(cf. equation (1.23)). Of special importance is the expectation of the product $(g(X, Y) = (X - E(X))(Y - E(Y))$, which is known as the *covariance* of X and Y,

$$\sigma_{XY} \triangleq E\left[(X - E(X))(Y - E(Y))\right],$$

and which is often normalized to yield the *correlation coefficient*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

and the *regression coefficient* $\in$ of X on Y $\ni$

$$r_{XY} \triangleq \rho_{XY} \frac{\sigma_X}{\sigma_Y} = \frac{\sigma_{XY}}{\sigma_Y^2}.$$

The *conditional* variance, covariance, and correlation coefficient, given $Z = z$, are defined in a similar manner, using the conditional distribution $P(x, y \mid z)$ in taking expectations. In particular, the *conditional correlation coefficient*, given $Z = z$, is defined as

$$\rho_{XY|z} = \frac{\sigma_{XY|z}}{\sigma_{X|z}\, \sigma_{Y|z}}. \tag{1.24}$$

Additional properties, specific to normal distributions, will be reviewed in Chapter 5 (Section 5.2.1).

The foregoing definitions apply to discrete random variables – that is, variables that take on finite or denumerable sets of values on the real line. The treatment of expectation and correlation is more often applied to continuous random variables, which are characterized by a *density function* $f(x)$ defined as follows:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

for any two real numbers a and b with $a < b$. If X is discrete, then $f(x)$ coincides with the probability function $P(x)$, once we interpret the integral through the translation

$$\int_{-\infty}^{\infty} f(x)dx \Longleftrightarrow \sum_x P(x). \tag{1.25}$$

Readers accustomed to continuous analysis should bear this translation in mind whenever summation is used in this book. For example, the expected value of a continuous random variable X can be obtained from (1.21), to read

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx,$$

with analogous translations for the variance, correlation, and so forth.

We now turn to define *conditional independence* relationships among variables, a central notion in causal modelling.

1.1.5 Conditional Independence and Graphoids

Definition 1.1.2 (Conditional Independence)
Let $V = \{V_1, V_2, \ldots\}$ be a finite set of variables. Let $P(\cdot)$ be a joint probability function over the variables in V, and let X, Y, Z stand for any three subsets of variables in V. The sets X and Y are said to be conditionally independent given Z *if*

$$P(x \mid y, z) = P(x \mid z) \quad \textit{whenever} \quad P(y, z) > 0. \tag{1.26}$$

In words, learning the value of Y does not provide additional information about X, once we know Z. (Metaphorically, Z "screens off" X from Y.)

Equation (1.26) is a terse way of saying the following: For any configuration x of the variables in the set X and for any configurations y and z of the variables in Y and Z satisfying $P(Y = y, Z = z) > 0$, we have

$$P(X = x \mid Y = y, Z = z) = P(X = x \mid Z = z). \tag{1.27}$$

We will use Dawid's (1979) notation $(X \perp\!\!\!\perp Y \mid Z)_P$ or simply $(X \perp\!\!\!\perp Y \mid Z)$ to denote the conditional independence of X and Y given Z; thus,

$$(X \perp\!\!\!\perp Y \mid Z)_P \quad \text{iff} \quad P(x \mid y, z) = P(x \mid z) \tag{1.28}$$

for all values x, y, z such that $P(y, z) > 0$. Unconditional independence (also called *marginal independence*) will be denoted by $(X \perp\!\!\!\perp Y \mid \emptyset)$; that is,

$$(X \perp\!\!\!\perp Y \mid \emptyset) \text{ iff } P(x \mid y) = P(x) \quad \text{whenever} \quad P(y) > 0 \tag{1.29}$$

("iff" is shorthand for "if and only if"). Note that $(X \perp\!\!\!\perp Y \mid Z)$ implies the conditional independence of all pairs of variables $V_i \in X$ and $V_j \in Y$, but the converse is not necessarily true.

The following is a (partial) list of properties satisfied by the conditional independence relation $(X \perp\!\!\!\perp Y \mid Z)$.

Symmetry: $(X \perp\!\!\!\perp Y \mid Z) \Longrightarrow (Y \perp\!\!\!\perp X \mid Z)$.

Decomposition: $(X \perp\!\!\!\perp YW \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid Z)$.

Weak union: $(X \perp\!\!\!\perp YW \mid Z) \Longrightarrow (X \perp\!\!\!\perp Y \mid ZW)$.

Contraction: $(X \perp\!\!\!\perp Y \mid Z)\ \&\ (X \perp\!\!\!\perp W \mid ZY) \Longrightarrow (X \perp\!\!\!\perp YW \mid Z)$.

Intersection: $(X \perp\!\!\!\perp W \mid ZY)\ \&\ (X \perp\!\!\!\perp Y \mid ZW) \Longrightarrow (X \perp\!\!\!\perp YW \mid Z)$.

(Intersection is valid in strictly positive probability distributions.)

The proof of these properties can be derived by elementary means from (1.28) and the basic axioms of probability theory.[4] These properties were called *graphoid axioms* by

[4] These properties were first introduced by Dawid (1979) and Spohn (1980) in a slightly different form, and were independently proposed by Pearl and Paz (1987) to characterize the relationships between graphs and informational relevance. Geiger and Pearl (1993) present an in-depth analysis.

to constants x.[10] *Denote by* $\boldsymbol{P}_*$ *the set of all interventional distributions* $P_x(v), X \subseteq V$, *including* $P(v)$, *which represents no intervention (i.e.,* $X = \phi$*). A DAG G is said to be a* causal Bayesian network *compatible with* $\boldsymbol{P}_*$ *if and only if the following three conditions hold for every* $P_x \in P_*$:

(i) $P_x(v)$ *is Markov relative to G;*

(ii) $P_x(v_i) = 1$ *for all* $V_i \in X$ *whenever* v_i *is consistent with* $X = x$;

(iii) $P_x(v_i \mid pa_i) = P(v_i \mid pa_i)$ *for all* $V_i \notin X$ *whenever* pa_i *is consistent with* $X = x$, i.e., each $P(v_i \mid pa_i)$ *remains invariant to interventions not involving* V_i.

Definition 1.3.1 imposes constraints on the interventional space $\boldsymbol{P}_*$ that permit us to encode this vast space economically, in the form of a single Bayesian network G. These constraints enable us to compute the distribution $P_x(v)$ resulting from any intervention $do(X = x)$ as a *truncated factorization*

$$P_x(v) = \prod_{\{i \mid V_i \notin X\}} P(v_i \mid pa_i) \qquad \text{for all } v \text{ consistent with } x, \tag{1.37}$$

which follows from Definition 1.3.1 and justifies the family deletion procedure on G, as in (1.36). It is not hard to show that, whenever G is a causal Bayes network with respect to $\boldsymbol{P}_*$, the following two properties must hold.

Property 1
For all i,

$$P(v_i \mid pa_i) = P_{pa_i}(v_i). \tag{1.38}$$

Property 2
For all i and for every subset S of variables disjoint of $\{V_i, PA_i\}$, *we have*

$$P_{pa_i, s}(v_i) = P_{pa_i}(v_i). \tag{1.39}$$

Property 1 renders every parent set PA_i *exogenous* relative to its child V_i, ensuring that the conditional probability $P(v_i \mid pa_i)$ coincides with the effect (on V_i) of setting PA_i to pa_i by external control. Property 2 expresses the notion of invariance; once we control its direct causes PA_i, no other interventions will affect the probability of V_i.

1.3.2 Causal Relationships and Their Stability

This mechanism-based conception of interventions provides a semantical basis for notions such as "causal effects" or "causal influence," to be defined formally and analyzed in Chapters 3 and 4. For example, to test whether a variable X_i has a causal influence on another variable X_j, we compute (using the truncated factorization formula of (1.37)) the (marginal) distribution of X_j under the actions $do(X_i = x_i)$ – namely, $P_{x_i}(x_j)$ for all

[10] The notation $P_x(v)$ will be replaced in subsequent chapters with $P(v \mid do(x))$ and $P(v \mid \hat{x})$ to facilitate algebraic manipulations.

R_1: Orient $b - c$ into $b \rightarrow c$ whenever there is an arrow $a \rightarrow b$ such that a and c are nonadjacent.

R_2: Orient $a - b$ into $a \rightarrow b$ whenever there is chain $a \rightarrow c \rightarrow b$.

R_3: Orient $a - b$ into $a \rightarrow b$ whenever there are two chains $a - c \rightarrow b$ and $a - d \rightarrow b$ such that c and d are nonadjacent.

R_4: Orient $a - b$ into $a \rightarrow b$ whenever there are two chains $a - c \rightarrow d$ and $c \rightarrow d \rightarrow b$ such that c and b are nonadjacent and a and d are adjacent.

Meek (1995) showed that these four rules are also sufficient, so that repeated application will eventually orient *all* arrows that are common to the equivalence class of D_0. Moreover, R_4 is not required if the starting orientation is limited to v-structures.

Another systematization is offered by an algorithm due to Dor and Tarsi (1992) that tests (in polynomial time) if a given partially oriented acyclic graph can be fully oriented without creating a new v-structure or a directed cycle. The test is based on recursively removing any vertex v that has the following two properties:

1. no edge is directed outward from v;
2. every neighbor of v that is connected to v through an undirected edge is also adjacent to all the other neighbors of v.

