Applied Regression Analysis and Generalized Linear Models, Third Edition

Supplement: Chapter 26 (Draft)

Causal Inferences From Observational Data: Directed Acyclic Graphs and Potential Outcomes

John Fox

Last Modified: 2023-03-31

Contents

Chapter 26

Causal Inferences From Observational Data: Directed Acyclic Graphs[1]

Among Sir R. A. Fisher's many seminal contributions to statistics was the fundamental technique of randomization in experimental design.[2] In the 1950s, the great British statistician, himself a smoker and consultant to the tobacco industry, notoriously maintained that there was no good evidence that tobacco smoking causes lung cancer. One of Fisher's arguments was that it is possible that the observed association of lung cancer with tobacco smoking is due to a common genetic cause. (Fisher made important contributions to genetics as well as to statistics.) Even without the benefit of hindsight, Fisher's position seems perverse (see, e.g., Stolley, 1991), but the more general difficulty of inferring causation from observational data is a real and continuing problem in epidemiology—witness, for example, the more recent controversy over the efficacy and safety of hormone-replacement therapy for post-menopausal women (e.g., Mayo Clinic, 2022).

Observational data are no less prominent in the social sciences than in epidemiology, and the issues that social scientists address, for example in the area of public policy, are probably even more difficult to disentangle: Does capital punishment decrease homicide? Does the availability of legal abortion decrease violent crime? Do social-welfare programs raise the standard of living of the poor?

[1]Some material in the chapter is adapted with permission from Fox (2008). The treatment of directed acyclic graphs here profited greatly from discussions with Georges Monette of York University in Toronto, who, along with Michael Friendly and several other members of the York University Statistical Consulting Service study group, also commented helpfully on drafts of the chapter.

[2]Several of Fisher's other contributions figure prominently in this text, including the theory of estimation, the method of maximum likelihood, the notion of degrees of freedom, and the analysis of variance for linear models.

The difficulty of drawing causal conclusions from observational data has been understood for a long time, as has the basic strategy of controlling statistically for potentially confounding variables, for example by multiple-regression analysis. Because it is always possible that a confounding prior cause has not been identified and observed, this strategy has a fundamental limitation not shared by randomized comparative experiments.[3] Although statisticians have intermittently addressed this issue, I think that it's fair to say that most statisticians prefer to construe regression equations as predictive rather than causal—for example, statisticians tend to call explanatory variables "predictors." In contrast, researchers who apply regression models to observational data, in the social sciences and more generally, typically want to give the models causal interpretations, even if they pay lip service to the oft-repeated dictum that "correlation doesn't imply causation."

I briefly discussed causal inference from observational data previously in the text (primarily in Sections 1.2, 6.3, 9.7, and 9.8). The purpose of the current chapter is to deepen the treatment of causal inferences from observational data by describing an approach to the topic that has achieved recent prominence: the use of directed acyclic graphs to represent causal relationships among variables. I'll also explain how directed acyclic graphs can elucidate the analysis of missing data, and the relationship between causal graphs and the potential-outcomes (or counterfactual) approach to understanding statistical causation.

The treatment of causal graphs in this chapter is relatively abbreviated, and I refer the interested reader to the recommended readings at the end of the chapter. In particular, the simple examples in the chapter are meant to clarify essential ideas concerning causal inference from observational data, but are insufficiently rich to develop the subject in detail.

26.1 Graphs

In the sense that it is used here, a *graph* is a labeled set of *nodes* (points), $\{A, B, C, \ldots\}$, connected in pairs by *edges* (visualized as line segments), for example, $\{A—B, B—C, \ \ldots\}$. Some examples of graphs appear in Figure 26.1. When a graph is represented visually, as here, the layout of the nodes on the page and the lengths of the edges connecting the nodes are irrelevant: Only the connections among the nodes created by the edges are important. Thus, panels (a) and (b) are different visual representations of the same graph.

The graph in panels (a) and (b) of Figure 26.1 is an *undirected graph*, while those in panels (c), (d), and (e) are *directed graphs*, with the edges represented by single-headed arrows. The node at the tail of each arrow in a directed graph is the *parent node* and that at the arrow head is the *child node*.

A *path* through a graph between two nodes is a sequence of consecutive edges connecting the nodes. All of the graphs in Figure 26.1 are *completely connected* in that there are paths between all pairs of nodes. A *directed path* (in a directed

[3] As explained in Section 1.2, however, statistical evidence of causation isn't completely unambiguous even in randomized experiments.

