies involving mediating variables to refute the genotype theory by showing, for example, that the mediating consequences of smoking (such as tar deposits) tend to increase, not decrease, the risk of cancer in smokers and nonsmokers

alike. The estimand of (3.29) could then be used for quantifying the causal effect of smoking on cancer.

3.4 A Calculus of Intervention

This section establishes a set of inference rules by which probabilistic sentences involving interventions and observations can be transformed into other such sentences, thus providing a syntactic method of deriving (or verifying) claims about interventions. Each inference rule will respect the interpretation of the $do(\cdot)$ operator as an intervention that modifies a select set of functions in the underlying model. The set of inference rules that emerge from this interpretation will be called *do calculus*.

We will assume that we are given the structure of a causal diagram G in which some of the nodes are observable while others remain unobserved. Our objective will be to facilitate the syntactic derivation of causal effect expressions of the form $P(y|\hat{x})$, where X and Y stand for any subsets of observed variables. By "derivation" we mean stepwise reduction of the expression $P(y|\hat{x})$ to an equivalent expression involving standard probabilities of observed quantities. Whenever such reduction is feasible, the causal effect of X on Y is identifiable (see Definition 3.2.4).

3.4.1 Preliminary Notationnotation

Let X, Y, and Z be arbitrary disjoint sets of nodes in a causal DAG G. We denote by $G_{\overline{X}}$ the graph obtained by deleting from G all arrows pointing to nodes in X. Likewise, we denote by $G_{\underline{X}}$ the graph obtained by deleting from G all arrows emerging from nodes in X. To represent the deletion of both incoming and outgoing arrows, we use the notation $G_{\overline{X}\underline{Z}}$ (see Figure 3.6 for an illustration). Finally, the expression $P(y|\hat{x}, z) \triangleq P(y, z|\hat{x})/P(z|\hat{x})$ stands for the probability of $Y = y$ given that X is held constant at x and that (under this condition) $Z = z$ is observed.

3.4.2 Inference Rules

The following theorem states the three basic inference rules of the proposed calculus. Proofs are provided in Pearl (1995a).

Theorem 3.4.1 (Rules of *do* Calculus)
Let G be the directed acyclic graph associated with a causal model as defined in (3.2)*, and let $P(\cdot)$ stand for the probability distribution induced by that model. For any disjoint subsets of variables X, Y, Z, and W we have the following rules.*

Rule 1 (*Insertion/deletion of observations*) :

$$P(y|\hat{x}, z, w) = P(y|\hat{x}, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}}}. \tag{3.31}$$

Rule 2 (*Action/observation exchange*) :

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, z, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}. \tag{3.32}$$

Rule 3 (*Insertion/deletion of actions*) :

$$P(y|\hat{x}, \hat{z}, w) = P(y|\hat{x}, w) \quad \textit{if} \quad (Y \perp\!\!\!\perp Z|X,\ W)_{G_{\overline{X},\ \overline{Z(W)}}}, \tag{3.33}$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$.

Each of these inference rules follows from the basic interpretation of the "hat" $\hat{x}$ operator as a replacement of the causal mechanism that connects X to its preaction parents by a new mechanism $X = x$ introduced by the intervening force. The result is a submodel characterized by the subgraph $G_{\overline{X}}$ (named "manipulated graph" in Spirtes et al. 1993).

Rule 1 reaffirms d-separation as a valid test for conditional independence in the distribution resulting from the intervention $do(X = x)$, hence the graph $G_{\overline{X}}$. This rule follows from the fact that deleting equations from the system does not introduce any dependencies among the remaining disturbance terms (see (3.2)).

Rule 2 provides a condition for an external intervention $do(Z = z)$ to have the same effect on Y as the passive observation $Z = z$. The

condition amounts to $\{X \cup W\}$ blocking all back-door paths from Z to Y (in $G_{\overline{X}}$), since $G_{\overline{X}\underline{Z}}$ retains all (and only) such paths.

Rule 3 provides conditions for introducing (or deleting) an external intervention $do(Z = z)$ without affecting the probability of $Y = y$. The validity of this rule stems, again, from simulating the intervention $do(Z = z)$ by the deletion of all equations corresponding to the variables in Z (hence the graph $G_{\overline{XZ}}$). The reason for limiting the deletion to nonancestors of W-nodes is provided with the proofs of Rules 1–3 in Pearl (1995a).

Corollary 3.4.2 *A causal effect $q = P(y_1, ..., y_k|\hat{x}_1, ..., \hat{x}_m)$ is identifiable in a model characterized by a graph G if there exists a finite sequence of transformations, each conforming to one of the inference rules in Theorem 3.4.1, that reduces q into a standard (i.e. "hat"-free) probability expression involving observed quantities.*

Whether Rules 1–3 are sufficient for deriving all identifiable causal effects remains an open question. However, the task of finding a sequence of transformations (if such exists) for reducing an arbitrary causal effect expression can be systematized and executed by efficient algorithms (Galles and Pearl 1995; Pearl and Robins 1995), to be discussed in Chapter 4. As we illustrate in Section 3.4.3, symbolic derivations using the hat notation are much more convenient than algebraic derivations that aim at eliminating latent variables from standard probability expressions (as in Section 3.3.2, equation(3.24)).

3.4.3 Symbolic Derivation of Causal Effects: An Example

We will now demonstrate how Rules 1–3 can be used to derive all causal effect estimands in the structure of Figure 3.5. Figure 3.6 displays the subgraphs that will be needed for the derivations that follow.

Task 1: Compute $P(z|\hat{x})$

This task can be accomplished in one step, since G satisfies the applicability condition for Rule 2. That is, $X \perp\!\!\!\perp Z$ in $G_{\underline{X}}$ (because the path $X \leftarrow U \rightarrow Y \leftarrow Z$ is blocked by the converging arrows at Y) and we can write

$$P(z|\hat{x}) = P(z|x). \tag{3.34}$$

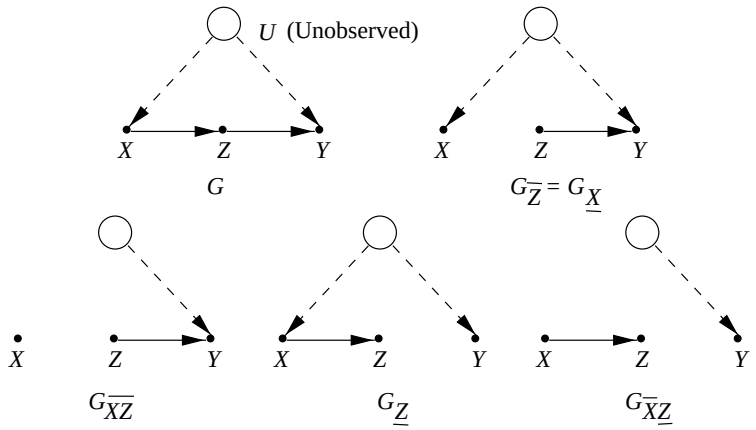

Figure 3.6: Subgraphs of G used in the derivation of causal effects.