A partially oriented acyclic graph has an admissible extension in a DAG if and only if all its vertices can be removed in this fashion. Thus, to find the maximally oriented pattern, we can (i) separately try the two orientations, $a \rightarrow b$ and $a \leftarrow b$, for every undirected edge $a - b$, and (ii) test whether both orientations, or just one, have extensions. The set of uniquely orientable arrows constitutes the desired maximally oriented pattern. Additional refinements can be found in Chickering (1995), Andersson et al. (1997), and Moole (1997).

Latent structures, however, require special treatment, because the constraints that a latent structure imposes upon the distribution cannot be completely characterized by any set of conditional independence statements. Fortunately, certain sets of those independence constraints can be identified (Verma and Pearl 1990); this permits us to recover valid fragments of latent structures.

2.6 RECOVERING LATENT STRUCTURES

When Nature decides to "hide" some variables, the observed distribution $\hat{P}$ need no longer be stable relative to the observable set O. That is, we are no longer guaranteed that, among the minimal latent structures compatible with $\hat{P}$, there exists one that has a DAG structure. Fortunately, rather ~~then~~ than having to search through this unbounded space of latent structures, the search can be confined to graphs with finite and well-defined structures. For every latent structure L, there is a dependency-equivalent latent structure (the projection) of L on O in which every unobserved node is a root node with exactly two observed children. We characterize this notion explicitly as follows.

over the space of possible structures to seek the one(s) with the highest posterior score. Methods based on this approach have the advantage of operating well under small-sample conditions, but they encounter difficulties in coping with hidden variables. The assumption of parameter independence, which is made in all practical implementations of the Bayesian approach, induces preferences toward models with fewer parameters and hence toward minimality. Likewise, parameter independence can be justified only when the parameters represent mechanisms that are free to change independently of one another – that is, when the system is autonomous and hence stable.

Postscript for the Second Edition

Work on causal discovery has been pursued vigorously by the TETRAD group at Carengie Mellon University and reported in Spirtes et al. (2000), Robins et al. (2003), Scheines (2002), and Moneta and Spirtes (2006).

Add sentence

Applications of causal discovery in economics are reported in Bessler (2002), Swanson and Granger (1997), and Demiralp and Hoover (2003). Gopnik et al. (2004) applied causal Bayesian networks to explain how children acquire causal knowledge from observations and actions (see also Glymour 2001).

Hoyer et al. (2006) and Shimizu et al. (2005, 2006) have proposed a new scheme of discovering causal directionality, based not on conditional independence but on functional composition. The idea is that in a linear model $X \rightarrow Y$ with non-Gaussian noise, variable Y is a linear combination of two independent noise terms. As a consequence, $P(y)$ is a convolution of two non-Gaussian distributions and would be, figuratively speaking, "more Gaussian" than $P(x)$. The relation of "more Gaussian than" can be given precise numerical measure and used to infer directionality of certain arrows.

Tian and Pearl (2001a,b) developed yet another method of causal discovery based on the detection of "shocks," or spontaneous local changes in the environment which act like "Nature's interventions," and unveil causal directionality toward the consequences of those shocks.

Verma and Pearl (1990) noted that two latent structures may entail the same set of conditional independencies and yet impose different equality constraints on the joint distributions. These constraints, dubbed "dormant independencies," were characterized systematically in Tian and Pearl (2002b) and Shpitser and Pearl (2008); they promise to provide a powerful new discovery tool for structure learning.

A program of benchmarks of causal discovery algorithms, named "Causality Workbench," has been reported by Guyon et al. (2008a,b; http://clopinet.com/causality). Regular contests are organized in which participants are given real data or data generated by a concealed causal model, and the challenge is to predict the outcome of a select set of interventions.

Spirtes, Glymour, Scheines, and Tillman (2010) summarize the current state of the art in causal discovery ☺

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

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

(an assumption that was termed "conditional ignorability" by Rosenbaum and Rubin 1983), then the causal effect $P^*(Y(x) = y)$ can readily be evaluated, using (3.52), to yield[12]

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

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 by-products 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).

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.

---

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

[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).

Phil showed special courage in printing my paper in *Biometrika* (Pearl 1995a), the journal founded by causality's worst adversary – Karl Pearson.

Postscript for the Second Edition

Complete identification results

A key identification condition, which generalizes all the criteria established in this chapter, has been derived by Jin Tian. It reads:

Theorem 3.6.1 (Tian and Pearl, 2002a)
A sufficient condition for identifying the causal effect $P(y \mid do(x))$ is that there exists no bi-directed path (i.e., a path composed entirely of bi-directed arcs) between X and any of its children.[15]

Remarkably, the theorem asserts that, as long as every child of X (on the pathways to Y) is not reachable from X via a bi-directed path, then, regardless of how complicated the graph, the causal effect $P(y \mid do(x))$ is identifiable. All identification criteria discussed in this chapter are special cases of the one defined in this theorem. For example, in Figure 3.5 $P(y \mid do(x))$ can be identified because the two paths from X to Z (the only child of X) are not bi-directed. In Figure 3.7, on the other hand, there is a path from X to Z_1 traversing only bi-directed arcs, thus violating the condition of Theorem 3.6.1, and $P(y \mid do(x))$ is not identifiable.

Note that all graphs in Figure 3.8 and none of those in Figure 3.9 satisfy the condition above. Tian and Pearl (2002a) further showed that the condition is both sufficient and necessary for the identification of $P(v \mid do(x))$, where V includes all variables except X. A necessary and sufficient condition for identifying $P(w \mid do(z))$, with W and Z two arbitrary sets, was established by Shpitser and Pearl (2006b). Subsequently, a complete graphical criterion was established for determining the identifiability of *conditional* interventional distributions, namely, expressions of the type $P(y \mid do(x), z)$ where X, Y, and Z are arbitrary sets of variables (Shpitser and Pearl 2006a).

These results constitute a complete characterization of causal effects in graphical models. They provide us with polynomial time algorithms for determining whether an arbitrary quantity invoking the $do(x)$ operator is identified in a given semi-Markovian model and, if so, what the estimand of that quantity is. Remarkably, one corollary of these results also states that the *do*-calculus is complete, namely, a quantity $Q = P(y \mid do(x), z)$ is identified if and only if it can be reduced to a *do*-free expression using the three rules of Theorem 3.4.1.[16] Tian and Shpitser (2010) provide a comprehensive summary of these results.

Add sentence

Applications and Critics

Gentle introductions to the concepts developed in this chapter are given in (Pearl 2003c) and (Pearl 2009a). Applications of causal graphs in epidemiology are reported in Robins

2009a

[15] Before applying this criterion, one may delete from the causal graph all nodes that are not ancestors of Y.

[16] This was independently established by Huang and Valtorta (2006).

(2001), Hernán et al. (2002), Hernán et al. (2004), Greenland and Brumback (2002), Greenland et al. (1999a,b) Kaufman et al. (2005), Petersen et al. (2006), Hernández-Díaz et al. (2006), VanderWeele and Robins (2007) and Glymour and Greenland (2008).

Interesting applications of the front-door criterion (Section 3.3.2) were noted in social science (Morgan and Winship 2007) and economics (Chalak and White 2006).

Some advocates of the "potential outcome" approach have been most resistant to accepting graphs or structural equations as the basis for causal analysis and, lacking these conceptual tools, were unable to address the issue of covariate selection (Rosenbaum 2002, p. 76; Rubin 2007, 2008a) and were led to dismiss important scientific concepts as "ill-defined," "deceptive," "confusing" (Holland 2001; Rubin 2004, 2008b), and worse (Rubin 2009). Lauritzen (2004) and Heckman (2005) have criticized this attitude; Pearl (2009) demonstrates its fallacies and illuminates its damaging consequences.

Equally puzzling are concerns of some philosophers (Cartwright 2007; Woodward 2003) and economists (Heckman 2005) that the *do*-operator is too local to model complex, real-life policy interventions, which sometimes affect several mechanisms at once and often involve conditional decisions, imperfect control, and multiple actions. These concerns emerge from conflating the mathematical definition of a relationship (e.g., causal effect) with the technical feasibility of testing that relationship in the physical world. While the *do*-operator is indeed an ideal mathematical tool (not unlike the *derivative* in differential calculus), it nevertheless permits us to specify and analyze interventional strategies of great complexity. Readers will find examples of such strategies in Chapter 4, and a further discussion of this issue in Chapter 11 (Sections 11.4.3–11.4.6 and Section 11.5.4).

Chapter Road Map to the Main Results

The three key results in this chapter are: 1. The control of confounding, 2. The evaluation of policies, and 3. The evaluation of counterfactuals.

1. The problem of controlling confounding bias is resolved through the back-door condition (Theorem 3.3.2, pp.79–80) – a criterion for selecting a set of covariates that, if adjusted for, would yield an unbiased estimate of causal effects.
2. The policy evaluation problem – to predict the effect of interventions from nonexperimental data – is resolved through the *do*-calculus (Theorem 3.4.1, pp. 85–86) and the graphical criteria that it entails (Theorem 3.3.4, p. 83; Theorem 3.6.1, p. 105). The completeness of *do*-calculus implies that any (nonparametric) policy evaluation problem that is not supported by an identifying graph, or an equivalent set of causal assumptions, can be proven "unsolvable."
3. Finally, equation (3.51) provides a formal semantics for counterfactuals, through which joint probabilities of counterfactuals can be defined and evaluated in the framework of scientific theories (see Chapter 7). This semantics will enable us to develop a number of techniques for counterfactual analyses (Chapters 8–11), including the Mediation Formula (equations (4.17)–(4.18)) a key tool for assessing causal pathways in nonlinear models.