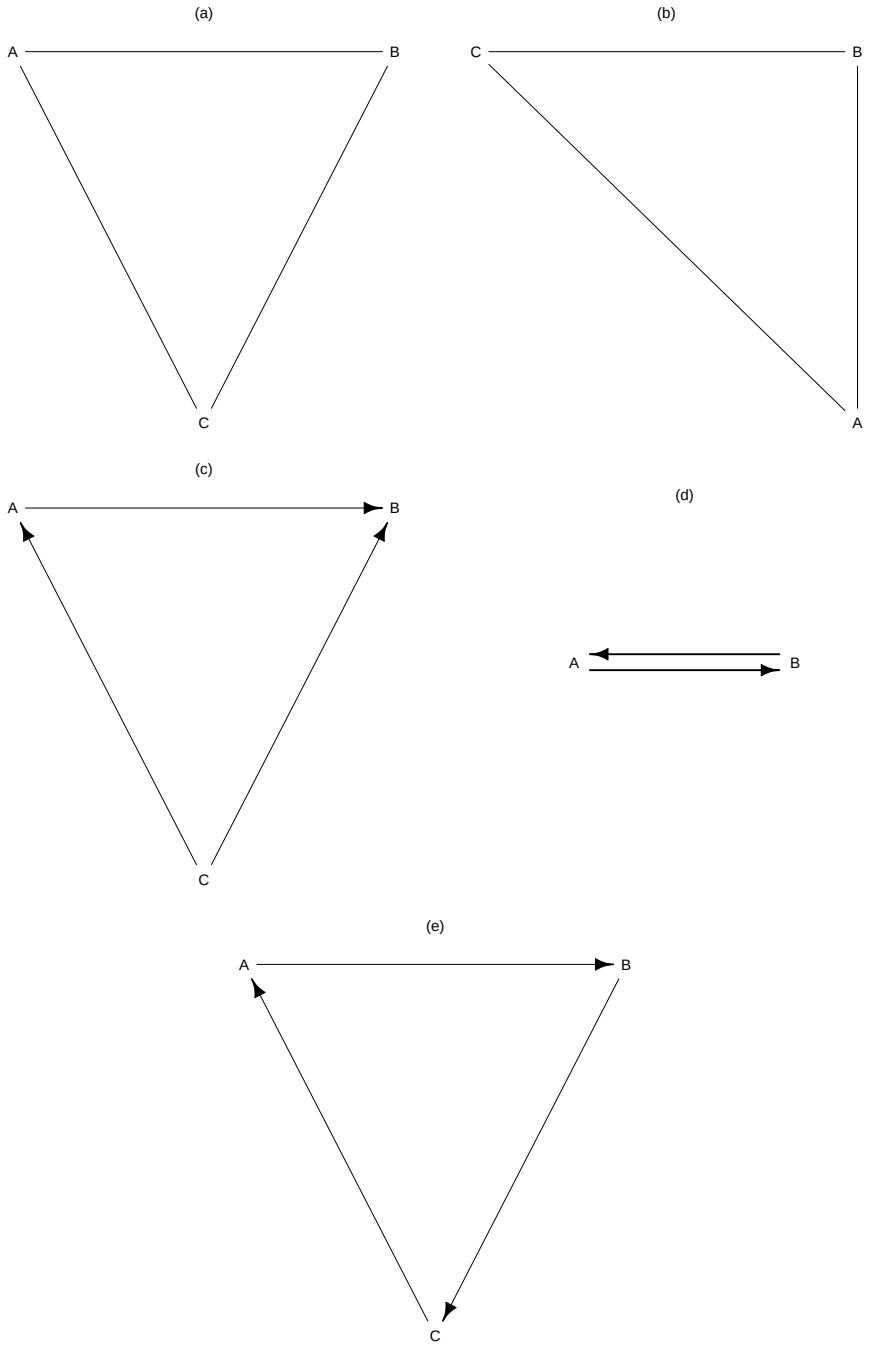

Figure 26.1 Examples of graphs: (a) and (b) an undirected graph; (c) a directed acyclic graph; (d) and (e) directed cyclic graphs.

graph) is a path all of whose arrows point in the same direction, in the sense that each intermediate node in the path is both a child and a parent. Thus, the graph in Figure 26.1(c) has directed paths from node A to B ($A \longrightarrow B$), and node C to B ($C \longrightarrow B$ and $C \longrightarrow A \longrightarrow B$), but not from node A to C or from node B to C. The initial node of a directed path is an *ancestor* of the terminal node; and the terminal node is a *descendant* of the initial node.

A directed graph is *acyclic* if it has no reciprocal paths (as in Figure 26.1(d)) or loops (as in Figure 26.1(e)). Thus, for example, Figure 26.1(c) represents a *directed acyclic graph* (or *DAG*). DAGs interpreted as causal graphs are the focus of this chapter.

> A graph is a labeled set of nodes (points) connected by edges (line segments). Graphs may be undirected or directed, in which case the edges are represented as single-headed arrows. The node at the head of each arrow in a directed graph is the parent node and that at the tail is the child node.
> A path through a graph between two nodes is a sequence of consecutive edges connecting the nodes, and a directed path is a path all of whose arrows point in the same direction. The initial node of a directed path is an ancestor of the terminal node, which is a descendant of the initial node. A directed graph is acyclic if it has no reciprocal paths or loops.

26.1.1 Causal Directed Acyclic Graphs

Graphs have a variety of applications, including, for example, to social-network analysis. The use of graphs to depict causal relationships among variables—where variables are represented as nodes in the graph, connected by arrows representing direct causal relationships—dates to the work of the geneticist Sewall Wright on *path analysis* (Wright, 1921). As originally formulated, Wright's path analysis was closely tied to linear least-squares regression, and more recent generalizations to *structural-equation models* (e.g., Duncan, 1975; Bollen, 1989) are also associated with parametric regression models. I'll have a bit more to say about path analysis later in the chapter, in connection with an example drawn from Blau and Duncan (1967),[4] who (along with Duncan, 1966) introduced Wright's method to sociologists. More recent work on causal directed acyclic graphs, most prominently by the computer scientist and philosopher Judea Pearl (Pearl, 2000, 2009; Pearl et al., 2016; Pearl and Mackenzie, 2018), is nonparametric, in that it doesn't require that the regression models corresponding to a graph have particular functional forms.[5]

[4] See Section 26.7.

[5] Indeed, it isn't even necessary to estimate causal relationships in a DAG by regression, though that is the strategy that I'll pursue in this chapter.

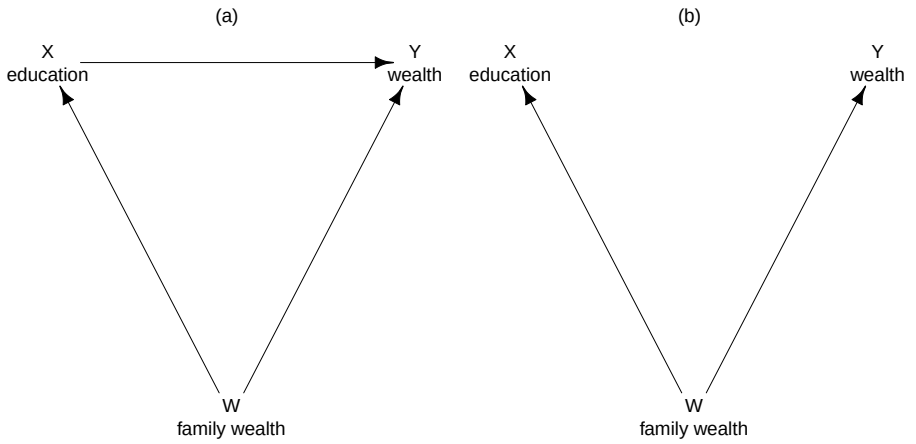

Figure 26.2 DAGs with W confounding the effect of X on Y. In (a) the confounding is partial; in (b) it is complete.

> Directed acyclic graphs (or DAGs) represent causal relationships among variables, where the variables are the nodes of the graph, and arrows connecting the variables represent direct effects, with the direct cause (or parent node) at the tail of an arrow and the effect (or child node) at the tip. To say that DAGs are acyclic implies that causation is unidirectional, with no reciprocal arrows or feedback loops.

Two examples of causal DAGs appear in Figure 26.2. The DAG in Figure 26.2(a) is familiar: It is essentially the same as Figure 1.1 (on page 7)[6] and Figure 6.2(a) (on page 121).[7] I use the following conventions (the first of which isn't universal) in drawing the DAGs in Figure 26.2:

- Observed variables are denoted by upper-case Latin letters, typically from near the end of the alphabet, with the causal explanatory variable of interest X and the response Y.
- The direct effect of one variable on another is represented by a directed arrow from the cause to the effect. Moreover, because it is acyclic, a DAG cannot have reciprocal direct effects, as in Figure 26.3(a), or a closed

[6] Page references in this chapter may be to pages within the chapter or to pages in the printed text (i.e., Chapters 1 through 24). References should generally be clear from the context; here, e.g., Figure 1.1 is from Chapter 1 of the text.