Task 2: Compute $P(y|\hat{z})$
Here we cannot apply Rule 2 to exchange $\hat{z}$ with z because $G_{\underline{Z}}$ contains a back-door path from Z to Y: $Z \leftarrow X \leftarrow U \rightarrow Y$. Naturally, we would like to block this path by measuring variables (such as X) that reside on that path. This involves conditioning and summing over all values of X:

$$P(y|\hat{z}) = \sum_x P(y|x,\hat{z})P(x|\hat{z}). \tag{3.35}$$

We now have to deal with two terms involving $\hat{z}$, $P(y|x,\hat{z})$ and $P(x|\hat{z})$. The latter can be readily computed by applying Rule 3 for action deletion:

$$P(x|\hat{z}) = P(x) \text{ if } (Z \perp\!\!\!\perp X)_{G_{\overline{Z}}}, \tag{3.36}$$

since X and Z are d-separated in $G_{\overline{Z}}$. (Intuitively, manipulating Z should have no effect on X, because Z is a descendant of X in G.) To reduce the former term, $P(y|x,\hat{z})$, we consult Rule 2:

$$P(y|x,\hat{z}) = P(y|x,z) \text{ if } (Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}}, \tag{3.37}$$

noting that X d-separates Z from Y in $G_{\underline{Z}}$. This allows us to write (3.35) as

$$P(y|\hat{z}) = \sum_x P(y|x,z)P(x) = E_x P(y|x,z), \tag{3.38}$$

which is a special case of the back-door formula (equation (3.19)). The legitimizing condition, $(Z \perp\!\!\!\perp Y|X)_{G_{\underline{Z}}}$, offers yet another graphical test for a set X to be sufficient for control of confounding (between Y and Z) that is equivalent to the ignorability condition of Rosenbaum and Rubin (1983).

Task 3: Compute $P(y|\hat{x})$

Writing

$$P(y|\hat{x}) = \sum_z P(y|z,\hat{x})P(z|\hat{x}), \tag{3.39}$$

we see that the term $P(z|\hat{x})$ was reduced in (3.34) but that no rule can be applied to eliminate the hat symbol $\hat{}$ from the term $P(y|z,\hat{x})$. However, we can legitimately add this symbol via Rule 2:

$$P(y|z,\hat{x}) = P(y|\hat{z},\hat{x}), \tag{3.40}$$

since the applicability condition $(Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}\underline{Z}}}$ holds (see Figure 3.6). We can now delete the action $\hat{x}$ from $P(y|\hat{z},\hat{x})$ using Rule 3, since $Y \perp\!\!\!\perp X|Z$ holds in $G_{\overline{XZ}}$. Thus, we have

$$P(y|z,\hat{x}) = P(y|\hat{z}), \tag{3.41}$$

which was calculated in (3.38). Substituting (3.38), (3.41), and (3.34) back into (3.39) finally yields

$$P(y|\hat{x}) = \sum_z P(z|x) \sum_{x'} P(y|x',z)P(x'), \tag{3.42}$$

which is identical to the front-door formula of (3.28).

Task 4: Compute $P(y,z|\hat{x})$

We have

$$P(y,z|\hat{x}) = P(y|z,\hat{x})P(z|\hat{x}).$$

The two terms on the r.h.s. were derived before in (3.34) and (3.41), from which we obtain

$$\begin{aligned} P(y,z|\hat{x}) &= P(y|\hat{z})P(z|x) \\ &= P(z|x)\textstyle\sum_{x'} P(y|x',z)P(x'). \end{aligned} \tag{3.43}$$

Task 5: Compute $P(x,y|\hat{z})$

We have

$$\begin{aligned} P(x,y|\hat{z}) &= P(y|x,\hat{z})P(x|\hat{z}) \\ &= P(y|x,z)P(x). \end{aligned} \tag{3.44}$$

The first term on the r.h.s. is obtained by Rule 2 (licensed by $G_{\underline{Z}}$) and the second term by Rule 3 (as in (3.36)).
Note that, in all the derivations, the graph G has provided both the license for applying the inference rules and the guidance for choosing the right rule to apply.

3.4.4 Causal Inference by Surrogate Experiments

Suppose we wish to learn the causal effect of X on Y when $P(y|\hat{x})$ is not identifiable and, for practical reasons of cost or ethics, we cannot control X by randomized experiment. The question arises of whether $P(y|\hat{x})$ can be identified by randomizing a surrogate variable Z that is easier to control than X. For example, if we are interested in assessing the effect of cholesterol levels (X) on heart disease (Y), a reasonable experiment to conduct would be to control subjects' diet (Z), rather than exercising direct control over cholesterol levels in subjects' blood.

Formally, this problem amounts to transforming $P(y|\hat{x})$ into expressions in which only members of Z obtain the hat symbol. Using Theorem 3.4.1, it can be shown that the following conditions are sufficient for admitting a surrogate variable Z:

(i) X intercepts all directed paths from Z to Y; and

(ii) $P(y|\hat{x})$ is identifiable in $G_{\overline{Z}}$.

Indeed, if condition (i) holds then we can write $P(y|\hat{x}) = P(y|\hat{x},\ \hat{z})$, because $(Y \perp\!\!\!\perp Z|X)_{G_{\overline{XZ}}}$. But $P(y|\hat{x},\ \hat{z})$ stands for the causal effect of X on Y in a model governed by $G_{\overline{Z}}$, which—by condition (ii), is identifiable. Translated to our cholesterol example, these condition require that there be no direct effect of diet on heart conditions and no confounding of cholesterol levels and heart disease, unless we can neutralize such confounding by additional measurements.

Figures 3.9(e) and 3.9(h) (in Section 3.5.2) illustrate models in which both conditions hold. With Figure 3.9(e), for example, we obtain this estimand

$$P(y|\hat{x}) = P(y|x,\ \hat{z}) = \frac{P(y,\ x|\hat{z})}{P(x|\hat{z})}. \tag{3.45}$$

This can be established directly by first applying Rule 3 to add $\hat{z}$,

$$P(y|\hat{x}) = P(y|\hat{x},\ \hat{z}) \text{ because } (Y \perp\!\!\!\perp Z|X)_{G_{\overline{X}\overline{Z}}},$$

and then applying Rule 2 to exchange $\hat{x}$ with x:

$$P(y|\hat{x},\ \hat{z}) = P(y|x,\ \hat{z}) \text{ because } (Y \perp\!\!\!\perp X|Z)_{G_{\underline{X}\overline{Z}}}.$$

According to (3.45), only one level of Z suffices for the identification of $P(y|\hat{x})$ for any values of y and x. In other words, Z need not be varied at all; it can simply be held constant by external means and, if the assumptions embodied in G are valid, the r.h.s. of (3.45) should attain the same value regardless of the (constant) level at which Z is being held. In practice, however, several levels of Z will be needed to ensure that enough samples are obtained for each desired value of X. For example, if we are interested in the difference $E(Y|\hat{x}) - E(Y|\hat{x}')$, where x and x' are two treatment levels, then we should choose two values z and z' of Z that maximize the number of samples in x and x' (respectively) and then estimate

$$E(Y|\hat{x}) - E(Y|\hat{x}') = E(Y|x,\hat{z}) - E(Y|x',\hat{z}').$$

3.5 Graphical Tests of Identifiability

Figure 3.7 shows simple diagrams in which $P(y|\hat{x})$ cannot be identified owing to the presence of a "bow" pattern—a confounding arc (dashed) embracing a causal link between X and Y. A confounding arc represents the existence in the diagram of a back-door path that contains only unobserved variables and has no converging arrows. For example, the path $X,\ Z_0,\ B,\ Z_3$ in Figure 3.1 can be represented as a confounding arc between X and Z_3. A bow pattern represents an equation $y = f_Y(x, u, \epsilon_Y)$, where U is unobserved and dependent on X. Such an equation does not permit the identification of causal effects, since any portion of the observed dependence between X and Y may always be attributed to spurious dependencies mediated by U.