4.5.5 Indirect Effects

Remarkably, the definition of the natural direct effect (4.11) can easily be turned around and provide an operational definition for the *indirect effect* – a concept shrouded in mystery and controversy, because it is impossible, using the $do(x)$ operator, to disable the direct link from X to Y so as to let X influence Y solely via indirect paths.

The natural indirect effect, IE, of the transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z to whatever value it would have attained had X been set to $X = x'$. Formally, this reads (Pearl 2001c):

$$IE_{x,x'}(Y) \triangleq E[(Y(x, Z(x'))) - E(Y(x))], \tag{4.14}$$

which is almost identical to the direct effect (equation (4.11)) save for exchanging x and x'.

Indeed, it can be shown that, in general, the total effect TE of a transition is equal to the *difference* between the direct effect of that transition and the indirect effect of the reverse transition. Formally,

$$TE_{x,x'}(Y) \triangleq E(Y(x) - Y(x')) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{4.15}$$

In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{4.16}$$

Since each term above is based on an independent operational definition, this quality constitutes a formal justification for the additive formula.

Note that the indirect effect has clear policy-making implications. For example: in a hiring discrimination environment, a policy maker may be interested in predicting the gender mix in the work force if gender bias is eliminated and all applicants are treated equally – say, the same way that males are currently treated. This quantity will be given by the indirect effect of gender on hiring, mediated by factors such as education and aptitude, which may be gender-dependent.

More generally, a policy maker may be interested in the effect of issuing a directive to a select set of subordinate employees, or in carefully controlling the routing of messages in a network of interacting agents. Such applications motivate the analysis of *path-specific effects*, that is, the effect of X on Y through a selected set of paths (Avin et al. 2005).

Note that in all these cases, the policy intervention invokes the selection of signals to be sensed, rather than variables to be fixed. Pearl (2001c) has suggested therefore that signal sensing is more fundamental to the notion of causation than manipulation; the latter being but a crude way of stimulating the former in experimental setup. (See Section 11.4.5.)

It is remarkable that counterfactual quantities like DE and IE that could not be expressed in terms of $do(x)$ operators, and appear therefore void of empirical content, can, under certain conditions, be estimated from empirical studies. A general analysis of those conditions is given in Shpitser and Pearl (2007).

We shall see additional examples of this "marvel of formal analysis" in Chapters 7, 9, and 11. It constitutes an unassailable argument in defense of counterfactual analysis, as expressed in Pearl (2000) against the stance of Dawid (2000).

New Page 132

4.5.5 Indirect Effects and the Mediation Formula

Remarkably, the definition of the natural direct effect (4.11) can easily be turned around and provide an operational definition for the *indirect effect* – a concept shrouded in mystery and controversy, because it is impossible, using the $do(x)$ operator, to disable the direct link from X to Y so as to let X influence Y solely via indirect paths.

The natural indirect effect, IE, of the transition from x to x' is defined as the expected change in Y affected by holding X constant, at $X = x$, and changing Z to whatever value it would have attained had X been set to $X = x'$. Formally, this reads (Pearl 2001c):

$$IE_{x,x'}(Y) \triangleq E[(Y(x, Z(x'))) - E(Y(x))], \tag{4.14}$$

We see that, in general, the total effect TE of a transition is equal to the *difference* between the direct effect of that transition and the indirect effect of the reverse transition:

$$TE_{x,x'}(Y) \triangleq E(Y(x') - Y(x)) = DE_{x,x'}(Y) - IE_{x',x}(Y). \tag{4.15}$$

In linear models, where reversal of transitions amounts to negating the signs of their effects, (4.15) provides a formal justification for the standard additive formula

$$TE_{x,x'}(Y) = DE_{x,x'}(Y) + IE_{x,x'}(Y). \tag{4.16}$$

In the simple case of unconfounded mediators, the natural direct and indirect effects are estimable through two regression equations called the Mediation Formula:

$$DE_{x,x'}(Y) = \sum_z [E(Y|x', z) - E(Y|x, z)]P(z|x). \tag{4.17}$$

$$IE_{x,x'}(Y) = \sum_z E(Y|x, z)[P(z|x') - P(z|x)] \tag{4.18}$$

These provide two ubiquitous measures of mediation effects, applicable to any nonlinear system, any distribution, and any type of variables (Pearl 2009b, 2010b).

Note that the indirect effect has clear policy-making implications. For example: in a hiring discrimination environment, a policy maker may be interested in predicting the gender mix in the work force if gender bias is eliminated and all applicants are treated equally—say, the same way that males are currently treated. This quantity will be given by the indirect effect of gender on hiring, mediated by factors such as education and aptitude, which may be gender-dependent. See (Pearl 2001c, 2010b) for more examples.

More generally, a policy maker may be interested in the effect of motivating a select set of subordinate employees, or of controlling the routing of messages in a network of interacting agents. Such applications motivate the analysis of *path-specific effects*, that is, the effect of X on Y through a selected set of paths (Avin et al. 2005).

In all these cases, the policy intervention invokes the selection of signals to be sensed, rather than variables to be fixed. Pearl (2001c) has suggested therefore that signal sensing is more fundamental to the notion of causation than manipulation; the latter being but a crude way of stimulating the former in experimental setup (see Section 11.4.5). A general characterization of counterfactuals that are empirically testable is given in Chapters 7, 9, 11, and in Shpitser and Pearl (2007).

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

5.3.3 Causal Effects: The Interventional Interpretation of Structural Equation Models

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

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

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

model M yields

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

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

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

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

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

and the parameter of interest is $\lambda = P(y_0 | x_0)$?

if X and Y are dichotomous, then the marginal probability $P(x)$ certainly "involves" parameters such as

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

~~as well as their ratio?~~ and λ "involves" their ratio ☺

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

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

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

The Mystical Error Term Revisited

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

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

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

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

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

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

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

5.6 Postscript for the Second Edition

5.6.1 An Econometric Awakening?

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

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

5.6.2 Identification in Linear Models

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

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

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

5.6.3 Robustness of Causal Claims

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

When several distinct sets of assumptions give rise to k distinct estimands for a parameter α, that parameter is called k-identified; the higher the k, the more robust are claims based on α, because equality among these estimands imposes $k - 1$ constraints on the covariance matrix which, if satisfied in the data, indicate an agreement among k distinct sets of assumptions, thus supporting their validity. A typical example emerges when several (independent) instrumental variables are available $Z_1, Z_2, \ldots, Z_k$ for a single link $X \rightarrow Y$, which yield the equalities $\alpha = r_{YZ_1}/r_{XZ_1} = r_{YZ_2}/r_{XZ_2} = \cdots = r_{YZ_k}/r_{XZ_k}$.

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

Acknowledgments

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

Thus, similarities and priorities – if they are ever needed – may be read into the $do(\cdot)$ operator as an afterthought (see discussion following (3.11) and Goldszmidt and Pearl 1992), but they are not basic to the analysis.

The structural account answers the mental representation question by offering a parsimonious encoding of knowledge from which causes, counterfactuals, and probabilities of counterfactuals can be derived by effective algorithms. However, this effectiveness is partly acquired by limiting the counterfactual antecedent to conjunction of elementary propositions. Disjunctive hypotheticals, such as "if Bizet and Verdi were compatriots," usually lead to multiple solutions and hence to nonunique probability assignments.

7.4.2 Axiomatic Comparison

If our assessment of interworld distances comes from causal knowledge, the question arises of whether that knowledge does not impose its own structure on distances, a structure that is not captured in Lewis's logic. Phrased differently: By agreeing to measure closeness of worlds on the basis of causal relations, do we restrict the set of counterfactual statements we regard as valid? The question is not merely theoretical. For example, Gibbard and Harper (1976) characterized decision-making conditionals (i.e., sentences of the form "If we do A, then B") using Lewis's general framework, whereas our $do(\cdot)$ operator is based on functions representing causal mechanisms; whether the two formalisms are identical is uncertain.[21]

We now show that the two formalisms are identical for recursive systems; in other words, composition and effectiveness hold with respect to Lewis's closest-world framework whenever recursiveness does. We begin by providing a version of Lewis's logic for counterfactual sentences (from Lewis 1973c).

Rules

(1) If A and $A \Rightarrow B$ are theorems, then so is B.

(2) If $(B_1 \ \& \ \ldots) \Rightarrow C$ is a theorem, then so is $((A \mathrel{\square\!\!\rightarrow} B_1)\ldots) \Rightarrow (A \mathrel{\square\!\!\rightarrow} C)$.

Axioms

(1) All truth-functional tautologies.

(2) $A \mathrel{\square\!\!\rightarrow} A$.

(3) $(A \mathrel{\square\!\!\rightarrow} B) \ \& \ (B \mathrel{\square\!\!\rightarrow} A) \Rightarrow (A \mathrel{\square\!\!\rightarrow} C) \equiv (B \mathrel{\square\!\!\rightarrow} C)$.

(4) $((A \vee B) \mathrel{\square\!\!\rightarrow} A) \vee ((A \vee B) \mathrel{\square\!\!\rightarrow} B) \vee$
$(((A \vee B) \mathrel{\square\!\!\rightarrow} C) \equiv (A \mathrel{\square\!\!\rightarrow} C) \ \& \ (B \mathrel{\square\!\!\rightarrow} C))$.

(5) $A \mathrel{\square\!\!\rightarrow} B) \Rightarrow A \Rightarrow B$.

(6) $A \ \& \ B) \Rightarrow A \mathrel{\square\!\!\rightarrow} B$.

