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 observed 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 (3.30) 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 (3.21), 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 (3.20)).

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 independent and identically distributted) i.i.d 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

[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.

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 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 ac-

[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).

[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).

ceptable 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.

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

$$x_i = f_i(pa_i, u_i), \quad i = 1, \ldots, n, \tag{3.52}$$

[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).

as in (3.5). 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

$$Y(x, u) \triangleq Y_{M_x}(u). \tag{3.53}$$

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.

(3.53) 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; e.g., 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]

$$X = x \implies Y(x) = Y \tag{3.54}$$

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 (3.54). 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 (3.54) 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,

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

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

$$Y(x) \perp\!\!\!\perp X | Z \tag{3.55}$$

(an assumption that was termed "conditional ignorability" by [Rosenbaum and Rubin, 1983], then the causal effect $P^*(Y(x) = y)$ can readily be evaluated, using (3.54), to yield[12]

$$\begin{aligned} P^*(Y(x) = y) &= \sum_z P^*(Y(x) = y|z)P(z) \\ &= \sum_z P^*(Y(x) = y|x, z)P(z) \\ &= \sum_z P^*(Y = y|x, z)P(z) \\ &= \sum_z P(y|x, z)P(z). \end{aligned} \tag{3.56}$$

The last expression contains no counterfactual quantities (thus permitting us to drop the asterisk from P^*) and coincides precisely with the adjustment formula of (3.21), which obtains from the back-door criterion. However, the assumption of conditional ignorability (3.55)—the key to the derivation of (3.56)—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

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

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).

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 (3.52). 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]; 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

$$Y(pa_Y) = Y(pa_Y, s). \tag{3.57}$$

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

$$Y(pa_Y) \perp\!\!\!\perp \{Z_1(pa_{Z_1}), \ldots, Z_k(pa_{Z_k})\}. \tag{3.58}$$

[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 (3.58).

[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.

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.

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

$$PA_X = \{\emptyset\},\ PA_Z = \{X\},\ PA_Y = \{Z\}. \tag{3.59}$$

Consequently, the exclusion restrictions translate into:

$$Z(x) = Z(y, x), \tag{3.60}$$
$$X(y) = X(z, y) = X(z) = X, \tag{3.61}$$
$$Y(z) = Y(z, x); \tag{3.62}$$

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

$$Z(x) \perp\!\!\!\perp \{Y(z), X\}. \tag{3.63}$$

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. (3.54)) 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 (3.53)). 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 (3.54) 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 (3.53).

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 (3.63) 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 (3.53)–(3.58) 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

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

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:

$$(Y(x) \perp\!\!\!\perp X_k | \overline{L}_k, \overline{X}_{k-1} = \overline{x}_{k-1}). \tag{3.64}$$

Robins further proved that, if (3.64) holds for every k, then the causal effect is given by

$$P(y|g = x) = \sum_{\overline{l}_K} P(y|\overline{l}_K, \overline{x}_K) \prod_{k=1}^{K} P(l_k|\overline{l}_{k-1}, \overline{x}_{k-1}), \tag{3.65}$$

an expression he called the "G-computation algorithm formula." This expression can be derived by applying condition (3.64) iteratively, as in the derivation of (3.56). If X is univariate, then (3.65) reduces to the standard adjustment formula

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

paralleling (3.56). Likewise, in the special structure of Figure 3.3, (3.65) reduces to (3.20).

To place this result in the context of our analysis in this chapter, we note that the class of semi-Markovian models satisfying assumption (3.64) 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 (3.64).[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 (3.16); the formula coincides with (3.65) after marginalizing over the uncontrolled covariates.

The structural analysis introduced in this chapter supports and generalizes Robins' 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 (ch3-robins:eq1) is inapplicable. Second, on the conceptual front, the structural framework represents a fundamental shift from the vocabulary of counterfactual independencies (e.g., (3.64)) 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 (3.56)), 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 (3.58) and (3.63)). 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, (3.64) can be obtained by applying the translation rule of (3.58) 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.