The presence of a bow pattern prevents the identification of $P(y|\hat{x})$ even when it is found in the context of a larger graph, as in Figure

3.7(b). This is in contrast to linear models, where the addition of an arc to a bow pattern can render $P(y|\hat{x})$ identifiable (see Chapter 5, Figure 5.9). For example, if Y is related to X via a linear relation $y = bx + u$, where U is an unobserved disturbance possibly correlated with X, then $b = \frac{\partial}{\partial x}E(Y|\hat{x})$ is not identifiable. However, adding an arc $Z \rightarrow X$ to the structure (i.e., finding a variable Z that is correlated with X but not with U) would facilitate the computation of $E(Y|\hat{x})$ via the instrumental variable formula (Bowden and Turkington 1984; see also Chapter 5):

$$b \triangleq \frac{\partial}{\partial x}E(Y|\hat{x}) = \frac{E(Y|z)}{E(X|z)} = \frac{r_{YZ}}{r_{XZ}}. \tag{3.46}$$

In nonparametric models, adding an instrumental variable Z to a bow pattern (Figure 3.7(b)) does not permit the identification of $P(y|\hat{x})$. This is a familiar problem in the analysis of clinical trials in which treatment assignment (Z) is randomized (hence, no link enters Z) but compliance is imperfect (see Chapter 8). The confounding arc between X and Y in Figure 3.7(b) represents unmeasurable factors that influence subjects' choice of treatment (X) as well as subjects' response to treatment (Y). In such trials, it is not possible to obtain an unbiased estimate of the treatment effect $P(y|\hat{x})$ without making additional assumptions on the nature of the interactions between compliance and response (as is done, for example, in the potential-outcome approach to instrumental variables developed in Imbens and Angrist 1994 and Angrist et al. 1996. Although the added arc $Z \rightarrow X$ permits us to calculate bounds on $P(y|\hat{x})$ (Robins 1989; sec. 1g; Manski 1990; Balke and Pearl 1997) and the upper and lower bounds may even coincide for certain types of distributions $P(x, y, z)$ (Section 8.2.4), there is no way of computing $P(y|\hat{x})$ for *every* positive distribution $P(x, y, z)$, as required by Definition 3.2.4.

In general, the addition of arcs to a causal diagram can impede, but never assist, the identification of causal effects in nonparametric models. This is because such addition reduces the set of d-separation conditions carried by the diagram; hence, if a causal effect derivation fails in the original diagram, it is bound to fail in the augmented diagram as well. Conversely, any causal effect derivation that succeeds in the augmented

diagram (by a sequence of symbolic transformations, as in Corollary 3.4.2) would succeed in the original diagram.

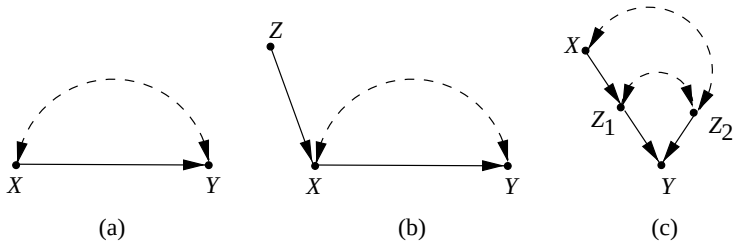

Figure 3.7: (a) A bow pattern: a confounding arc embracing a causal link $X \rightarrow Y$, thus preventing the identification of $P(y|\hat{x})$ even in the presence of an instrumental variable Z, as in (b). (c) A bowless graph that still prohibits the identification of $P(y|\hat{x})$.

Our ability to compute $P(y_1|\hat{x})$ and $P(y_2|\hat{x})$ for pairs (Y_1, Y_2) of singleton variables does not ensure our ability to compute joint distributions, such as $P(y_1, y_2|\hat{x})$. Figure 3.7(c), for example, shows a causal diagram where both $P(z_1|\hat{x})$ and $P(z_2|\hat{x})$ are computable yet $P(z_1, z_2|\hat{x})$ is not. Consequently, we cannot compute $P(y|\hat{x})$. It is interesting to note that this diagram is the smallest graph that does not contain a bow pattern and still presents an uncomputable causal effect.

Another interesting feature demonstrated by Figure 3.7(c) is that computing the effect of a joint intervention is often easier than computing the effects of its constituent singleton interventions.[6] Here, it is possible to compute $P(y|\hat{x}, \hat{z_2})$ and $P(y|\hat{x}, \hat{z_1})$, yet there is no way of computing $P(y|\hat{x})$. For example, the former can be evaluated by invoking Rule 2 in $G_{\overline{X}\underline{Z_2}}$, giving

$$\begin{aligned} P(y|\hat{x}, \hat{z_2}) &= \sum_{z_1} P(y|z_1, \hat{x}, \hat{z_2}) P(z_1|\hat{x}, \hat{z_2}) \\ &= \sum_{z_1} P(y|z_1, x, z_2) P(z_1|x). \end{aligned} \tag{3.47}$$

However, Rule 2 cannot be used to convert $P(z_1|\hat{x}, z_2)$ into $P(z_1|x, z_2)$ because, when conditioned on Z_2, X and Z_1 are d-connected in $G_{\underline{X}}$

[6] This was brought to my attention by James Robins, who has worked out many of these computations in the context of sequential treatment management (Robins 1986, p. 1423).

(through the dashed lines). A general approach to computing the effect of joint interventions is developed in Pearl and Robins (1995); this is described in Chapter 4 (Section 4.4).

3.5.1 Identifying Models

Figure 3.8 shows simple diagrams in which the causal effect of X on Y is identifiable (where X and Y are single variables). Such models are called "identifying" because their structures communicate a sufficient number of assumptions (missing links) to permit the identification of the target quantity $P(y|\hat{x})$. Latent variables are not shown explicitly in these diagrams; rather, such variables are implicit in the confounding arcs (dashed). Every causal diagram with latent variables can be converted to an equivalent diagram involving measured variables interconnected by arrows and confounding arcs. This conversion corresponds to substituting out all latent variables from the structural equations of (3.2) and then constructing a new diagram by connecting any two variables X_i and X_j by (i) an arrow from X_j to X_i whenever X_j appears in the equation for X_i and (ii) a confounding arc whenever the same ϵ term appears in both f_i and f_j. The result is a diagram in which all unmeasured variables are exogenous and mutually independent.

Several features should be noted from examining the diagrams in Figure 3.8.

1. Since the removal of any arc or arrow from a causal diagram can only assist the identifiability of causal effects, $P(y|\hat{x})$ will still be identified in any edge subgraph of the diagrams shown in Figure 3.8. Likewise, the introduction of mediating observed variables onto any edge in a causal graph can assist, but never impede, the identifiability of any causal effect. Therefore, $P(y|\hat{x})$ will still be identified from any graph obtained by adding mediating nodes to the diagrams shown in Figure 3.8.

2. The diagrams in Figure 3.8 are maximal in the sense that the introduction of any additional arc or arrow onto an existing pair of nodes would render $P(y|\hat{x})$ no longer identifiable.

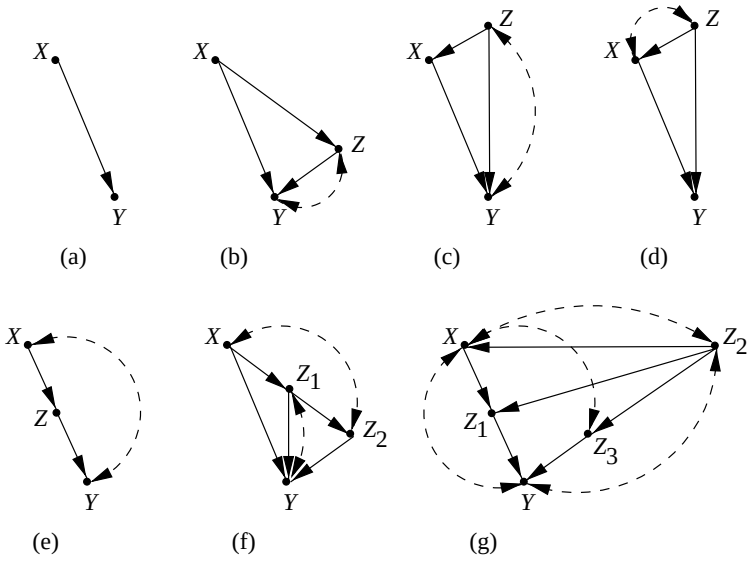

Figure 3.8: Typical models in which the effect of X on Y is identifiable. Dashed arcs represent confounding paths, and Z represents observed covariates.