---

[21] Ginsberg and Smith (1987) and Winslett (1988) have also advanced theories of actions based on closest-world semantics; they have not imposed any special structure for the distance measure to reflect causal considerations. Pearl [2010c] discusses the counterfactual interpretation of do(A) or B ☺

Ital.

In Chapter 9 we will continue the analysis of causal attribution in specific events, and we will establish conditions under which the probability of correct attribution can be identified from both experimental and nonexperimental data.

8.4 A TEST FOR INSTRUMENTS

As defined in Section 8.2, our model of imperfect experiment rests on two assumptions: Z is randomized, and Z has no side effect on Y. These two assumptions imply that Z is independent of U, a condition that economists call "exogeneity" and which qualifies Z as an instrumental variable (see Sections 5.4.3 and 7.4.5) relative to the relation between X and Y. For a long time, experimental verification of whether a variable Z is exogenous or instrumental has been thought to be impossible (Imbens and Angrist 1994), since the definition involves unobservable factors (or disturbances, as they are usually called) such as those represented by U.[6] The notion of exogeneity, like that of causation itself, has been viewed as a product of subjective modeling judgment, exempt from the scrutiny of nonexperimental data.

The bounds presented in (8.14a,b) tell a different story. Despite its elusive nature, exogeneity can be given an empirical test. The test is not guaranteed to detect all violations of exogeneity, but it can (in certain circumstances) screen out very bad would-be instruments.

By insisting that each upper bound in (8.14b) be higher than the corresponding lower bound in (8.14a), we obtain the following testable constraints on the observed distribution:

$$\begin{aligned} &P(y_0, x_0 \mid z_0) + P(y_1, x_0 \mid z_1) \leq 1, \\ &P(y_0, x_1 \mid z_0) + P(y_1, x_1 \mid z_1) \leq 1, \\ &P(y_1, x_0 \mid z_0) + P(y_0, x_0 \mid z_1) \leq 1, \\ &P(y_1, x_1 \mid z_0) + P(y_0, x_1 \mid z_1) \leq 1. \end{aligned} \tag{8.21}$$

If any of these inequalities is violated, the investigator can deduce that at least one of the assumptions underlying our model is violated as well. If the assignment is carefully randomized, then any violation of these inequalities must be attributed to some direct influence that the assignment process has on subjects' responses (e.g., a traumatic experience). Alternatively, if direct effects of Z on Y can be eliminated – say, through an effective use of a placebo – then any observed violation of the inequalities can safely be attributed to spurious correlation between Z and U: namely, to assignment bias and hence loss of exogeneity. Richardson and Robins (2010) discuss the power of these tests.

The Instrumental Inequality

The inequalities in (8.21), when generalized to multivalued variables, assume the form

$$\max_x \sum_y [\max_z P(y, x \mid z)] \leq 1, \tag{8.22}$$

[6] The tests developed by economists (Wu 1973) merely compare estimates based on two or more instruments and, in case of discrepency, do not tell us objectively which estimate is incorrect.

constraints on the contingencies were too liberal. This led to a further refinement (Halpern and Pearl 2005a,b) and to the definition given below:

Definition 10.4.2 (Actual Causation) (Halpern and Pearl 2005)
$X = x$ is an actual cause of $Y = y$ in a world $U = u$ if the following three conditions hold:

AC1. $X(u) = x, Y(u) = y$

AC2. *There is a partition of V into two subsets, Z and W, with $X \subseteq Z$ and a setting x' and w of the variables in X and W, respectively, such that if $Z(u) = z^*$, then both of the following conditions hold:*

(a) $Y_{x',w} \neq y$.

(b) $Y_{x,w,z^*} = y$ *for all subsets W' of W and all subsets Z' of Z, with the setting w of W' and z^* of Z' equal to the setting of those variables in $W = w$ and $Z = z^*$, respectively.*

AC3. *W is minimal; no subset of X satisfies conditions AC1 and AC2.*

The assignment $W = w$ acts as a contingency against which $X = x$ is given the counterfactual test, as expressed in AC2(a).

AC2 (b) limits the choice of contingencies. Roughly speaking, it says that if the variables in X are reset to their original values, then $Y = y$ must hold, even under the contingency $W = w$ and even if some variables in Z are given their original values (i.e., the values in z^*).

In the case of the voting machine, if we identify $W = w$ with $V_2 = 0$, and $Z = z^*$ with $V_1 = 1$, we see that $V_i = 1$ qualifies as a cause under AC2; we no longer require that M remains invariant to the contingency $V_2 = 0$; the invariance of $Y = 1$ suffices.

This definition, though it correctly solves most problems posed in the literature (Hiddleston 2005; Hall 2007; Hitchcock 2007, 2008), still suffers from one deficiency; it must rule out certain contingencies as unreasonable. Halpern (2008) has offered a solution to this problem by appealing to the notion of "normality" in default logic (Spohn 1988; Kraus et al. 1990; Pearl 1990b); only those contingencies should be considered which are at the same level of "normality" as their counterparts in the actual world.

Halpern and Hitchcock (2010) summarize the state of the art of the structural approach to actual causation, and discuss its sensitivity to choice of variables.

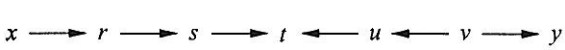

Figure 11.1 A graph containing a collider at t.

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

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

At the request of many who have had difficulties switching from algebraic to graphical thinking, I am including a gentle introduction to *d*-separation, supplementing the formal definition given in Chapter 1, pp. 16–18. ☹ See also Hayduk et al☺ 2003☺☹

Introduction

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

Unconditional Separation

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

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

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

bold / ital.

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

$$y = c_1 p + c_2 r + c_3 x + \epsilon,$$

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

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

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

Question to Author:

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

Author's Reply:

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

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

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

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

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

Now let us rotate the coordinates using the transformation

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

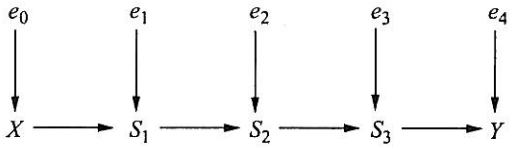

Figure 11.4 Showing the noise factors on the path from X to Y.

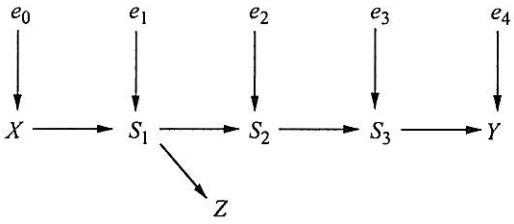

Figure 11.5 Conditioning on Z creates dependence between X and e_1, which biases the estimated effect of X on Y.

Author's Answer:

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

"|" too heavy

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

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

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

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

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

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

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

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

for some e_i, $i > 0$

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

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

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

Further Questions from This Reader:

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

1. It does not address cases such as

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

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

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

Author's Answer

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

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

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

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

which allows us to write

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

The graphical demystification of "strong ignorability" also helps explain why the probability of causation $P(Y_{x'} = y' \mid x, y)$ and, in fact, any counterfactual expression conditioned on y, would not permit such a derivation and is, in general, non-identifiable (see Chapter 9) and Shpitser and Pearl 2007

11.3.3 Alternative Proof of the Back-Door Criterion

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

Proof of the Back-Door Criterion

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

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

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

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

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

(i) $(Y \perp\!\!\!\perp T \mid X, Z)$

(ii) $(X \perp\!\!\!\perp Z \mid T)$.

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

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

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

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

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

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

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

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

Replace

See scan of pp. 346-347 from *Causality*, 2nd edition, reprinted 2010.

11.3.4 Data vs. Knowledge in Covariate Selection

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

$$V \perp\!\!\!\perp X \mid Z_1, W_1 \qquad V \perp\!\!\!\perp W_2 \qquad Z_2 \perp\!\!\!\perp W_1 \mid W_2$$
$$V \perp\!\!\!\perp Y \mid X, Z_2, W_2 \qquad Z_1 \perp\!\!\!\perp Z_2 \mid V, W_1, W_2 \qquad Z_1 \perp\!\!\!\perp W_2 \mid W_1$$
$$V \perp\!\!\!\perp W_1 \qquad X \perp\!\!\!\perp Z_2 \mid Z, W_1, W_2 \qquad X \perp\!\!\!\perp \{V, Z_2\} \mid Z_1, W_1, W_2.$$

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

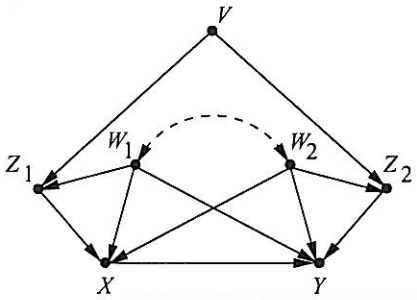

Figure 11.9 A model that is dependence-wise indistinguishable from that of Figure 11.8 (b), in which the irreducible sets $\{Z_1, W_1, W_2\}$, $\{W_1, W_2, V\}$, and $\{W_1, W_2, Z_2\}$ are admissible.

Replace

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

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

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

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

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

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

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

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

See scan of pp. 346-347 from *Causality*, 2nd edition, reprinted 2010.

Causality, 2nd edition, corrected reprint 2010, pp. 346-347.