[7] There is this difference, however: I described Figure 1.1 as "an informal 'causal model,'" while DAGs are *formal* causal models with an associated statistical theory that I will partially develop in this chapter.

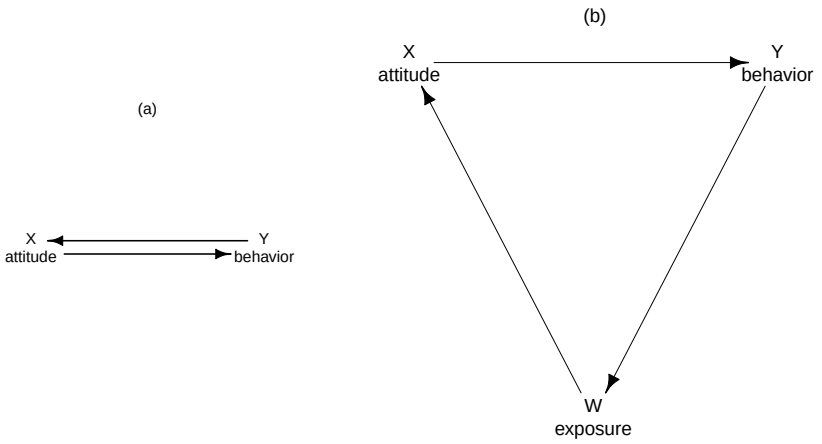

Figure 26.3 Graphs that contain causal cycles and hence are not DAGs: (a) reciprocal causation; (b) a causal loop.

causal loop, as in Figure 26.3(b). Variables that are causes but never effects (i.e., variables to which no arrows point, and which, hence, have no parents in the DAG), such as W in the DAGs in Figures 26.2(a) and (b), are termed *exogenous*, while variables that are effects and possibly, but not necessarily, causes (i.e., variables to which arrows point and which, hence, have parents), such as X and Y, are termed *endogenous*.

In Figure 26.3, and more generally in this chapter, I name the variables in causal graphs, not to provide serious applications but simply to make it easier to think concretely about the graphs.[8] Figure 26.3(a), for example, specifies that individuals' attitude (say to members of another race) affects their behavior, and that their behavior affects their attitude. Figure 26.3(b) is similar, except that the effect of behavior on attitude is mediated by exposure. It is therefore possible to imagine causal graphs that are not DAGs.[9]

26.2 Confounders and Mediators

In Figure 26.2, I imagine that we're interested in estimating the effect of individuals' education (X) on their wealth (Y), and that the wealth of their families of origin (W) is a common prior cause of education and wealth. Figures 26.2(a) and (b) illustrate *confounding*, where the *confounder* W is causally prior to both X and Y.

[8] For an application of DAGs, see Section 26.7.

[9] It is also sometimes possible to estimate such causal systems, as in nonrecursive structural-equation models, a topic not discussed in this text (but see the references at the end of the chapter).

A confounder creates a *back-door path* connecting X and Y, a kind of undirected path that generates spurious (i.e., non-causal) association between these two variables. I'll elaborate the idea of back-door paths presently.

In Figure 26.2(a), confounding is partial, in the sense that there is still a true causal source of association between Y and X, represented by the directed arrow linking the two. Suppose that the direct effect of X on Y is positive: An increase in X tends to produce an increase in Y.[10] Depending on the signs of the effects of W on X and Y, the overall relationship between X and Y could be positive, negative, or absent. The key—and familiar—point is that the overall statistical association between Y and X is not the same as the effect of X on Y due to the presence of the confounder W. Figure 26.2(b) is similar, except that confounding is complete and there is *no* effect of X and Y.

The pattern of the path connecting X and Y in Figure 26.2(b) is called a causal fork, and its qualitative influence on the association of X and Y is essentially unchanged if other variables appear in the branches of the fork; for example, $X \longleftarrow U \longleftarrow W \longrightarrow V \longrightarrow Y$ generates spurious association between X and Y.

We already know how to deal with an observed confounder like W in Figure 26.2: We can control statistically for W to estimate the partial relationship between Y and X. If all of the partial relationships are linear, we could perform a multiple linear regression of Y on X and W, taking the coefficient of X as an estimate of the effect of X on Y. In the case of Figure 26.2(b), we'd expect an estimate close to 0 (and a population regression coefficient of precisely 0). More generally—that is, whether or not the regressions are linear—the DAG in Figure 26.2(b) implies that Y is statistically independent of X given W, which is often symbolized as $(Y \perp\!\!\!\perp X) \,|\, W$.

> A confounder creates a back-door path connecting a cause X and effect Y, which in turn generates spurious (i.e., non-causal) association between these two variables. We can estimate the effect of X on Y by controlling statistically for the confounder. The path $X \longleftarrow W \longrightarrow Y$ is called a causal fork.

[10] In the general context of DAGs, effects don't necessarily have a consistent sign (i.e., aren't necessarily monotone) and certainly need not be linear. For example, the effect of X on Y might be quadratic. It is, however, simpler to consider, at least initially, monotone or even linear effects.

We can go further: A DAG such as Figure 26.2(a), with arrows $X \longrightarrow Y$ and $W \longrightarrow Y$ implies that the probability distribution of Y given X and W depends jointly on the values x of X and w of W, but the function specifying this dependence is quite general and need not be additive in X and W; that is, $p[y|(X = x, W = w)] = f_Y(y; x, w)$ (where p is a probability for discrete Y or probability density for continuous Y), and so, for example, X and W might interact in and arbitrary manner in determining Y. This is true even if we focus on the conditional mean of Y, that is, the regression of Y on X and W, where $\mu \equiv E[Y|(X = x, W = w)] = f_\mu(x, w)$, and where $f_\mu(\cdot)$ isn't constrained parametrically by the DAG to take a particular functional form.