3. Although most of the diagrams in Figure 3.8 contain bow patterns, none of these patterns emanates from X (as is the case in Figures 3.9(a) and (b) to follow). In general, a necessary condition for the identifiability of $P(y|\hat{x})$ is the absence of a confounding arc between X and any child of X that is an ancestor of Y.

4. Diagrams (a) and (b) in Figure 3.8 contain no back-door paths between X and Y and thus represent experimental designs in which there is no confounding bias between the treatment (X) and the response (Y); hence, $P(y|\hat{x}) = P(y|x)$. Likewise, diagrams (c) and (d) in Figure 3.8 represent designs in which observed covariates Z block every back-door path between X and Y (i.e., X is "conditionally ignorable" given Z, in the language of Rosenbaum and Rubin 1983); hence, $P(y|\hat{x})$ is obtained by standard adjustment for Z (as in (3.19)):

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

5. For each of the diagrams in Figure 3.8, we readily obtain a formula for $P(y|\hat{x})$ by using symbolic derivations patterned after those in Section 3.4.3. The derivation is often guided by the graph topology. For example, diagram (f) in Figure 3.8 dictates the following derivation. Writing

$$P(y|\hat{x}) = \sum_{z_1,z_2} P(y|z_1, z_2, \hat{x})P(z_1, z_2|\hat{x}),$$

we see that the subgraph containing $\{X, Z_1, Z_2\}$ is identical in structure to that of diagram (e), with (Z_1, Z_2) replacing (Z, Y), respectively. Thus, $P(z_1, z_2|\hat{x})$ can be obtained from (3.43). Likewise, the term $P(y|z_1, z_2, \hat{x})$ can be reduced to $P(y|z_1, z_2, x)$ by Rule 2, since $(Y \perp\!\!\!\perp X|Z_1, Z_2)_{G_{\underline{X}}}$. We therefore have

$$\begin{aligned} P(y|\hat{x}) &= \sum_{z_1,z_2} P(y|z_1, z_2, x)P(z_1|x) \\ &\quad \times \sum_{x'} P(z_2|z_1, x')P(x'). \end{aligned} \tag{3.48}$$

Applying a similar derivation to diagram (g) of Figure 3.8 yields

$$\begin{aligned} P(y|\hat{x}) &= \sum_{z_1}\sum_{z_2}\sum_{x'} P(y|z_1,\ z_2,\ x')P(x'|z_2) \\ &\quad \times P(z_1|z_2,\ x)P(z_2) \end{aligned} \tag{3.49}$$

Note that the variable Z_3 does not appear in (3.48), which means that Z_3 need not be measured if all one wants to learn is the causal effect of X on Y.

6. In diagrams (e), (f), and (g) of Figure 3.8, the identifiability of $P(y|\hat{x})$ is rendered feasible through observed covariates Z that are affected by the treatment X (since members of Z are descendants of X). This stands contrary to the warning—repeated in most of the literature on statistical experimentation—to refrain from adjusting for concomitant observations that are affected by the treatment (Cox 1958; Rosenbaum 1984; Pratt and Schlaifer 1988; Wainer 1989). It is commonly believed that a concomitant Z that is affected by the treatment must be excluded from the

analysis of the total effect of the treatment (Pratt and Schlaifer 1988). The reason given for the exclusion is that the calculation of total effects amounts to integrating out Z, which is functionally equivalent to omitting Z to begin with. Diagrams (e), (f), and (g) show cases where the total effects of X are indeed the target of investigation and, even so, the measurement of concomitants that are affected by X (e.g., Z or Z_1) is still necessary. However, the adjustment needed for such concomitants is nonstandard, involving two or more stages of the standard adjustment of (3.19) (see (3.28), (3.48), and (3.49)).

7. In diagrams (b), (c), and (f) of Figure 3.8, Y has a parent whose effect on Y is not identifiable; even so, the effect of X on Y is identifiable. This demonstrates that local identifiability is not a necessary condition for global identifiability. In other words, to identify the effect of X on Y we need not insist on identifying each and every link along the paths from X to Y.

3.5.2 Nonidentifying Models

Figure 3.9 presents typical diagrams in which the total effect of X on Y, $P(y|\hat{x})$, is not identifiable. Noteworthy features of these diagrams are as follows.

1. All graphs in Figure 3.9 contain unblockable back-door paths between X and Y, that is, paths ending with arrows pointing to X that cannot be blocked by observed nondescendants of X. The presence of such a path in a graph is, indeed, a necessary test for nonidentifiability (see Theorem 3.3.2). That is not a sufficient test is demonstrated by Figure 3.8(e), in which the back-door path (dashed) is unblockable and yet $P(y|\hat{x})$ is identifiable.

2. A sufficient condition for the nonidentifiability of $P(y|\hat{x})$ is the existence of a confounding path between X and any of its children on a path from X to Y, as shown in Figures 3.9(b) and (c). A stronger sufficient condition is that the graph contain any of the patterns shown in Figure 3.9 as an edge subgraph.

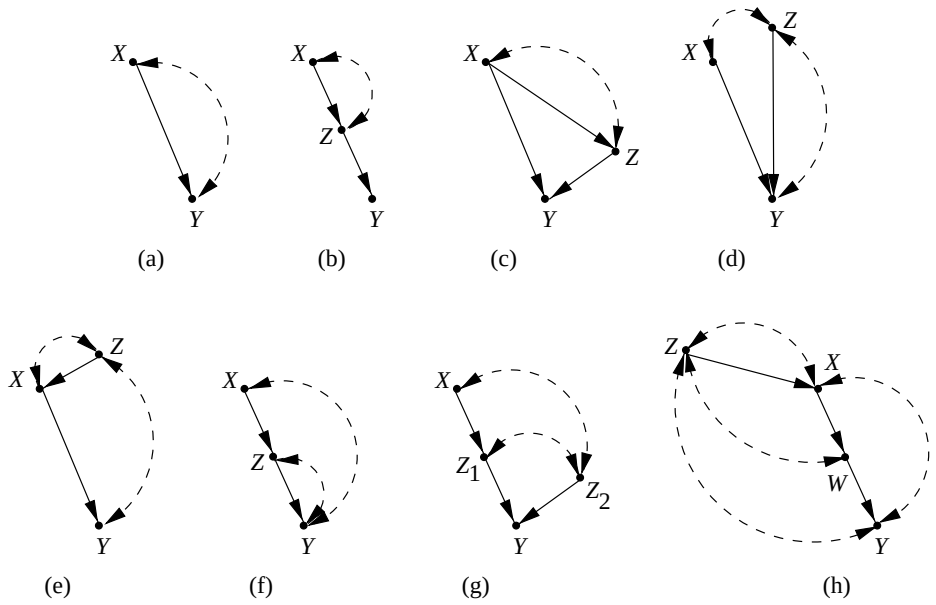

Figure 3.9: Typical models in which $P(y|\hat{x})$ is not identifiable.

3. Graph (g) in Figure 3.9 (same as Figure 3.7(c)) demonstrates that local identifiability is not sufficient for global identifiability. For example, we can identify $P(z_1|\hat{x})$, $P(z_2|\hat{x})$, $P(y|\hat{z}_1)$, and $P(y|\hat{z}_2)$ but not $P(y|\hat{x})$. This is one of the main differences between nonparametric and linear models; in the latter, all causal effects can be determined from the structural coefficients and each coefficient represents the causal effect of one variable on its immediate successor.