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

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

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

The graph depicted in Figure 11.8(b) demonstrates the difficulties commonly faced by social and health scientists. Suppose our target is to estimate $P(y \mid do(x))$ given measurements on $\{X, Y, Z_1, Z_2, W_1, W_2, V\}$, but having no idea of the underlying graph structure. The conventional wisdom is to start with all available covariates $C = \{Z_1, Z_2, W_1, W_2, V\}$, and test if a proper subset of C would yield an equivalent estimand upon adjustment. Statistical methods for such reduction are described in Greenland et al. (1999b), Geng et al. (2002), and Wang et al. (2008). For example, V and Z_2 can be removed from C by successively applying conditions C_1 and C_2, thus producing an irreducible subset, $\{Z_1, W_1, W_2\}$, c-equivalent to the original covariate set C. However, this subset is inadmissible for adjustment because, like C, it does not satisfy the back-door criterion. While a theorem due to Tian et al. (1998) assures us that any c-equivalent subset of a set C can be reached from C by a step-at-a-time removal method, going through a sequence of c-equivalent subsets, the problem of covariate selection is that, lacking the graph structure, we do not know which (if any) of the many subsets of C is admissible. The next subsection discusses how external knowledge, as well as more refined analysis of the data at hand, can be brought to bear on the problem.

11.3.4 Data vs. Knowledge in Covariate Selection

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

$$V \perp\!\!\!\perp \{W_1, W_2\} \qquad X \perp\!\!\!\perp \{V, Z_2\} | \{Z_1, W_2, W_1\}$$
$$Z_1 \perp\!\!\!\perp \{W_2, Z_2\} | \{V, W_1\} \qquad V \perp\!\!\!\perp Y | \{X, Z_2, W_2, Z_1, W_1\}$$
$$Z_2 \perp\!\!\!\perp \{W_1, Z_1, X\} | \{V, W_2\} \qquad V \perp\!\!\!\perp Y | \{Z_2, W_2, Z_1, W_1\}$$

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

Causality, 2nd edition, corrected reprint 2010, pp. 346-347.

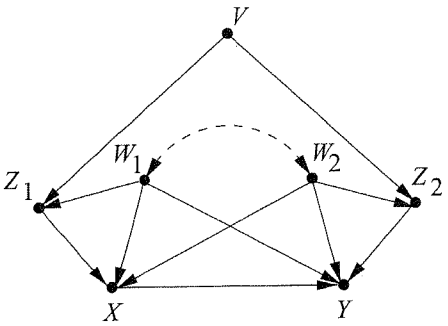

Figure 11.9 A model that is almost indistinguishable from that of Figure 11.8(b), save for advertising one additional independency $Z_1 \perp\!\!\!\perp Y | X, W_1, W_2, Z_2$. It deems three sets to be admissible (hence c-equivalent): $\{V, W_1, W_2\}$, $\{Z_1, W_1, W_2\}$, and $\{W_1, W_2, Z_2\}$, and would be rejected therefore if any pair of them fails the c-equivalence test.

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

The model depicted in Figure 11.9 is a tough contender to that of Figure 11.8(b); it satisfies all the conditional independencies implied by the latter, plus one more: $Z_1 \perp\!\!\!\perp Y | X, W_1, W_2, Z_2$, which is not easy to detect or test. Yet, contrary to Figure 11.8(b), it deems three sets $\{Z_1, W_1, W_2\}$, $\{V, W_1, W_2\}$, and $\{Z_2, W_1, W_2\}$ to be admissible, hence c-equivalent; testing for the c-equivalence of the three sets should decide between the two contesting models.

Substantive causal knowledge may provide valuable information for such decisions. For example, the model of Figure 11.9 can be ruled out if we have good reasons to believe that variable W_2 cannot have any influence on X (e.g., it may occur *later* than X), or that W_1 could not possibly have direct effect on Y.

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

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

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

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

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

The Controversy Surrounding Propensity Score

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

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

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

But why would anyone play down the cautionary note of Rosenbaum and Rubin when doing so would violate the golden rule of causal analysis: No causal claim can be established by a purely statistical method, be it propensity scores, regression, stratification, or any other distribution-based design. The answer, I believe, rests with the language that Rosenbaum and Rubin used to formulate the condition of admissibility, i.e., equation (11.11). The condition was articulated in the ~~restricted~~ cryptic language of potential-outcome, stating that the set S must render X "strongly ignorable," i.e., $\{Y_1, Y_0\} \perp\!\!\!\perp X \mid S$. As stated several times in this book, the opacity of "ignorability" is the Achilles' heel of the potential-outcome approach – no mortal can apply this condition to judge whether it holds even in simple problems, with all causal relationships correctly specified, let alone in partially specified problems that involve dozens of variables.[10]

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

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

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

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

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

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

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

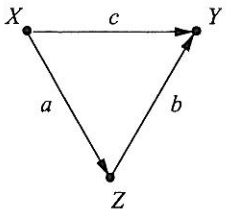

Figure 11.13 Demonstrating an indirect effect of X on Y via Z.

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

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

Me: "That's right."

Student: "Then what is being held constant when we interpret an indirect effect?"

Me: "Not sure what you mean."

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

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

Author's Reply:

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

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

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

replace ¶1

Author's Afterthought:

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

This counterfactual definition leads to the Mediation Formula (4.18), which extends path analysis to nonlinear models and enables us to estimate indirect effects for categorical variables using ordinary regression ☺

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

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

Question to Author:

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

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

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

Author's Reply:

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

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

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

11.4.4 Is the $do(x)$ Operator Universal?

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

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

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

Author's Answer:

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

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

There are three points to notice:

1. The additive manipulation *can* be represented in the $do(\ )$ framework – we merely apply the $do(\ )$ operator to the instrument I, and not to X itself. This is a different kind of manipulation that needs to be distinguished from $do(x)$ because, as is shown below, the effect on Y may be different.
2. In many cases, scientists are not satisfied with estimating the effect of the instrument on Y, but are trying hard to estimate the effect of X itself, which is often more meaningful or more transportable to other situations. (See p. 261 for discussion of the effect of "intention to treat.") ~~2~~ 1 and p@ 363 for an example of
3. Consider the nonrecursive example where LISREL fails $y = bx + e_1 + I, x = ay + e_2$, (p. 164). If we interpret "total effects" as the response of Y to a unit change of the instrument I, then LISREL's formula obtains: The effect of I on Y is $b/(1 - ab)$. However, if we adhere to the notion of "per unit change in X," we get back the do-formula: The effect of X on Y is b, not $b/(1 - ab)$, even though the manipulation is done through an instrument. In other words, we change I from 0 to 1 and observe the changes in X and in Y; if we divide the change in Y by the change in X, we get b, not $b/(1 - ab)$.

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

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

Ital.

Shpitser and Pearl (2009) provide a necessary and sufficient graphical condition for identifying the effect of the "add z to X" operator.

Bill Shipley Further Asked:

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

Author's Reply:

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

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

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

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

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

11.5 CAUSAL ANALYSIS IN LINEAR STRUCTURAL MODELS

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

***Question to Author*:**

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

Author's Reply:

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

Lemma 11.5.1 (*Graphical identification of direct effects*)
Let c stand for the path coefficient assigned to the arrow $X \rightarrow Y$ in a causal graph G. Parameter c is identified if there exists a pair (W, Z), where W is a single node in G (not excluding $W = X$), and Z is a (possibly empty) set of nodes in G, such that:

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

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

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

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

More general identification methods are reported in Brito and Pearl (2002a,b,c; 2006) and surveyed in Brito & 2010☺

11.5.2 The Causal Interpretation of Structural Coefficients

Question to Author:

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

> Page 245 of L. Hayduk, *Structural Equation Modeling with LISREL: Essentials and Advances*, 1987, [states] that a slope can be interpreted as: the magnitude of the change in y that would be predicted to accompany a unit change in x with the other variables in the equation left untouched at their original values.
>
> O.D. Duncan, *Introduction to Structural Equation Models* (1975) pages 1 and 2 are pretty clear on b as causal. More precisely, it says that "a change of one unit in x ... produces a change of b units in y" (page 2). I suspect that H. M. Blalock's book

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

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

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

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

$$q = b_1 p + d_1 i + u_1$$
$$p = b_2 q + d_2 w + u_2. \tag{11.38}$$

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

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

$$E(Q_{p_1} \mid p_0, i, w) = E(Y \mid p_0, i, w) + b_1(p_1 - p_0)$$
$$= r_p p_0 + r_i i + r_w w + b_1(p_1 - p_0), \tag{11.39}$$

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

Equation (11.39) replaces equation (7.17) on page 217. Note that the parameters of the price equation, $p = b_2 q + d_2 w + u_2$, enter (11.39) only via the regression coefficients. Thus, they need not be calculated explicitly in cases where they are estimated directly by least square.

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

Conjecture:

Any counterfactual query of the form $P(Y_x \mid e)$ is empirically identifiable ~~when Y is monotonic relative to X~~ in every constant effect model, i.e., $Y_{x_1}(u) - Y_{x_2}(u)$ is constant over u.

It is good to end on a challenging note.

11.7.2 The Meaning of Counterfactuals

Question to Author:

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

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

ADD REFERENCES TO

Bibliography

[Brito, 2010] C. Brito. Instrumental sets. In R. Dechter, H. Geffner, and J.Y. Halpern, editors, *Heuristics, Probability and Causality*, pages 295–308. College Publications, London, 2010.

[Halpern and Hitchcock, 2010] J.Y. Halpern and C. Hitchcock. Actual causation and the art of modeling. In R. Dechter, H. Geffner, and J.Y. Halpern, editors, *Heuristics, Probability and Causality*, pages 383–406. College Publications, London, 2010.