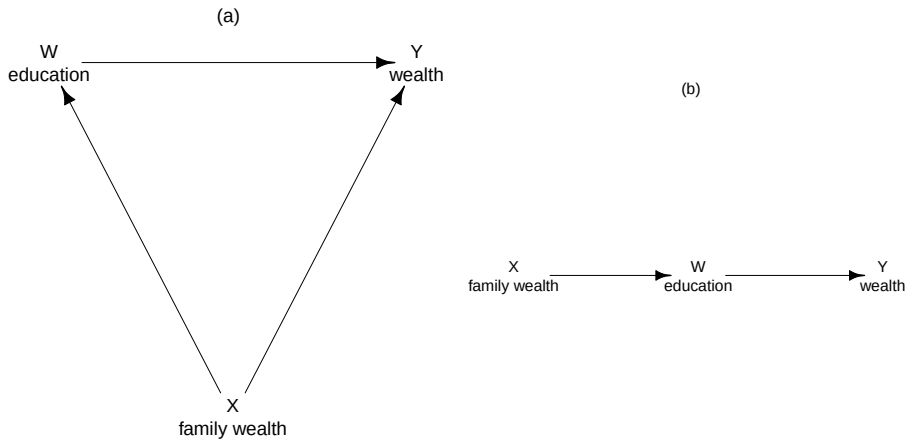

Figure 26.4 DAGs with W mediating the effect of X on Y. In (a) mediation is partial; in (b) it is complete.

The DAG in Figure 26.4(a) is also familiar, representing causal *mediation*, where W (education) intervenes between X (family wealth) and Y (wealth). Indeed, Figure 26.4(a) is really just a rearrangement of Figure 26.2(a), with W and X switching roles: Recall that I reserve X to represent the cause of interest. In Figure 26.4(a), mediation is partial, and there is a direct effect of X on Y, in addition to the indirect effect through W; in Figure 26.4(b), in contrast, there is *no* direct effect, and mediation is complete.

The pattern of the directed path linking X to Y in Figure 26.4(b) is called a causal *chain*, and, as was true for forks, the qualitative influence of a causal chain is unaffected by interpolating additional variables between X and Y. For example, in $X \longrightarrow U \longrightarrow W \longrightarrow V \longrightarrow Y$, X and Y are associated because of the indirect effect of X on Y through W (and U and V).

An important point (made previously in Sections 1.2, 6.3, and 9.7) is that the DAGs in Figures 26.2 and 26.4 have identical observable implications, although widely divergent interpretations. Both Figure 26.2(b) and Figure 26.4(b), for example, imply that $(Y \perp\!\!\!\perp X) \,|\, W$, but in the first case we would explain away the association of Y and X as *spurious* (i.e., non-causal), while in the second case we would elaborate the *mechanism* through which X affects Y. Indeed, if, as I assume, our object is to estimate the effect of X on Y, we *should not* control statistically for the mediator W in the DAGs in Figures 26.4(a) and (b).

A mediator is a variable that intervenes between a cause X and effect Y. We should not control statistically for a mediator if we want to estimate the effect of X on Y. The path $X \longrightarrow W \longrightarrow Y$ is called a causal chain. Confounders and mediators can't be distinguished solely on statistical grounds.

26.3 Closing Back-Door Paths

I have explained that a single confounder, as in Figures 26.2(a) and (b) (on page 5), opens a back-door path linking the potential cause of interest X and effect Y, producing a spurious source of statistical association between X and Y. Back-door paths can be more complex, as illustrated in Figure 26.5(a), where a back-door path linking X and Y passes through the two prior variables V and W, and in Figure 26.5(b), where there are two back-door paths, through V, W and U, and through W and U. In all of the examples in this chapter, involving relatively simple DAGs, it is easy to discern the back-door paths. The more general identification of back-door paths and other features of DAGs can be more complex, however.[11] Nevertheless, all back-door paths end in arrows pointing directly to X and Y, traverse causally prior variables, and reverse direction once at an eventual confounder.

In the elementary cases illustrated in Figure 26.2, we can obtain unbiased estimators of the effect of X on Y (that is, no effect in the case of Figure 26.2(b)) by controlling statistically for the confounder W. Similarly, in the simple cases illustrated in Figure 26.4, we know that we *should not* control for the mediator W, for to do so would eliminate the indirect effect of X on Y through W. More generally, to estimate the effect of X on Y we must *close* (or *block*) all of the back-door paths connecting the two variables in the DAG. That, however, does not in general require that we control for *all* variables in the DAG that are causally prior to X and Y. Any back-door path can be blocked by controlling statistically for at least one causally prior variable along it, and blocking all back-door paths connecting X and Y serves to identify the effect of X on Y. We should never, however, control for a variable that intervenes causally between X and Y, even if it's on a backdoor path (see an example below). This result (or conditions equivalent to it) is called the *back-door criterion*.

Consider, for example, the DAG in Figure 26.5(a): To block the back-door path in this DAG, we can control either for V, or for W, or for both. Any of these choices would produce an unbiased estimator of the effect of X on Y. Which choice is best? Suppose, again for simplicity, that all of the regressions are linear and that we use multiple least-squares regression. Because regressing Y on X and V, on X and W, and on X, V, and W all yield unbiased estimators

[11] For additional details, consult the recommended readings at the end of the chapter. Moreover, the rules for analyzing DAGs are well understood and are instantiated in statistical software; see, in particular daggity (Textor et al., 2017), which I used in preparing the material in this chapter.