3.6 Discussion

3.6.1 Qualifications and Extensions

The methods developed in this chapter facilitate the drawing of quantitative causal inferences from a combination of qualitative causal assumptions (encoded in the diagram) and nonexperimental observations. The causal assumptions in themselves cannot generally be tested in nonexperimental studies, unless they impose constraints on the ob-

served distributions. The most common type of constraints appears in the form of conditional independencies, as communicated through the d-separation conditions in the diagrams. Another type of constraints takes the form of numerical inequalities. In Chapter 8, for example, we show that the assumptions associated with instrumental variables (Figure 3.7(b)) are subject to falsification tests in the form of inequalities on conditional probabilities (Pearl 1995b). Still, such constraints permit the testing of merely a small fraction of the causal assumptions embodied in the diagrams; the bulk of those assumptions must be substantiated from domain knowledge as obtained from either theoretical considerations (e.g., that falling barometers do not cause rain) or related experimental studies. For example, the experimental study of Moertel et al. (1985), which refuted the hypothesis that vitamin C is effective against cancer, can be used as a substantive assumption in observational studies involving vitamin C and cancer patients; it would be represented as a missing link (between vitamin C and cancer) in the associated diagram. In summary, the primary use of the methods described in this chapter lies not in testing causal assumptions but in providing an effective language for making those assumptions precise and explicit. Assumptions can thereby be isolated for deliberation or experimentation and then (once validated) be integrated with statistical data to yield quantitative estimates of causal effects.

An important issue that will be considered only briefly in this book (see Section 8.5) is sampling variability. The mathematical derivation of causal effect estimands should be considered a first step toward supplementing these estimands with confidence intervals and significance levels, as in traditional analysis of controlled experiments. We should remark, though, that having obtained nonparametric estimands for causal effects does not imply that one should refrain from using parametric forms in the estimation phase of the study. For example, if the assumptions of Gaussian, zero-mean disturbances and additive interactions are deemed reasonable, then the estimand given in (3.28) can be converted to the product $E(Y|\hat{x}) = r_{ZX} r_{YZ \cdot X} x$, where $r_{YZ \cdot X}$ is the standardized regression coefficient (Section 5.3.1); the estimation problem then reduces to that of estimating regression coefficients (e.g., by least squares). More sophisticated estimation techniques can be found in Rosenbaum and Rubin (1983), Robins (1989, sec. 17), and Robins

et al. (1992, pp. 331–3). For example, the "propensity score" method of Rosenbaum and Rubin (1983) was found to be quite useful when the dimensionality of the adjusted covariates is high. In a more recent scheme called "marginal models," Robins (1999) shows that, rather than estimating individual factors in the adjustment formula of (3.19), it is often more advantageous to use $P(y|\hat{x}) = \sum_z \frac{P(x,y,z)}{P(x|z)}$, where the preintervention distribution remains unfactorized. One can then separately estimate the denominator $P(x|z)$, weigh individual samples by the inverse of this estimate, and treat the weighted samples as if they were drawn at random from the postintervention distribution $P(y|\hat{x})$. Postintervention parameters, such as $\frac{\partial}{\partial x}E(Y|\hat{x})$, can then be estimated by ordinary least squares. This method is especially advantageous in longitudinal studies with time-varying covariates, as in the process control problem discussed in Section 3.2.3 (see (3.18)).

Several extensions of the methods proposed in this chapter are noteworthy. First, the identification analysis for atomic interventions can be generalized to complex policies in which a set X of controlled variables is made to respond in a specified way to some set Z of covariates via functional or stochastic strategies, as in Section 3.2.3. In Chapter 4 (Section 4.2), it is shown that identifying the effect of such policies is equivalent to computing the expression $P(y|\hat{x}, z)$.

A second extension concerns the use of the intervention calculus (Theorem 3.4.1) in nonrecursive models, that is, in causal diagrams involving directed cycles or feedback loops. The basic definition of causal effects in term of "wiping out" equations from the model (Definition 3.2.1) still carries over to nonrecursive systems (Strotz and Wold 1960; Sobel 1990), but then two issues must be addressed. First, the analysis of identification must ensure the stability of the remaining submodels (Fisher 1970). Second, the d-separation criterion for DAGs must be extended to cover cyclic graphs as well. The validity of d-separation has been established for nonrecursive linear models (Spirtes 1995) as well as for nonlinear systems involving discrete variables (Pearl and Dechter 1996). However, the computation of causal effect estimands will be harder in cyclic nonlinear systems, because symbolic reduction of $P(y|\hat{x})$ to hat-free expressions may require the solution of nonlinear equations. In Chapter 7 (Section 7.2.1) we demonstrate the evaluation

of policies and counterfactuals in nonrecursive linear systems (see also Balke and Pearl (1995)).

A third extension concerns generalizations of intervention calculus (Theorem 3.4.1) to situations where the data available is not obtained under i.i.d. (independent and identically distributed) sampling. One can imagine, for instance, a physician who prescribes a certain treatment to patients only when the fraction of survivors among previous patients drops below some threshold. In such cases, it is required to estimate the causal effect $P(y|\hat{x})$ from non-independent samples. Vladimir Vovk (1996) gave conditions under which the rules of Theorem 3.4.1 will be applicable when sampling is not i.i.d., and he went on to cast the three inference rules as a logical production system.

3.6.2 Diagrams as a Mathematical Language

The benefit of incorporating substantive background knowledge into probabilistic inference was recognized as far back as Thomas Bayes (1763) and Pierre Laplace (1814), and its crucial role in the analysis and interpretation of complex statistical studies is generally acknowledged by most modern statisticians. However, the mathematical language available for expressing background knowledge has remained in a rather pitiful state of development.

Traditionally, statisticians have approved of only one way of combining substantive knowledge with statistical data: the Bayesian method of assigning subjective priors to distributional parameters. To incorporate causal information within this framework, plain causal statements such as "Y is not affected by X" must be converted into sentences or events capable of receiving probability values (e.g. counterfactuals). For instance, to communicate the innocent assumption that mud does not cause rain, we would have to use a rather unnatural expression and say that the probability of the counterfactual event "rain if it were not muddy" is the same as the probability of "rain if it were muddy." Indeed, this is how the potential-outcome approach of Neyman and Rubin has achieved statistical legitimacy: causal judgments are expressed as constraints on probability functions involving counterfactual variables (see Section 3.6.3).

Causal diagrams offer an alternative language for combining data

with causal information. This language simplifies the Bayesian route by accepting plain causal statements as its basic primitives. Such statements, which merely indicate whether a causal connection between two variables of interest exists, are commonly used in ordinary discourse and provide a natural way for scientists to communicate experience and organize knowledge.[7] It can be anticipated, therefore, that the language of causal graphs will find applications in problems requiring substantial domain knowledge.

The language is not new. The use of diagrams and structural equations models to convey causal information has been quite popular in the social sciences and econometrics. Statisticians, however, have generally found these models suspect, perhaps because social scientists and econometricians have failed to provide an unambiguous definition of the empirical content of their models—that is, to specify the experimental conditions, however hypothetical, whose outcomes would be constrained by a given structural equation. (Chapter 5 discusses the bizarre history of structural equations in the social sciences and economics). As a result, even such basic notions as "structural coefficients" or "missing links" become the object of serious controversy (Freedman 1987; Goldberger 1992) and misinterpretations (Whittaker 1990, p. 302; Wermuth 1992; Cox and Wermuth 1993).