[Hayduk *et al.*, 2003] L. Hayduk, G. Cummings, R. Stratkotter, M. Nimmo, K. Grygoryev, D. Dosman, M. Gillespie, H. Pazderka-Robinson, and K. Boadu. Pearl's D-separation: One more step into causal thinking. *Structural Equation Modeling*, 10(2):289–311, 2003.

[Pearl, 2009a] J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, <http://www.bepress.com/ijb/vol6/iss2/7/>, 2009.

[Pearl, 2009b] J. Pearl. Remarks on the method of propensity scores. *Statistics in Medicine*, 28:1415–1416, 2009. <http://ftp.cs.ucla.edu/pub/stat_ser/r345-sim.pdf>.

[Pearl, 2010a] J. Pearl. The mediation formula: A guide to the assessment of causal pathways in non-linear models. Technical Report R-363, <http://ftp.cs.ucla.edu/pub/stat_ser/r363.pdf>, Department of Computer Science, University of California, Los Angeles, CA, 2010.

[Pearl, 2010b] J. Pearl. On a class of bias-amplifying variables that endanger effect estimates. In P. Grünwald and P. Spirtes, editors, *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 417–424. AUAI, Corvallis, OR, 2010.

[Pearl, 2010c] J. Pearl. Physical and metaphysical counterfactuals. Technical Report R-359, <http://ftp.cs.ucla.edu/pub/stat_ser/r359.pdf>, Department of Computer Science, University of California, Los Angeles, CA, 2010. Submitted to *Review of Symbolic Logic*.

[Richardson and Robins, 2010] T.S. Richardson and J. Robins. Analysis of the binary instrumental variable model. In R. Dechter, H. Geffner, and J.Y. Halpern, editors, *Heuristics, Probability and Causality*, pages 415–440. College Publications, London, 2010.

[Shpitser and Pearl, 2009] I. Shpitser and J. Pearl. Effects of treatment on the treated: Identification and generalization. In J. Bilmes and A. Ng, editors, *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Montreal, Quebec, 2009.

[Spirtes et al., 2010] P. Spirtes, C. Glymour, R. Scheines, and R Tillman. Automated search for causal relations: Theory and practice. In R. Dechter, H. Geffner, and J.Y. Halpern, editors, *Heuristics, Probability and Causality*, pages 467–506. College Publications, London, 2010.

[Tian and Shpitser, 2010] J. Tian and I. Shpitser. On identifying causal effects. In R. Dechter, H. Geffner, and J.Y. Halpern, editors, *Heuristics, Probability and Causality*, pages 523–540. College Publications, London, 2010.

Bowden and Turkington, 1984 R.J. Bowden and D.A. Turkington. *Instrumental Variables*. Cambridge University Press, Cambridge, England, 1984.

Breckler, 1990 S.J. Breckler. Applications of covariance structure modeling in psychology: Cause for concern? *Psychological Bulletin*, 107(2):260–273, 1990.

Breslow and Day, 1980 N.E. Breslow and N.E. Day. *Statistical Methods in Cancer Research; Vol. 1, The Analysis of Case-Control Studies*. IARC, Lyon, 1980.

Brito and Pearl, 2002a C. Brito and J Pearl. Generalized instrumental variables. In A. Darwiche and N. Friedman, editors, *Uncertainty in Artificial Intelligence, Proceedings of the Eighteenth Conference*, pages 85–93. Morgan Kaufmann, San Francisco, 2002.

Brito and Pearl, 2002b C. Brito and J Pearl. A graphical criterion for the identification of causal effects in linear models. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 533–538. AAAI Press/The MIT Press, Menlo Park, CA, 2002.

Brito and Pearl, 2002c C. Brito and J Pearl. A new identification condition for recursive models with correlated errors. *Journal of Structural Equation Modeling*, 9(4):459–474, 2002.

Brito and Pearl, 2006 C. Brito and J Pearl. Graphical condition for identification in recursive SEM. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 47–54. AUAI Press, Corvallis, OR, 2006.

Brito, 2010 >

Butler, 2002 S.F. Butler. Book review: A structural approach to the understanding of causes, effects, and judgment. *Journal of Mathematical Psychology*, 46:629–635, 2002.

Byerly, 2000 H.C. Byerly. Book reviews: Causality: Models, Reasoning, and Inference. *Choice*, 548, November 2000.

Cai and Kuroki, 2006 Z. Cai and M. Kuroki. Variance estimators for three 'probabilities of causation'. *Risk Analysis*, 25(6):1611–1620, 2006.

Cai et al., 2008 Z. Cai, M. Kuroki, J. Pearl, and J. Tian. Bounds on direct effect in the presence of confound intermediate variables. *Biometrics*, 64:695–701, 2008.

Campbell and Stanley, 1966 D.T. Campbell and J.C. Stanley. *Experimental and Quasi-Experimental Designs for Research*. R. McNally and Co., Chicago, IL, 1966.

Cartwright, 1983 N. Cartwright. *How the Laws of Physics Lie*. Clarendon Press, Oxford, 1983.

Cartwright, 1989 N. Cartwright. *Nature's Capacities and Their Measurement*. Clarendon Press, Oxford, 1989.

Cartwright, 1995a N. Cartwright. False idealisation: A philosophical threat to scientific method. *Philosophical Studies*, 77:339–352, 1995.

Cartwright, 1995b N. Cartwright. Probabilities and experiments. *Journal of Econometrics*, 67:47–59, 1995.

Cartwright, 1999 N. Cartwright. Causality: Independence and determinism. In A Gammerman, editor, *Causal Models and Intelligent Data Management*, pages 51–63. Springer-Verlag, Berlin, 1999.

Cartwright, 2007 N. Cartwright. *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press, New York, NY, 2007.

Chajewska and Halpern, 1997 U. Chajewska and J.Y. Halpern. Defining explanation in probabilistic systems. In D. Geiger and P.P. Shenoy, editors, *Uncertainty in Artificial Intelligence 13*, pages 62–71. Morgan Kaufmann, San Francisco, CA, 1997.

Chakraborty, 2001 R. Chakraborty. A rooster crow does not cause the sun to rise: Review of Causality: Models, Reasoning, and Inference. *Human Biology*, 110(4):621–624, 2001.

Chalak and White, 2006 K. Chalak and H. White. An extended class of instrumental variables for the estimation of causal effects. Technical Report Discussion Paper, UCSD, Department of Economics, July 2006.

Cheng, 1992 P.W. Cheng. Separating causal laws from causal facts: Pressing the limits of statistical relevance. *Psychology of Learning and Motivation*, 30:215–264, 1992.

Cheng, 1997 P.W. Cheng. From covariation to causation: A causal power theory. *Psychological Review*, 104(2):367–405, 1997.

Chickering and Pearl, 1997 D.M. Chickering and J. Pearl. A clinician's tool for analyzing non-compliance. *Computing Science and Statistics*, 29(2):424–431, 1997.

Greenland and Robins, 1988 S. Greenland and J.M Robins. Conceptual problems in the definition and interpretation of attributable fractions. *American Journal of Epidemiology,* 128:1185–1197, 1988.

Greenland et al., 1989 S. Greenland, H. Morgenstern, C. Poole, and J.M. Robins. Re: 'Confounding confounding'. *American Journal of Epidemiology,* 129:1086–1089, 1989.

Greenland et al., 1999a S. Greenland, J. Pearl, and J.M Robins. Causal diagrams for epidemiologic research. *Epidemiology,* 10(1):37–48, 1999.

Greenland et al., 1999b S. Greenland, J.M. Robins, and J. Pearl. Confounding and collapsibility in causal inference. *Statistical Science,* 14(1):29–46, February 1999.

Greenland, 1998 S. Greenland. Confounding. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics,* page 905–6. J. Wiley, New York, 1998.

Gursoy, 2002 K. Gursoy. Book reviews: Causality: Models, Reasoning, and Inference. *IIE Transactions,* 34:583, 2002.

Guyon et al., 2008a I. Guyon, C. Aliferis, G.F. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causation and prediction challenge. *JMLR Workshop and Conference Proceedings*, volume 3: WCCI 2008 causality challenge, Hong Kong, June 3–4 2008.

Guyon et al., 2008b I. Guyon, C. Aliferis, G.F. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov. Design and analysis of the causality pot-luck challenge. *JMLR Workshop and Conference Proceedings*, volume 5: NIPS 2008 causality workshop, Whistler, Canada, December 12 2008.

Haavelmo, 1943 T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica,* 11:1–12, 1943. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis,* Cambridge University Press, 477–490, 1995.

Haavelmo, 1944 T. Haavelmo. The probability approach in econometrics (1944)*. Supplement to *Econometrica,* 12:12–17, 26–31, 33–39, 1944. Reprinted in D.F. Hendry and M.S. Morgan (Eds.), *The Foundations of Econometric Analysis,* Cambridge University Press, New York, 440–453, 1995.

Hadlock, 2005 C.R. Hadlock. Book reviews: Causality: Models, Reasoning, and Inference. *Journal of the American Statistical Association,* 100:1095–1096, 2005.

Hall, 2004 N. Hall. Two concepts of causation. In N. Hall, J. Collins, and L.A. Paul, editors, *Causation and Counterfactuals*, Chapter 9. MIT Press, Cambridge, MA, 2004.

Halpern and Hitchcock, 2010 >

Hall, 2007 N. Hall. Structural equations and causation. *Philosophical Studies,* 132:109–136, 2007.