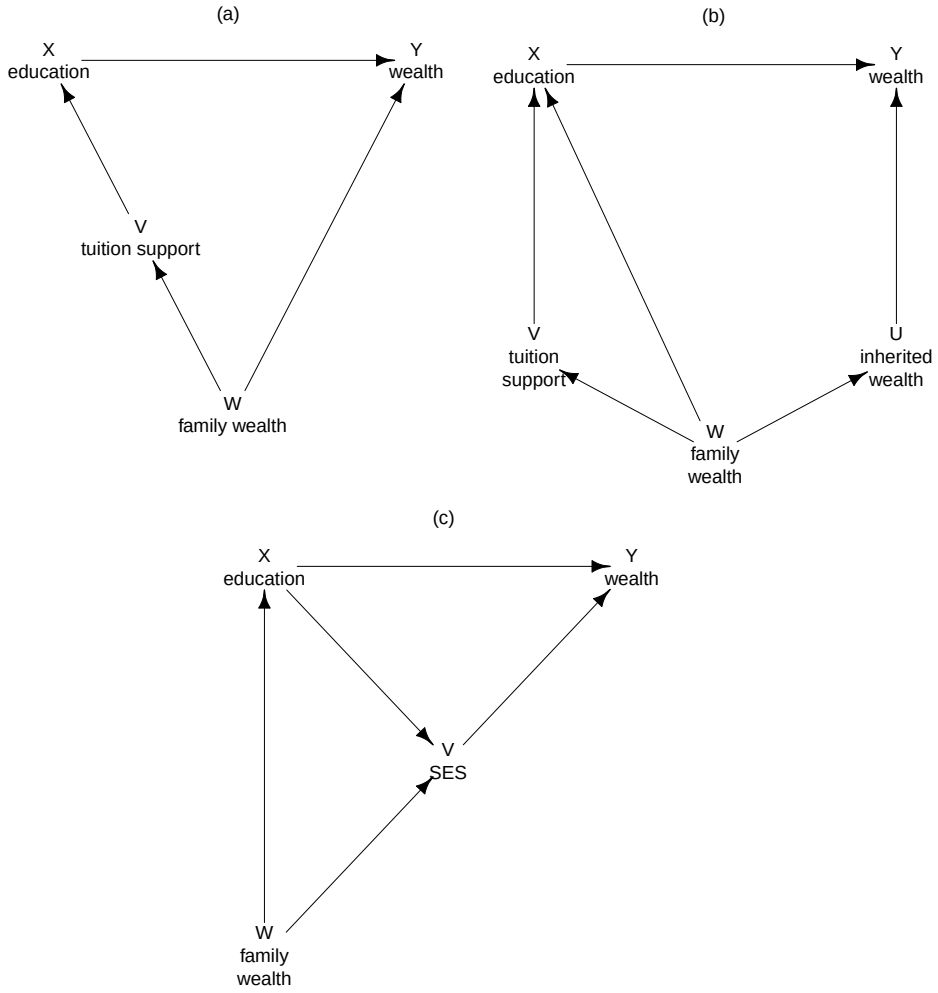

Figure 26.5 DAGs with back-door paths: (a) a single back-door path between X and Y through V and W; (b) two back-door paths, through V, W, and U, and through W and U. (c) a variable, V, that's simultaneously on a fork and a causal chain.

of the effect, say β_X, of X on Y, we would prefer the regression that produces the most efficient—that is, the lowest-variance—estimator of β_X.

Recall (adapting Equation 6.3 on page 113) the formula for the sampling variance of the least-squares regression coefficient B_X (estimating β_X):

$$V(B_X) = \frac{1}{1 - R_X^2} \times \frac{\sigma_\varepsilon^2}{\sum (X_i - \overline{X})^2} \tag{26.1}$$

where R_X^2 is the squared multiple correlation for the regression of X on the other explanatory variables in the main regression, and σ_ε^2 is the error variance for the regression of Y on X and the other explanatory variables.[12]

There are, therefore, two relevant factors that affect the sampling variance of B_X: (1) the degree of collinearity between X and the other explanatory variables in the regression model; and (2) the size of the error variance.[13] On both grounds, we should control for W and ignore V:

- Because there is no direct arrow between V and Y, the population regression coefficient for V in the multiple regression of Y on X, V, and W is 0; to include the irrelevant regressor V could only decrease the precision of estimation, to the degree that V is correlated with W and X.

- Moreover, we should prefer to regress Y on X and W rather than to regress Y on X and V, because V is causally closer than W to X, and hence will be more correlated than W with X (thus producing a larger R_X^2).

- Finally, V is more remote causally than W from Y (its effect on Y is transmitted solely through W) and so including V rather than W in the regression would increase the error variance.

[12] As mentioned in Section 13.1, another way to think about the sampling variance of B_X is in terms of the added-variable plot (AV plot) for X, where the variable on the horizontal axis is the residual, say $\widetilde{X}$, from the regression of X on the other regressors and the variable on the vertical axis, , say $\widetilde{Y}$, is the residual for the regression of Y on the other regressors (also see Section 11.6.1). Then

$$\widehat{V}(B_X) = \frac{S_E^2}{\sum \widetilde{X}_i^2}$$

where S_E^2 is the residual variance from the original regression, also representing the spread round the regression line in the AV plot. To obtain a precise estimate of β_X, we want the residual variation to be as small as possible and the conditional variation in X to be as large as possible, as long as B_X is an unbiased estimator of β_X. See Exercise 26.3(b) for an application of AV plots to these ideas.

[13] The third factor in Equation 26.1, $\sum (X_i - \overline{X})^2 = (n-1)S_X^2$, doesn't change from one regression to another, and so isn't relevant to the current context.

To obtain an unbiased estimator of the effect of X on Y we must close (i.e., block) all of the back-door paths connecting the two variables in the DAG (the back-door criterion). Closing all back-door paths does not in general require that we control for all variables in the DAG that are causally prior to X and Y. It's generally advantageous to control for the antecedent variable or variables that are sufficient to close all back-door paths and that, in doing so, produce the most precise estimate of the effect of X on Y. When we have a choice, we therefore prefer to control for antecedent variables that are close to Y and remote from X.

It is important to understand, however, that these conclusions assume that the causal structure of the DAG is correct. If, for example, were there an additional direct arrow (not shown) between V and Y, then it might well be advantageous to control for V, which would block *both* back-door paths—the path through V alone *and* the path through V and W. Additionally, in this case, it may be (but is not necessarily) best to control for V *and* W, inasmuch as adding W to the regression would both decrease the error variance, thus increasing the precision of estimation of β_X, and increase collinearity, thus decreasing the precision of estimation. The DAG doesn't imply which of these opposing consequences is larger.

Even though it isn't necessary, and may be suboptimal, to control for all prior variables in a DAG such as Figure 26.5(a), it is in a sense safest to do so. Suppose, for example, that this DAG incorrectly omits direct arrows from V to Y and from W to X (again not shown). In this case, we should control for both V and W to close all back-door paths so as to obtain an unbiased estimator of the effect of X on Y.

Figure 26.5(b) elaborates Figure 26.5(a) by adding a direct path from W to Z and a variable U that wholly mediates the effect of W on Y. There are additional back-door paths linking X to Y in this DAG, but no new principles, and I leave its analysis as an exercise for the reader.[14]

Figure 26.5(c) includes the variable V, which is simultaneously on a fork linking X and Y ($X \longleftarrow W \longrightarrow V \longrightarrow Y$) and a causal chain between the two focal variables ($X \longrightarrow V \longrightarrow Y$). In a case such as this, we should *not* control for the mediator V, for doing so would block the indirect effect of X on Y through V. We can instead close the back-door path by controlling for W, which is antecedent to both X and Y.

26.4 Colliders and Descendants

The phenomena highlighted by DAGs are specific instances of a more general, and familiar, idea: that partial and marginal relationships can differ. The role of a DAG is to identify which partial (or possibly marginal) relationships can

[14] See Exercise 26.1.