To a large extent, this history of controversy and miscommunication stems from the absence of an adequate mathematical notation for defining basic notions of causal modeling. For example, standard probabilistic notation cannot express the empirical content of the coefficient b in the structural equation $y = bx + \epsilon_Y$, even if one is prepared to assume that ϵ_Y (an unobserved quantity) is uncorrelated with X.[8] Nor can any probabilistic meaning be attached to the analyst's excluding

[7] Remarkably, many readers of this chapter (including two referees of this book) classified the methods presented here as belonging to the "Bayesian camp" and as depending on a "good prior." This classification is misleading. The method does depend on subjective assumptions (e.g., mud does not cause rain), but such assumptions are causal, not statistical, and cannot be expressed as prior probabilities on parameters of joint distributions.

[8] Voluminous literature on the subject of "exogeneity" (e.g. Richard 1980; Engle et al. 1983; Hendry 1995) has emerged from economists' struggle to give statistical interpretation to the causal assertion "X and ϵ_Y are uncorrelated" (Aldrich 1993; see Section 5.4.3).

from the equation variables that are highly correlated with X or Y but do not "directly affect" Y.[9]

The notation developed in this chapter gives these (causal) notions a clear empirical interpretation, because it permits one to specify precisely what is being held constant and what is merely measured in a given experiment. (The need for this distinction was recognized by many researchers, most notably Pratt and Schlaifer 1988 and Cox 1992). The meaning of b is simply $\frac{\partial}{\partial x}E(Y|\hat{x})$, that is, the rate of change (in x) of the expectation of Y in an experiment where X is held at x by external control. This interpretation holds regardless of whether ϵ_Y and X are correlated (e.g., via another equation $x = ay + \epsilon_X$). Likewise, the analyst's decision as to which variables should be included in a given equation can be based on a hypothetical controlled experiment: A variable Z is excluded from the equation for Y if (for every level of ϵ_Y) Z has no influence on Y when all other variables (S_{YZ}), are held constant; this implies $P(y|\hat{z}, \hat{s}_{YZ}) = P(y|\hat{s}_{YZ})$. Specifically, variables that are excluded from the equation $y = bx + \epsilon_Y$ are not conditionally independent of Y given measurements of X but instead are *causally* irrelevant to Y given settings of X. The operational meaning of the "disturbance term" ϵ_Y is likewise demystified: ϵ_Y is defined as the difference $Y - E(Y|\hat{s}_Y)$. Two disturbance terms, ϵ_X and ϵ_Y, are correlated if $P(y|\hat{x}, \hat{s}_{XY}) \neq P(y|x, \hat{s}_{XY})$, and so on (see Chapter 5, Section 5.4 for further elaboration).

The distinctions provided by the hat notation clarify the empirical basis of structural equations and should make causal models more acceptable to empirical researchers. Moreover, since most scientific knowledge is organized around the operation of "holding X fixed" rather than "conditioning on X," the notation and calculus developed in this chapter should provide an effective means for scientists to communicate substantive information and to infer its logical consequences.

[9]The bitter controversy between Goldberger (1992) and Wermuth (1992) revolves around Wermuth's insistence on giving a statistical interpretation to the zero coefficients in structural equations (see Section 5.4.1).

3.6.3 Translation from Graphs to Potential Outcomes

This chapter uses two representations of causal information: graphs and structural equations, where the former is an abstraction of the latter. Both representations have been controversial for almost a century. On the one hand, economists and social scientists have embraced these modeling tools, but they continue to question and debate the causal content of the parameters they estimate (see Sections 5.1 and 5.4 for details); as a result, the use of structural models in policy-making contexts is often viewed with suspicion. Statisticians, on the other hand, reject both representations as problematic (Freedman 1987) if not meaningless (Wermuth 1992; Holland 1995), and they sometimes resort to the Neyman-Rubin potential-outcome notation when pressed to communicate causal information (Rubin 1990).[10] A detailed formal analysis of the relationships between the structural and potential-outcome approaches is offered in Chapter 7 (Section 7.4.4) and proves their mathematical equivalence. In this section we highlight commonalities and differences between the two approaches as they pertain to the elicitation of causal assumptions.

The primitive object of analysis in the potential-outcome framework is the unit-based response variable, denoted $Y(x, u)$ or $Y_x(u)$, read: "the value that Y would obtain in unit u, had X been x." This counterfactual entity has natural interpretation in structural equations models. Consider a general structural model M that contains a set of equations

$$x_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n, \tag{3.50}$$

as in (3.4). Let U stand for the vector $(U_1, \ldots, U_n)$ of background variables, let X and Y be two disjoint subsets of observed variables, and let M_x be the submodel created by replacing the equations corresponding to variables in X with $X = x$, as in Definition 3.2.1. The structural interpretation of $Y(x, u)$ is given by

$$Y(x, u) \triangleq Y_{M_x}(u). \tag{3.51}$$

[10] A parallel framework was developed in the econometrics literature under the rubric "switching regression" (Manski 1995, p. 38), which Heckman (1996) attributed to Roy (1951) and Quandt (1958).

That is, $Y(x, u)$ is the (unique) solution of Y under the realization $U = u$ in the submodel M_x of M. Although the term *unit* in the potential-outcome literature normally stands for the identity of a specific individual in a population, a unit may also be thought of as the set of attributes that characterize that individual, the experimental conditions under study, the time of day, and so on—all of which are represented as components of the vector u in structural modeling. In fact, the only requirements on U are (i) that it represent as many background factors as needed to render the relations among endogenous variables deterministic and (ii) that the data consist of independent samples drawn from $P(u)$. The identity of an individual person in an experiment is often sufficient for this purpose because it represents the anatomical and genetic makings of that individual, which are often sufficient for determining that individual's response to treatments or other programs of interest.

Equation (3.51) forms a connection between the opaque English phrase "the value that Y would obtain in unit u, had X been x" and the physical processes that transfer changes in X into changes in Y. The formation of the submodel M_x explicates precisely how the hypothetical phrase "had X been x" could be realized, as well as what process must give in to make $X = x$ a reality.

Given this interpretation of $Y(x, u)$, it is instructive to contrast the methodologies of causal inference in the counterfactual versus structural frameworks. If U is treated as a random variable then the value of the counterfactual $Y(x, u)$ becomes a random variable as well, denoted as $Y(x)$ or Y_x. The potential-outcome analysis proceeds by imagining the observed distribution $P(x_1, \ldots, x_n)$ as the marginal distribution of an augmented probability function P^* defined over both observed and counterfactual variables. Queries about causal effects (written $P(y|\hat{x})$ in our structural analysis) are phrased as queries about the marginal distribution of the counterfactual variable of interest, written $P^*(Y(x) = y)$. The new hypothetical entities $Y(x)$ are treated as ordinary random variables; for example, they are assumed to obey the axioms of probability calculus, the laws of conditioning, and the axioms of conditional independence. Moreover, these hypothetical entities are assumed to be connected to observed variables via consistency

constraints (Robins 1986) such as[11]

$$X = x \implies Y(x) = Y, \tag{3.52}$$

which states that, for every u, if the actual value of X turns out to be x, then the value that Y would take on if X were x is equal to the actual value of Y. Thus, whereas the structural approach views the intervention $do(x)$ as an operation that changes the model (and the distribution) but keeps all variables the same, the potential-outcome approach views the variable Y under $do(x)$ to be a different variable, $Y(x)$, loosely connected to Y through relations such as (3.52). In Chapter 7 we show, using the structural interpretation of $Y(x, u)$, that it is indeed legitimate to treat counterfactuals as random variables in all respects and, moreover, that consistency constraints like (3.52) follow as theorems from the structural interpretation.

To communicate substantive causal knowledge, the potential-outcome analyst must express causal assumptions as constraints on P^*, usually in the form of conditional independence assertions involving counterfactual variables. For example, to communicate the understanding that—in a randomized clinical trial with imperfect compliance (see Figure 3.7(b))—the way subjects react (Y) to treatments (X) is statistically independent of the treatment assignment (Z), the potential-outcome analyst would write $Y(x) \perp\!\!\!\perp Z$. Likewise, to convey the understanding that the assignment is randomized and hence independent of how subjects comply with the assignment, the potential-outcome analyst would use the independence constraint $Z \perp\!\!\!\perp X(z)$.