Halpern and Pearl, 1999 J.Y. Halpern and J. Pearl. Actual causality. Technical Report R-266, University of California Los Angeles, Cognitive Systems Lab, Los Angeles, 1999.

Halpern and Pearl, 2000 J.Y. Halpern and J. Pearl. Causes and explanations. Technical Report R-266, Cognitive Systems Laboratory, Department of Computer Science, University of California, Los Angeles, CA, March 2000. Online at ⟨www.cs.ucla.edu/~judea/⟩.

Halpern and Pearl, 2001a J.Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach—Part I: Causes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence,* pages 194–202. Morgan Kaufmann, San Francisco, CA, 2001.

Halpern and Pearl, 2001b J.Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach—Part II: Explanations. In *Proceedings of the International Joint Conference on Artificial Intelligence,* pages 27–34. Morgan Kaufmann, CA, 2001.

Halpern and Pearl, 2005a J.Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach—Part I: Causes. *British Journal of Philosophy of Science,* 56:843–887, 2005.

Halpern and Pearl, 2005b J.Y. Halpern and J. Pearl. Causes and explanations: A structural-model approach—Part II: Explanations. *British Journal of Philosophy of Science,* 56:843–887, 2005.

Halpern, 1998 J.Y. Halpern. Axiomatizing causal reasoning. In G.F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence,* pages 202–210. Morgan Kaufmann, San Francisco, CA, 1998. Also, *Journal of Artificial Intelligence Research* 12:3, 17–37, 2000.

Halpern, 2008 J.Y. Halpern. Defaults and normality in causal structures. In G. Brewka and J. Lang, editors, *Proceedings of the Eleventh International Conference on Principles of Knowledge Representation and Reasoning (KR 2008),* page 198–208. Morgan Kaufmann, San Mateo, CA, 2008.

Hauck et al., 1991 W.W. Hauck, J.M. Heuhaus, J.D. Kalbfleisch, and S. Anderson. A consequence of omitted covariates when estimating odds ratios. *Journal of Clinical Epidemiology,* 44(1):77–81, 1991.

Hausman, 1998 D.M. Hausman. *Causal Asymmetries.* Cambridge University Press, New York, 1998.

Hayduk, 1987 L.A. Hayduk. *Structural Equation Modeling with LISREL, Essentials and Advances*. Johns Hopkins University Press, Baltimore, 1987.

Heckerman and Shachter, 1995 D. Heckerman and R. Shachter. Decision-theoretic foundations for causal reasoning. *Journal of Artificial Intelligence Research*, 3:405–430, 1995.

Heckerman et al., 1994 D. Heckerman, D. Geiger, and D. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. In R. Lopez de Mantaras and D. Poole, editors, *Uncertainty in Artificial Intelligence 10*, pages 293–301. Morgan Kaufmann, San Mateo, CA, 1994.

Heckerman et al., 1995 Guest Editors: David Heckerman, Abe Mamdani, and Michael P. Wellman. Real-world applications of Bayesian networks. *Communications of the ACM*, 38(3):24–68, March 1995.

Heckerman et al., 1999 D. Heckerman, C. Meek, and G.F. Cooper. A Bayesian approach to causal discovery. In C. Glymour and G. Cooper, editors, *Computation, Causation, and Discovery*, The MIT Press, Cambridge, MA, 143–167, 1999.

Heckman and Honoré, 1990 J.J. Heckman and B.E. Honoré. The empirical content of the Roy model. *Econometrica*, 58:1121–1149, 1990.

Heckman and Robb, 1986 J.J. Heckman and R.R. Robb. Alternative methods for solving the problem of selection bias in evaluating the impact of treatments on outcomes. In H. Wainer, editor, *Drawing Inference From Self Selected Samples*, pages 63–107. Springer-Verlag, New York, NY, 1986.

Heckman and Vytlacil, 1999 J.J. Heckman and E.J. Vytlacil. Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the National Academy of Sciences, USA*, 96(8):4730–4734, April 1999.

Heckman and Vytlacil, 2007 J.J. Heckman and E.J. Vytlacil. *Handbook of Econometrics*, volume 6B, Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation, pages 4779–4874. Elsevier B.V., 2007.

Heckman et al., 1998 J.J. Heckman, H. Ichimura, and P. Todd. Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65:261–294, 1998.

Heckman, 1992 J.J. Heckman. Randomization and social policy evaluation. In C. Manski and I. Garfinkle, editors, *Evaluations: Welfare and Training Programs*, pages 201–230. Harvard University Press, Cambridge, MA, 1992.

Heckman, 1996 J.J. Heckman. Comment on 'Identification of causal effects using instrumental variables'. *Journal of the American Statistical Association*, 91(434):459–462, June 1996.

Heckman, 2000 J.J. Heckman. Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics*, 115(1):45–97, 2000.

Heckman, 2003 J.J. Heckman. Conditioning causality and policy analysis. *Journal of Econometrics*, 112(1):73–78, 2003.

Heckman, 2005 J.J. Heckman. The scientific model of causality. *Sociological Methodology*, 35:1–97, 2005.

Heise, 1975 D.R. Heise. *Causal Analysis*. John Wiley and Sons, New York, 1975.

Hendry and Morgan, 1995 D.F. Hendry and M.S. Morgan. *The Foundations of Econometric Analysis*. Cambridge University Press, Cambridge, 1995.

Hendry, 1995 David F. Hendry. *Dynamic Econometrics*. Oxford University Press, New York, 1995.

Hennekens and Buring, 1987 C.H. Hennekens and J.E. Buring. *Epidemiology in Medicine*. Little, Brown, Boston, 1987.

Hernán et al., 2002 M.A. Hernán, S. Hernández-Díaz, M.M. Werler, and A.A. Mitchell. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *American Journal of Epidemiology*, 155(2):176–184, 2002.

Hernán et al., 2004 M.A. Hernán, S. Hernández-Díaz, and J.M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, 2004.

Hernández-Díaz et al., 2006 S. Hernández-Díaz, E.F. Schisterman, and Hernán M.A. The birth weight "paradox" uncovered? *American Journal of Epidemiology*, 164(11):1115–1120, 2006.

Hesslow, 1976 G. Hesslow. Discussion: Two notes on the probabilistic approach to causality. *Philosophy of Science*, 43:290–292, 1976.

Hiddleston, 2005 E. Hiddleston. Causal powers. *British Journal for Philosophy of Science*, 56:27–59, 2005.

Pearl, 2005b J. Pearl. Influence diagrams – historical and personal perspectives. *Decision Analysis*, 2(4):232–234, 2005.

Pearl, 2008 J. Pearl. The mathematics of causal relations. Technical Report TR-338, http://ftp.cs.ucla.edu/pub/stat_ser/r338.pdf, Department of Computer Science, University of California, Los Angeles, CA, 2008. Presented at the American Psychopathological Association (APPA) Annual Meeting, NYC, March 6–8, 2008.

Pearl, 2009 J. Pearl. Remarks on the method of propensity scores. *Statistics in Medicine*, 28:1415–1416, 2009. See also http://ftp.cs.ucla.edu/put/stat_ser/r345-sim.pdf.

Pearson et al., 1899 K. Pearson, A. Lee, and L. Bramley-Moore. Genetic (reproductive) selection: Inheritance of fertility in man. *Philosophical Transactions of the Royal Society A*, 73:534–539, 1899.

Peikes et al., 2008 D.N. Peikes, L. Moreno, and S.M. Orzol. Propensity scores matching: A note of caution for evaluators of social programs. *The American Statistician*, 62(3):222–231, 2008.

Peng and Reggia, 1986 Y. Peng and J.A. Reggia. Plausibility of diagnostic hypotheses. In *Proceedings of the Fifth National Conference on AI (AAAI-86)*, pages 140–145, Philadelphia, 1986.

Petersen et al., 2006 M.L. Petersen, S.E. Sinisi, and M.J. van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(3):276–284, 2006.

Poole, 1985 D. Poole. On the comparison of theories: Preferring the most specific explanations. In *Proceedings of the Ninth International Conference on Artificial Intelligence (IJCAI-85)*, pages 144–147, Los Angeles, CA, 1985.

Popper, 1959 K.R. Popper. *The Logic of Scientific Discovery*. Basic Books, New York, 1959.

Pratt and Schlaifer, 1988 J.W. Pratt and R. Schlaifer. On the interpretation and observation of laws. *Journal of Econometrics*, 39:23–52, 1988.

Price, 1991 H. Price. Agency and probabilistic causality. *British Journal for the Philosophy of Science*, 42:157–176, 1991.

Price, 1996 H. Price. *Time's arrow and Archimedes' point: New directions for the physics of time*. Oxford University Press, New York, 1996.

Program, 1984 Lipid Research Clinic Program. The Lipid Research Clinics Coronary Primary Prevention Trial results, parts I and II. *Journal of the American Medical Association*, 251(3):351–374, January 1984.

Rebane and Pearl, 1987 G. Rebane and J. Pearl. The recovery of causal poly-trees from statistical data. In *Proceedings of the Third Workshop on Uncertainty in AI*, pages 222–228, Seattle, WA, 1987.

Reichenbach, 1956 H. Reichenbach. *The Direction of Time*. University of California Press, Berkeley, 1956.

Reiter, 1987 R. Reiter. A theory of diagnosis from first principles. *Artificial Intelligence*, 32(1):57–95, 1987.

Richard, 1980 J.F. Richard. Models with several regimes and changes in exogeneity. *Review of Economic Studies*, 47:1–20, 1980.