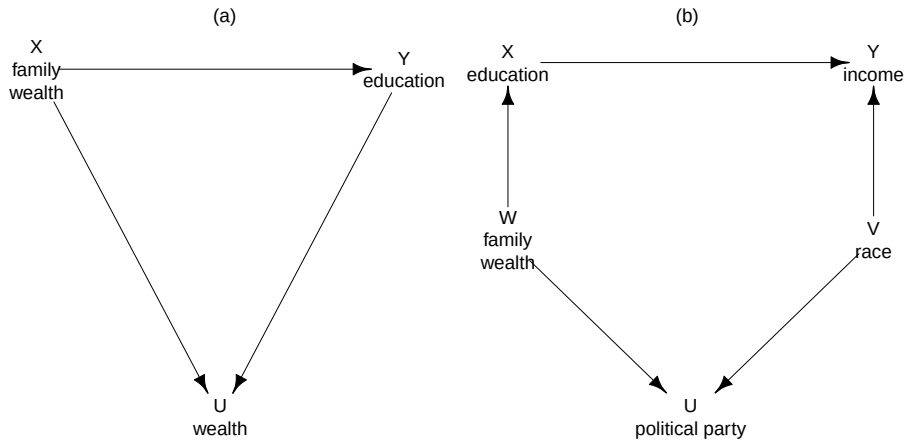

Figure 26.6 DAGs with colliders: (a) a consequent collider; (b) a possibly antecedent collider.

reasonably be given causal interpretations. Our examination of DAGs has thus far suggested that to estimate the effect of X on Y we may want to control for (some) causally prior variables to close back-door paths that induce spurious sources of association between X and Y, and that we should generally not control for variables that intervene causally between X and Y. I believe that these conclusions are reasonably intuitive.

26.4.1 Colliders

So-called *colliders* are an additional, and less intuitive, class of variables for which we should *not* control in estimating the effect of X on Y. Colliders are variables that block sources of non-causal association between X and Y, and controlling for a collider opens a non-causal path between the two focal variables.

Two examples of colliders appear in Figure 26.6. The variable U is a collider in both panels (a) and (b):[15] In each case, there is a path in the DAG linking X and Y through U, with two arrows pointing *towards* U (hence the term "collider"). That both arrows point towards U blocks association from flowing through this path, and controlling for U in effect *unblocks* the path, biasing the estimator of the effect of X on Y.

As I said, I believe that the idea that a collider blocks a non-causal path is less intuitive than the notion that a confounder creates a back-door path. As a consequence, using the DAG in Figure 26.6(b), I generated simulated data to illustrate this phenomenon. In particular, I drew $n = 1000$ independent observations on the variables W, V, X, Y, and U according to the following

[15] The story told by Figure 26.6(b) is not entirely credible, in that there almost surely should be an arrow between race (V) and family wealth (W). I omitted this arrow to create a simpler example, but see Exercise 26.2.

Table 26.1 Several Regressions for Y in the Simulated Data With Collider U

Coefficient	Model (Explanatory Variables)					
(SE(B))	1. (X)	2. (X, U)	3. (X, V)	4. (X, U, V)	5. (X, U, W)	6. (X, U, V, W)
B_X	1.033	0.166	0.998	0.956	0.930	0.930
	(0.032)	(0.017)	(0.008)	(0.023)	(0.042)	(0.031)
B_U		0.888		0.044	0.946	0.024
		(0.012)		(0.023)	(0.011)	(0.033)
B_V			0.993	0.949		0.968
			(0.008)	(0.024)		(0.033)
B_W					−0.877	0.039
					(0.045)	(0.046)
S_E	1.035	0.407	0.256	0.255	0.346	0.255

scheme:

$$
\begin{aligned}
W &\sim N(0, 1) \\
V &\sim N(0, 1) \\
X &\sim N(W,\ 0.25^2) \\
Y &\sim N(X + V,\ 0.25^2) \\
U &\sim N(W + V,\ 0.25^2)
\end{aligned}
$$

By construction, the effect of X on Y is the population regression coefficient $\beta_X = 1$. The DAG in Figure 26.6(b) suggests that we can obtain an unbiased estimator of β_X by regressing Y on X alone. Because V is a direct cause of Y and not a cause of X, it is also advantageous to include the covariate V in the regression to reduce the size of the error variance and hence increase the precision of estimation of β_X. In contrast, including W, a direct cause of X but not of Y, in the regression would serve only to decrease the precision of estimation. Finally, including the collider U in a regression along with X should bias the estimator of β_X.

Table 26.1 summarizes the results of fitting several least-squares regressions to the simulated data, and the coefficient of X in each model is graphed in Figure 26.7,[16] along with a 95% confidence interval.[17]

- Model 1 includes X alone and, as expected, provides an estimate B_X close to $\beta_X = 1$.
- Also as expected, adding the collider U to Model 2 biases the coefficient of X, which is now much smaller.

[16] I'm grateful to Michael Friendly of York University for suggesting that I draw this graph.

[17] I've omitted the model that includes X and W: See Exercise 26.3, which slightly elaborates this example.

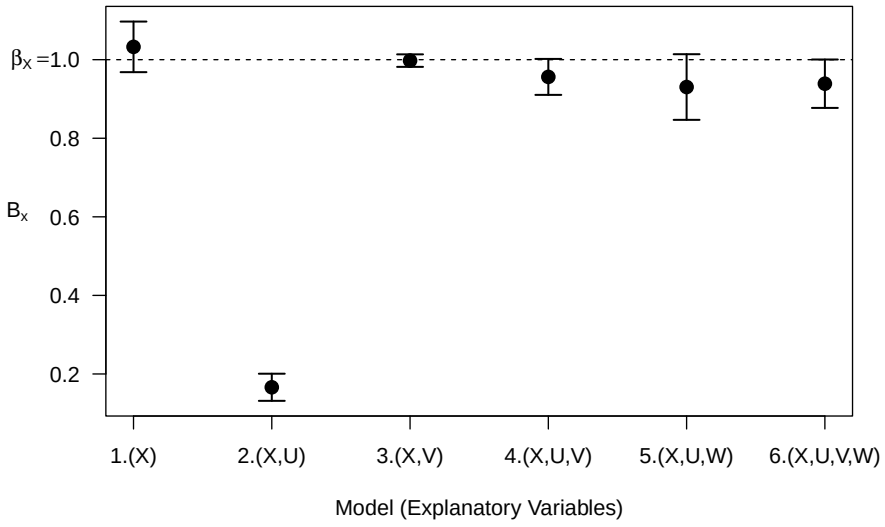

Figure 26.7 Coefficients B_X of X and 95% confidence intervals for the models in Table 26.1, estimating $\beta_X = 1$.

- In contrast, including the covariate V along with X in Model 3 yields an unbiased estimator of β_X that has a much smaller standard error (i.e., is much more precise) than the estimate from Model 1. The regression standard error S_E for this model is close to the error standard deviation $\sigma_Y = 0.25$ in the population regression equation that generated the data for Y, which, recall, included X and V.
- Models 4, 5, and 6, include the focal explanatory variable X, the collider U, and one or both of the other variables V and W on the path with the collider. Because they block the spurious path opened by including the collider, these models yield unbiased estimators of the effect of X, but none provides an estimate of β_X as precise as that in Model 2, which includes only X and V.