A collection of constraints of this type might sometimes be sufficient to permit a unique solution to the query of interest; in other cases, only bounds on the solution can be obtained. For example, if one can plausibly assume that a set Z of covariates satisfies the conditional independence

$$Y(x) \perp\!\!\!\perp X \,|\, Z \tag{3.53}$$

(an assumption that was termed "conditional ignorability" by (Rosenbaum and Rubin 1983), then the causal effect $P^*(Y(x) = y)$ can readily

[11] Gibbard and Harper (1976, p. 156) expressed this constraint as $A \supset [(A \mathbin{\Box\!\!\rightarrow} S) \equiv S]$.

be evaluated, using (3.52), to yield[12]

$$\begin{aligned} P^*(Y(x)=y) &= \sum_z P^*(Y(x)=y|z)P(z) \\ &= \sum_z P^*(Y(x)=y|x,z)P(z) \\ &= \sum_z P^*(Y=y|x,z)P(z) \\ &= \sum_z P(y|x,z)P(z). \end{aligned} \tag{3.54}$$

The last expression contains no counterfactual quantities (thus permitting us to drop the asterisk from P^*) and coincides precisely with the adjustment formula of (3.19), which obtains from the back-door criterion. However, the assumption of conditional ignorability (equation (3.53))—the key to the derivation of (3.54)—is not straightforward to comprehend or ascertain. Paraphrased in experimental metaphors, this assumption reads: The way an individual with attributes Z would react to treatment $X = x$ is independent of the treatment actually received by that individual.

Section 3.6.2 explains why this approach may appeal to some statisticians, even though the process of eliciting judgments about counterfactual dependencies has been extremely difficult and error-prone; instead of constructing new vocabulary and new logic for causal expressions, all mathematical operations in the potential-outcome framework are conducted within the safe confines of probability calculus. The drawback lies in the requirement of using independencies among counterfactual variables to express plain causal knowledge. When counterfactual variables are not viewed as byproducts of a deeper, process-based model, it is hard to ascertain whether *all* relevant counterfactual independence judgments have been articulated,[13] whether the judgments articulated are redundant, or whether those judgments are self-consistent. The elicitation of such counterfactual judgments can be systematized by using the following translation from graphs (see Section 7.1.4 for additional relationships).

[12]Gibbard and Harper (1976, p. 157) used the "ignorability assumption" $Y(x) \perp\!\!\!\perp X$ to derive the equality $P(Y(x)=y) = P(y|x)$.

[13]A typical oversight in the example of Figure 3.7(b) has been to write $Z \perp\!\!\!\perp Y(x)$ and $Z \perp\!\!\!\perp X(z)$ instead of $Z \perp\!\!\!\perp \{Y(x), X(z)\}$, as dictated by (3.56).

Graphs encode substantive information in both the equations and the probability function $P(u)$; the former is encoded as missing arrows, the latter as missing dashed arcs. Each parent-child family (PA_i, X_i) in a causal diagram G corresponds to an equation in the model M of (3.50). Hence, missing arrows encode exclusion assumptions, that is, claims that adding excluded variables to an equation will not change the outcome of the hypothetical experiment described by that equation. Missing dashed arcs encode independencies among disturbance terms in two or more equations. For example, the absence of dashed arcs between a node Y and a set of nodes $\{Z_1, \ldots, Z_k\}$ implies that the corresponding background variables, U_Y and $\{U_{Z_1}, \ldots, U_{Z_k}\}$, are independent in $P(u)$.

These assumptions can be translated into the potential-outcome notation using two simple rules (Pearl 1995a, p. 704)pearl:95; the first interprets the missing arrows in the graph, the second, the missing dashed arcs.

1. *Exclusion restrictions:* For every variable Y having parents PA_Y and for every set of variables S disjoint of PA_Y, we have

$$Y(pa_Y) = Y(pa_Y, s). \tag{3.55}$$

2. *Independence restrictions:* If $Z_1, \ldots, Z_k$ is any set of nodes not connected to Y via dashed arcs, we have[14]

$$Y(pa_Y) \perp\!\!\!\perp \{Z_1(pa_{Z_1}), \ldots, Z_k(pa_{Z_k})\}. \tag{3.56}$$

The independence restriction translates the independence between U_Y and $\{U_{Z_1}, \ldots, U_{Z_k}\}$ into independence between the corresponding potential-outcome variables. This follows from the observation that, once we set their parents, the variables in $\{Y, Z_1, \ldots, Z_k\}$ stand in functional relationships to the U terms in their corresponding equations.

[14] The restriction is in fact stronger, jointly applying to all instantiations of the PA variables. For example, $X \perp\!\!\!\perp Y(pa_Z)$ should be interpreted as $X \perp\!\!\!\perp \{Y(pa'_Z), Y(pa''_Z), Y(pa'''_Z), \ldots\}$, where $pa'_Z, pa''_Z, pa'''_Z, \ldots$ are the values that the set PA_Z may take on.

As an example, the model shown in Figure 3.5 displays the following parent sets:

$$PA_X = \{\emptyset\},\ PA_Z = \{X\},\ PA_Y = \{Z\}. \tag{3.57}$$

Consequently, the exclusion restrictions translate into:

$$\begin{aligned} Z(x) &= Z(y,x), && (3.58)\\ X(y) &= X(z,y) = X(z) = X, && (3.59)\\ Y(z) &= Y(z,x); && (3.60) \end{aligned}$$

the absence of a dashed arc between Z and $\{Y, X\}$ translates into the independence restriction

$$Z(x) \perp\!\!\!\perp \{Y(z), X\}. \tag{3.61}$$

Given a sufficient number of such restrictions on P^*, the analyst attempts to compute causal effects $P^*(Y(x) = y)$ using standard probability calculus together with the logical constraints (e.g. (3.52)) that couple counterfactual variables with their measurable counterparts. These constraints can be used as axioms, or rules of inference, in attempting to transform causal effect expressions of the form $P^*(Y(x) = y)$ into expressions involving only measurable variables. When such a transformation is found, the corresponding causal effect is identifiable, since P^* then reduces to P.

The question naturally arises of whether the constraints used by potential-outcome analysts are *complete*—that is, whether they are sufficient for deriving every valid statement about causal processes, interventions, and counterfactuals. To answer this question, the validity of counterfactual statements need be defined relative to more basic mathematical objects, such as possible worlds (Section 1.4.4) or structural equations (equation (3.51)). In the standard potential-outcome framework, however, the question of completeness remains open, because $Y(x, u)$ is taken as a primitive notion and because consistency constraints such as (3.52) although they appear plausible for the English expression "had X been x"—are not derived from a deeper mathematical object. This question of completeness is settled in Chapter 7, where

a necessary and sufficient set of axioms is derived from the structural semantics given to $Y(x, u)$ by (3.51).

In assessing the historical development of structural equations and potential-outcome models, one cannot overemphasize the importance of the conceptual clarity that structural equations offer vis-à-vis the potential-outcome model. The reader may appreciate this importance by attempting to judge whether the condition of (3.61) holds in a given familiar situation. This condition reads: "the value that Z would obtain had X been x is jointly independent of both X and the value that Y would obtain had Z been z." (In the structural representation, the sentence reads: "Z shares no cause with either X or Y, except for X itself, as shown in Figure 3.5.") The thought of having to express, defend, and manage formidable counterfactual relationships of this type may explain why the enterprise of causal inference is currently viewed with such awe and despair among rank-and-file epidemiologists and statisticians—and why economists and social scientists continue to use structural equations instead of the potential-outcome alternatives advocated in Holland (1988), Angrist et al. (1996), and Sobel (1998). On the other hand, the algebraic machinery offered by the potential-outcome notation, once a problem is properly formalized, can be quite powerful in refining assumptions, deriving probabilities of counterfactuals, and verifying whether conclusions follow from premises—as we demonstrate in Chapter 9. The translation given in (3.51)–(3.56) should help researchers combine the best features of the two approaches.