Richardson, 1996 T. Richardson. A discovery algorithm for directed cyclic graphs. In E. Horvitz and F. Jensen, editors, *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 454–461. Morgan Kaufmann, San Francisco, CA, 1996.

Rigdon, 2002 E.E. Rigdon. New books in review: Causality: Models, Reasoning, and Inference and Causation, Prediction, and Search. *Journal of Marketing Research*, XXXIX: 137–140, 2002.

Robert and Casella, 1999 C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Verlag, New York, NY, 1999.

Robertson, 1997 D.W. Robertson. The common sense of cause in fact. *Texas Law Review*, 75(7): 1765–1800, 1997.

Robins and Greenland, 1989 J.M. Robins and S. Greenland. The probability of causation under a stochastic model for individual risk. *Biometrics*, 45:1125–1138, 1989.

Robins and Greenland, 1991 J.M. Robins and S. Greenland. Estimability and estimation of expected years of life lost due to a hazardous exposure. *Statistics in Medicine*, 10:79–93, 1991.

Robins and Greenland, 1992 J.M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2): 143–155, 1992.

Shipley, 1997 B. Shipley. An inferential test for structural equation models based on directed acyclic graphs and its nonparametric equivalents. Technical report, Department of Biology, University of Sherbrooke, Canada, 1997. Also in *Structural Equation Modelling*, 7:206–218, 2000.

Shipley, 2000a B. Shipley. Book reviews: Causality: Models, Reasoning, and Inference. *Structural Equation Modeling*, 7(4):637–639, 2000.

Shipley, 2000b B. Shipley. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference*. Cambridge University Press, New York, 2000.

Shoham, 1988 Y. Shoham. *Reasoning About Change: Time and Causation from the Standpoint of Artificial Intelligence*. MIT Press, Cambridge, MA, 1988.

Shpitser and Pearl, 2006a I. Shpitser and J Pearl. Identification of conditional interventional distributions. In R. Dechter and T.S. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 437–444. AUAI Press, Corvallis, OR, 2006.

Shpitser and Pearl, 2006b I. Shpitser and J Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1219–1226. AAAI Press, Menlo Park, CA, 2006.

Shpitser and Pearl, 2007 I. Shpitser and J Pearl. What counterfactuals can be tested. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, pages 352–359. AUAI Press, Vancouver, BC Canada, 2007. Also, *Journal of Machine Learning Research*, 9:1941–1979, 2008.

Shpitser and Pearl, 2008 I. Shpitser and J Pearl. Dormant independence. In *Proceedings of the Twenty-Third Conference on Artificial Intelligence*, pages 1081–1087. AAAI Press, Menlo Park, CA, 2008.

Shpitser and Pearl, 2009

Shrier, 2009 I. Shrier. Letter to the Editor: Propensity scores. *Statistics in Medicine*, 28:1317–1318, 2009.

Simon and Rescher, 1966 H.A. Simon and N. Rescher. Cause and counterfactual. *Philosophy and Science*, 33:323–340, 1966.

Simon, 1953 H.A. Simon. Causal ordering and identifiability. In Wm. C. Hood and T.C. Koopmans, editors, *Studies in Econometric Method*, pages 49–74. Wiley and Sons, Inc., New York, NY, 1953.

Simpson, 1951 E.H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13:238–241, 1951.

Sims, 1977 C.A. Sims. Exogeneity and causal ordering in macroeconomic models. In *New Methods in Business Cycle Research: Proceedings from a Conference, November 1975*, pages 23–43. Federal Reserve Bank, Minneapolis, 1977.

Singh and Valtorta, 1995 M. Singh and M. Valtorta. Construction of Bayesian network structures from data – a brief survey and an efficient algorithm. *International Journal of Approximate Reasoning*, 12(2): 111–131, 1995.

Sjölander, 2009 A. Sjölander. Letter to the Editor: Propensity scores and M-structures. *Statistics in Medicine*, 28:1416–1423, 2009.

Skyrms, 1980 B. Skyrms. *Causal Necessity*. Yale University Press, New Haven, 1980.

Smith and Todd, 2005 J. Smith and P. Todd. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125:305–353, 2005.

Sobel, 1990 M.E. Sobel. Effect analysis and causation in linear structural equation models. *Psychometrika*, 55(3):495–515, 1990.

Sober and Barrett, 1992 E. Sober and M. Barrett. Conjunctive forks and temporally asymmetric inference. *Australian Journal of Philosophy*, 70:1–23, 1992.

Sober, 1985 E. Sober. Two concepts of cause. In P. Asquith and P. Kitcher, editors, *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, volume II, pages 405–424. Philosophy of Science Association, East Lansing, MI, 1985.

Sommer et al., 1986 A. Sommer, I. Tarwotjo, E. Djunaedi, K. P. West, A. A. Loeden, R. Tilden, and L. Mele. Impact of vitamin A supplementation on childhood mortality: A randomized controlled community trial. *The Lancet*, 327:1169–1173, 1986.

Sosa and Tooley, 1993 E. Sosa and M. Tooley (Eds.). *Causation*. Oxford readings in Philosophy. Oxford University Press, Oxford, 1993.

Spiegelhalter et al., 1993 D.J. Spiegelhalter, S.L. Lauritzen, P.A. Dawid, and R.G. Cowell. Bayesian analysis in expert systems (with discussion). *Statistical Science*, 8:219–283, 1993.

Spirtes etal., 2010

Suppes, 1970 P. Suppes. *A Probabilistic Theory of Causality.* North-Holland Publishing Co., Amsterdam, 1970.

Suppes, 1988 P. Suppes. Probabilistic causality in space and time. In B. Skyrms and W.L. Harper, editors, *Causation, Chance, and Credence.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

Swanson and Granger, 1997 N.R. Swanson and C.W.J. Granger. Impulse response functions based on a causal approach to residual orthogonalization in vector autoregressions. *Journal of the American Statistical Association*, 92:357–367, 1997.

Swanson, 2002 N.R. Swanson. Book reviews: Causality: Models, Reasoning, and Inference. *Journal of Economic Literature,* XL:925–926, 2002.

Tian and Pearl, 2000 J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28:287–313, 2000.

Tian and Pearl, 2001a J. Tian and J. Pearl. Causal discovery from changes. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence,* pages 512–521. Morgan Kaufmann, San Francisco, CA, 2001.

Tian and Pearl, 2001b J. Tian and J. Pearl. Causal discovery from changes: A Bayesian approach. Technical Report R-285, Computer Science Department, UCLA, February 2001.

Tian and Pearl, 2002a J. Tian and J. Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence,* pages 567–573. AAAI Press/The MIT Press, Menlo Park, CA, 2002.

Tian and Pearl, 2002b J. Tian and J Pearl. On the testable implications of causal models with hidden variables. In A. Darwiche and N. Friedman, editors, *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 519–527. Morgan Kaufmann, San Francisco, CA, 2002.

Tian and Shpitser, 2010

Tian et al., 1998 J. Tian, A. Paz, and J. Pearl. Finding minimal separating sets. Technical Report R-254, University of California, Los Angeles, CA, 1998.

Tian et al., 2006 J. Tian, C. Kang, and J. Pearl. A characterization of interventional distributions in semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence,* pages 1239–1244. AAAI Press, Menlo Park, CA, 2006.

Tversky and Kahneman, 1980 A. Tversky and D. Kahneman. Causal schemas in judgments under uncertainty. In M. Fishbein, editor, *Progress in Social Psychology,* pages 49–72. Lawrence Erlbaum, Hillsdale, NJ, 1980.

VanderWeele and Robins, 2007 T.J. VanderWeele and J.M. Robins. Four types of effect modification: A classification based on directed acyclic graphs. *Epidemiology,* 18(5):561–568, 2007.

Verma and Pearl, 1988 T. Verma and J. Pearl. Causal networks: Semantics and expressiveness. In *Proceedings of the Fourth Workshop on Uncertainty in Artificial Intelligence,* pages 352–359, Mountain View, CA, 1988. Also in R. Shachter, T.S. Levitt, and L.N. Kanal (Eds.), *Uncertainty in AI 4,* Elesevier Science Publishers, 69–76, 1990.

Verma and Pearl, 1990 T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference on Uncertainty in Artificial Intelligence,* pages 220–227, Cambridge, MA, July 1990. Also in P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer (Eds.), *Uncertainty in Artificial Intelligence 6,* Elsevier Science Publishers, B.V, 255–268, 1991.

Verma and Pearl, 1992 T. Verma and J. Pearl. An algorithm for deciding if a set of observed independencies has a causal explanation. In D. Dubois, M.P. Wellman, B. D'Ambrosio, and P. Smets, editors, *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence,* pages 323–330. Morgan Kaufmann, Stanford, CA, 1992.

Verma, 1993 T.S. Verma. Graphical aspects of causal models. Technical Report R-191, UCLA, Computer Science Department, 1993.

Wainer, 1989 H. Wainer. Eelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions. *Journal of Educational Statistics,* 14:121–140, 1989.

Wang et al., 2009 X. Wang, Z. Geng, H. Chen, and X. Xie. Detecting multiple confounders. *Journal of Statistical Planning and Inference,* 139: 1073–1081, 2009.

Wasserman, 2004 L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference.* Springer Science+Business Media, Inc., New York, NY, 2004.

instrumental inequality, 274

Mediation
see indirect effects
Mediation Formula, 106, 132, 358

Neyman-Rubin model
see potential outcome framework

Ital.