> Colliders are variables that block sources of non-causal association between X and Y. Controlling for a collider opens a non-causal path between the two focal variables, biasing the estimated effect of X and Y.

26.4.2 Descendants

A descendant of a variable V in a causal DAG is a variable U affected directly or indirectly by V. Conversely, in this circumstance, V is an ancestor of U. When the focal causal variables X and Y are common ancestors of a variable U (and there is a directed path from X to U that doesn't go through Y), then U is a collider, as illustrated in the DAG in Figure 26.8, and previously in the DAG in Figure 26.6(a) (page 13).

Figure 26.8 also shows three other direct descendants: V of X, T of Y, and W of the collider U (and, hence, indirectly of X and Y). We know that the consequence of controlling for a collider is to bias the estimator of the effect of X on Y. Are there consequences of controlling for the other kinds of descendants?

- Controlling for a descendant of X, such as V in Figure 26.8, doesn't bias the estimator of the effect of X on Y, but it impairs its efficiency by decreasing conditional variation in X.
- Controlling for a descendant of a collider, such as the descendant W of U, biases the estimate of the effect of X on Y, because W serves as a partial proxy for U. That is, we can think of W as an imperfect measure of the collider U with a measurement-error component.
- Controlling for a descendant of Y, such as T, biases the estimator of the effect of X for a similar reason: T is an version of Y with an increased error component, and it's not sensible to control for the response.

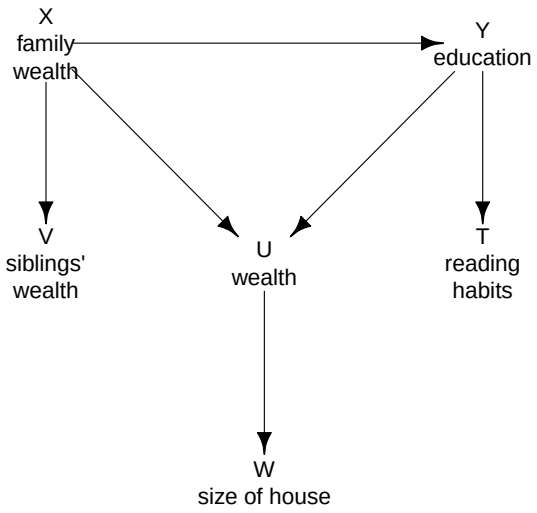

Figure 26.8 A DAG with four different kinds of descendants: a collider U, which is a common descendant of X and Y; a descendant V of X; a descendant T of Y; and a descendant W of the collider U.

The general lesson to be drawn from these observations is that we should not in general control for consequences of X and Y. Although controlling for consequences of the response ("symptoms") can improve prediction of Y, often dramatically, the goals of causal inference and prediction are different and frequently contradictory.

The fictional story told by the substantive variable names in Figure 26.8 isn't entirely credible. In particular, we might well think of size of house as a *cause*, rather than a *consequence*, of wealth, reversing the direction of the arrow between U and W. Suppose we make this change to the DAG and then control for W, regressing Y on X and W to assess the effect of X on Y. In the modified DAG (which isn't shown), $W \perp\!\!\!\perp X$ and $W \perp\!\!\!\perp Y$, so controlling for W has no expected effect on the regression.[18]

> Controlling for descendants of both X and Y or of Y alone biases the estimate of the effect of X on Y, while controlling for descendants of X alone makes the estimate of the effect of X less precise. We should therefore avoid controlling for descendants of the focal causal variables.

[18] For an illustration of this point, and more generally of the phenomena concerning controlling for descendants in this section, see Exercise 26.4.

26.4.3 Selection Bias and Colliders

DAGs with colliders illuminate how self-selection can produce biased estimates of causal effects, and, conversely, the process of self-selection can illuminate why controlling statistically for a collider induces bias. To illustrate, I'll adapt an example from Section 9.8 on instrumental-variables estimation.[19]

Suppose that to estimate the effect of private-school versus public-school attendance on the academic performance of economically disadvantaged students, a researcher randomly assigns a group of volunteer students to two experimental conditions: The students in the treatment condition receive vouchers to attend well resourced private schools, while those in the control condition do not. Unlike in Section 9.8, I'll assume here that compliance is perfect, so that all students who receive vouchers attend private schools, while all students who do not receive vouchers attend public schools. On the other hand, not all students participate in the evaluative phase of the study, and self-selection of participation (denoted $S = 1$ for participants and $S = 0$ for non-participants) is related both to type of school attended ($X = 1$ for those who attended private schools, $X = 0$ for those who attended public schools) and academic achievement at the end of the study (assessed by a standardized exam, Y).

I generated simulated data for this experiment according to the following scheme:

- Half of $n = 5000$ students were picked at random to attend private schools (for whom $X_i = 1$), and the other half attended public schools ($X_i = 0$).
- The response was generated so that private-school students had slightly better test scores on average than their public-school counterparts, $Y_i \sim N(0.1 \times X_i, 1)$.
- Participation in the study was a function of both X and Y (as depicted in the DAG in Figure 26.9), with the probability of selection $\Pr(S)$ following the logistic-regression equation

$$\Pr(S_i = 1) = \frac{1}{1 + \exp[-(2.5X_i + Y_i)]}$$

  Selection is therefore strongly dependent on both type of school attended and students' test scores. For example, 50 percent of public-school students and 89 percent of private-school students participated in the study.

Table 26.2 shows group mean academic achievement ($\overline{Y}_j, j = 0, 1$), the difference in means ($\overline{Y}_1 - \overline{Y}_0$), the standard error of the difference, and the p-value from an independent-samples t-test of the difference for the full data set and for two subsets of the data: the students who participated in the study, and

[19] Also see the discussion of DAGs and instrumental variables in Section 26.6.1. Although the material on instrumental variables is in starred parts of the text, the example in the current section doesn't depend on an understanding of instrumental variables.

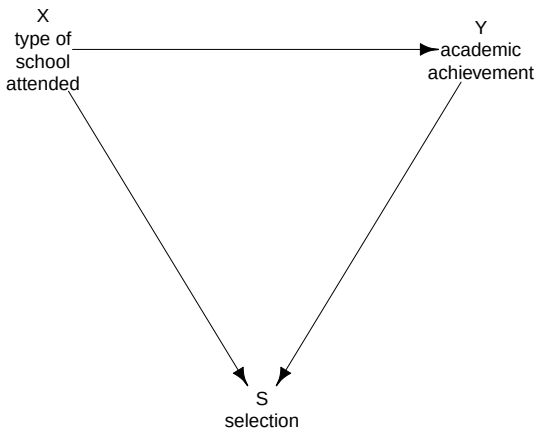

Figure 26.9 The selection variable S is a collider, affected by both X and Y. Looking only at self-selected subjects, for whom $S = 1$, controls for the collider and consequently biases the estimate of the effect of X on Y.