3.6.4 Relations to Robins's G-Estimation

Among the investigations conducted in the potential-outcome framework, the one closest in spirit to the structural analysis described in this chapter is Robins's work on "causally interpreted structured tree graphs" (Robins 1986, 1987). Robins was the first to realize the potential of Neyman's counterfactual notation $Y(x)$ as a general mathematical language for causal inference, and he used it to extend Rubin's (1978) "time-independent treatment" model to studies with direct and indirect effects and time-varying treatments, concomitants, and outcomes.

Robins considered a set $V = \{V_1, \ldots, V_M\}$ of temporally ordered

discrete random variables (as in Figure 3.3) and asked under what conditions one can identify the effect of control policy $g : X = x$ on outcomes $Y \subseteq V \backslash X$, where $X = \{X_1, \ldots, X_K\} \subseteq V$ are the temporally ordered and potentially manipulable treatment variables of interest. The causal effect of $X = x$ on Y was expressed as the probability

$$P(y|g = x) \stackrel{\Delta}{=} P\{Y(x) = y\},$$

where the counterfactual variable $Y(x)$ stands for the value that outcome variables Y would take had the treatment variables X been x.

Robins showed that $P(y|g = x)$ is identified from the distribution $P(v)$ if each component X_k of X is "assigned at random, given the past," a notion explicated as follows. Let L_k be the variables occurring between X_{k-1} and X_k, with L_1 being the variables preceding X_1. Write $\overline{L}_k = (L_1, \ldots, L_k)$, $L = \overline{L}_K$, and $\overline{X}_k = (X_1, \ldots, X_k)$, and define $\overline{X}_0$, $\overline{L}_0$, $\overline{V}_0$ to be identically zero. The treatment $X_k = x_k$ is said to be *assigned at random, given the past*, if the following relation holds:

$$(Y(x) \perp\!\!\!\perp X_k | \overline{L}_k, \overline{X}_{k-1} = \overline{x}_{k-1}). \tag{3.62}$$

Robins further proved that, if (3.62) holds for every k, then the causal effect is given by

$$P(y|g = x) = \sum_{\overline{l}_K} P(y|\overline{l}_K, \overline{x}_K) \prod_{k=1}^{K} P(l_k|\overline{l}_{k-1}, \overline{x}_{k-1}), \tag{3.63}$$

an expression he called the "G-computation algorithm formula." This expression can be derived by applying condition (3.62) iteratively, as in the derivation of (3.54). If X is univariate, then (3.63) reduces to the standard adjustment formula

$$P(y|g = x) = \sum_{l_1} P(y|x, l_1) P(l_1),$$

paralleling (3.54). Likewise, in the special structure of Figure 3.3, (3.63) reduces to (3.18).

To place this result in the context of our analysis in this chapter, we note that the class of semi-Markovian models satisfying assumption

(3.62) corresponds to complete DAGs in which all arrowheads pointing to X_k originate from observed variables. Indeed, in such models, the parents $PA_k = \overline{L}_k, \overline{X}_{k-1}$ of variable X_k satisfy the back-door condition of Definition 3.3.1,

$$(X_k \perp\!\!\!\perp Y | PA_k)_{G_{\underline{X}_k}},$$

which implies (3.62).[15] This class of models falls under Theorem 3.2.5, which states that all causal effects in this class are identifiable and are given by the truncated factorization formula of (3.14); the formula coincides with (3.63) after marginalizing over the uncontrolled covariates.

The structural analysis introduced in this chapter supports and generalizes Robins's result from a new theoretical perspective. First, on the technical front, this analysis offers systematic ways of managing models with unmeasured confounders (i.e., unobserved parents of control variables, as in Figures 3.8(d)–(g)), where Robins's starting assumption (3.62) is inapplicable. Second, on the conceptual front, the structural framework represents a fundamental shift from the vocabulary of counterfactual independencies (e.g. (3.62)) to the vocabulary of processes and mechanisms, from which human judgment of counterfactuals originates. Although expressions of counterfactual independencies can be engineered to facilitate algebraic derivations of causal effects (as in (3.54)), articulating the right independencies for a problem or assessing the assumptions behind such independencies may often be the hardest part of the problem. In the structural framework, the counterfactual expressions themselves are derived (if needed) from a mathematical theory (as in (3.56) and (3.61)). Still, Robins's pioneering research has proven (i) that algebraic methods can handle causal analysis in complex multistage problems and (ii) that causal effects in such problems can be reduced to estimable quantities (see also Sections 3.6.1 and 4.4).

[15] Alternatively, (3.62) can be obtained by applying the translation rule of (3.56) to graphs with no confounding arcs between X_k and $\{Y, PA_k\}$. Note, however, that the implication goes only one way; Robins's condition is the weakest assumption needed for identifying the causal effect.

Postscript

The work recounted in this chapter sprang from two simple ideas that totally changed my attitude toward causality. The first idea arose in the summer of 1990, while I was working with Tom Verma on "A Theory of Inferred Causation" (Pearl and Verma 1991; see also Chapter 2). We played around with the possibility of replacing the parents-child relationship $P(x_i|pa_i)$ with its functional counterpart $x_i = f_i(pa_i, u_i)$ and, suddenly, everything began to fall into place: we finally had a mathematical object to which we could attribute familiar properties of physical mechanisms instead of those slippery epistemic probabilities $P(x_i|pa_i)$ with which we had been working so long in the study of Bayesian networks. Danny Geiger, who was writing his dissertation at that time, asked with astonishment: "Deterministic equations? Truly deterministic?" Although we knew that deterministic structural equations have a long history in econometrics, we viewed this representation as a relic of the past. For us at UCLA in the early 1990s, the idea of putting the semantics of Bayesian networks on a deterministic foundation seemed a heresy of the worst kind.

The second simple idea came from Peter Spirtes's lecture at the International Congress of Philosophy of Science (Uppsala, Sweden, 1991). In one of his slides, Peter illustrated how a causal diagram would change when a variable is manipulated. To me, that slide of Spirtes's—when combined with the deterministic structural equations—was the key to unfolding the manipulative account of causation and led to most of the explorations described in this chapter.

I should really mention another incident that contributed to this chapter. In early 1993 I read the fierce debate between Arthur Goldberger and Nanny Wermuth on the meaning of structural equations (Goldberger 1992; Wermuth 1992). It suddenly hit me that the century-old tension between economists and statisticians stems from simple semantic confusion: Statisticians read structural equations as statements about $E(Y|x)$, while economists read them as $E(Y|do(x))$. This would explain why statisticians claim that structural equations have no meaning and why economists retort that statistics has no substance. I wrote a technical report, "On the Statistical Interpretation of Structural Equations" (Pearl 1993c), hoping to see the two camps embrace

in reconciliation. Nothing of the sort happened. The statisticians in the dispute continued to insist that anything that is not interpreted as $E(Y|x)$ simply lacks meaning. The economists, in contrast, are still trying to decide if it was $do(x)$ that they have been meaning to say all along.

Encouraging colleagues receive far too little credit in official channels, considering the immense impact they have on the encouraged. I must take this opportunity to acknowledge four colleagues who saw clarity shining through the $do(x)$ operator before it gained popularity: Steffen Lauritzen, David Freedman, James Robins, and Philip Dawid. Phil showed special courage in printing my paper in *Biometrika* (Pearl 1995a), the journal founded by causality's worst adversary—Karl Pearson.