Table 26.2 Mean Academic Achievement by Type of School Attended, for the Full Data Set, for Participants, and for Non-participants

	Full Data Set	Participants	Non-participants
Public-School Mean $\overline{Y}_0$	0.015	0.403	−0.379
Private-School Mean $\overline{Y}_1$	0.108	0.207	−0.721
Difference in Means $\overline{Y}_1 - \overline{Y}_0$	0.094	−0.196	−0.342
SE(Difference)	0.020	0.033	0.062
p-Value for t-Test	$< .001$	$\ll .0001$	$\ll .0001$

those who didn't participate.[20] Of course, in a real data set we would only be able to observe the participants—an advantage of using simulated data for this example.

As expected, for the full data set, the estimates $\overline{Y}_0$ for public-school students and $\overline{Y}_1$ for private-school students are close to the population means of $\mu_0 = 0$ and $\mu_1 = 0.1$, respectively, and the difference in means $\overline{Y}_1 - \overline{Y}_0 = 0.094$ is associated with a small p-value. For both subsets of students, however, the sign of the difference is reversed: $\overline{Y}_1 - \overline{Y}_0 = -0.196$ for participants and -0.342 for non-participants. Both of these subset differences are associated with very small p-values. This is, therefore, an instance of Simpson's paradox: Not only are the partial relationships of academic achievement to type of school controlling for participation different from the marginal relationship, but the partial and marginal relationships differ in direction.

[20]The t-test is equivalent to a test for the coefficient of X in a dummy-variable regression of Y on X, where the coefficient of X is just the difference in group means.

How are we to understand these results? As mentioned, participation is strongly related to both type of school attended and academic achievement. Almost all private-school students participated in the study, and so both relatively low- and high-achieving private-school students were included. For public-school students, participation was driven more by academic achievement, and so relatively low-achieving public-school students tended to be excluded. Similarly, the private-school students who *did not* participate were predominantly the lowest achievers among their peers. The differential selection of public- and private-school students accounts for the reversal in sign of the relationship between academic achievement and type of school attended within the two participation subsets.

Selection bias in estimating the effect of X on Y can be understood as controlling for a collider, in which we examine the partial relationship between X and Y within one category of the collider—that is, for self-selected subjects.

26.5 Statistical Independencies and d-Separation

Two variables, say X and Y, in a DAG are said to be *d-separated* if (1) there are no causal forks or chains connecting the variables, and (2) if any other paths connecting the variables are blocked by colliders. Variables that are d-separated are marginally (i.e., unconditionally) statistically independent, $X \perp\!\!\!\perp Y$. Variables that aren't d-separated are *d-connected* and marginally dependent.[21]

Now imagine that we examine the relationship between X and Y holding a set of other variables $\mathbf{W} = \{W_1, W_2, \ldots, W_k\}$ constant (i.e., conditioning on the values of the variables in $\mathbf{W}$). Then X and Y are *conditionally d-separated* given $\mathbf{W}$ if (1) one or more variables in $\mathbf{W}$ block every fork and chain connecting X and Y, and (2) no path connecting X and Y is unblocked by a collider or the descendant of a collider in $\mathbf{W}$, unless that path is also blocked by one or more other variables in $\mathbf{W}$. Variables that are conditionally d-separated are conditionally independent, $(X \perp\!\!\!\perp Y)|\mathbf{W}$. The conditioning variables $\mathbf{W}$ need not be unique; that is, there may be more than one set of conditioning variables that satisfy these criteria.

Consider, the slightly more complicated DAG in Figure 26.10:

- There are no unblocked arrows connecting C_1 and C_2, and so these two variables are d-separated, and $C_1 \perp\!\!\!\perp C_2$.[22]
- There are two back-door paths connecting X and Y: $X \longleftarrow C_1 \longrightarrow Y$ and $X \longleftarrow C_2 \longrightarrow C_3 \longrightarrow Y$. Similarly, there are two causal chains

[21] The "d" in d-separated and d-connected represents "*d*irectional."
[22] Also see Exercise 26.5(a).

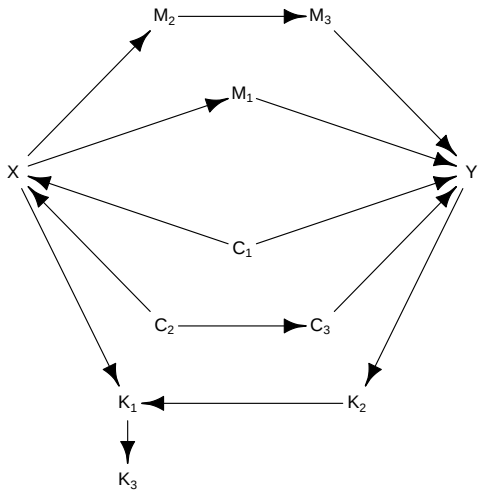

Figure 26.10 A DAG with mediators, confounders, and colliders.

connecting X and Y: $X \longrightarrow M_1 \longrightarrow Y$ and $X \longrightarrow M_2 \longrightarrow M_3 \longrightarrow Y$, and so, for example, controlling for C_1, C_2, M_1, and M_2 is sufficient to d-separate X and Y conditionally. That is, $(Y \perp\!\!\!\perp X) | \{C_1, C_2, M_1, M_2\}$.[23]

- As mentioned, C_1 and C_2 are d-separated. From the point of view of these two variables X is a collider, and so if we control for X, C_1 and C_2 are conditionally d-connected; it's almost surely the case that C_1 and C_2 are *not* independent controlling for X.[24]

26.6 DAGs With Unobserved Variables

In my opinion, perhaps the most important contribution of DAGs to data analysis is their ability to assist us in reasoning about potential confounders that aren't present in our observed data. Recall (from Sections 6.3 and 9.7) that a key assumption in interpreting a regression causally is that the explanatory variables in the regression are independent of (or, in linear regression, at least uncorrelated with) the regression error—where the error represents the omitted causes of the response. It is always possible in observational data that this assumption fails in an unknown manner, but it is also possible that the assumption fails in a *known* manner. That is, we may understand what (some of) the omitted causes of Y are, and this understanding, cast in the form of a DAG,

[23] Also see Exercise 26.5(b).
[24] Also see Exercise 26.5(d).

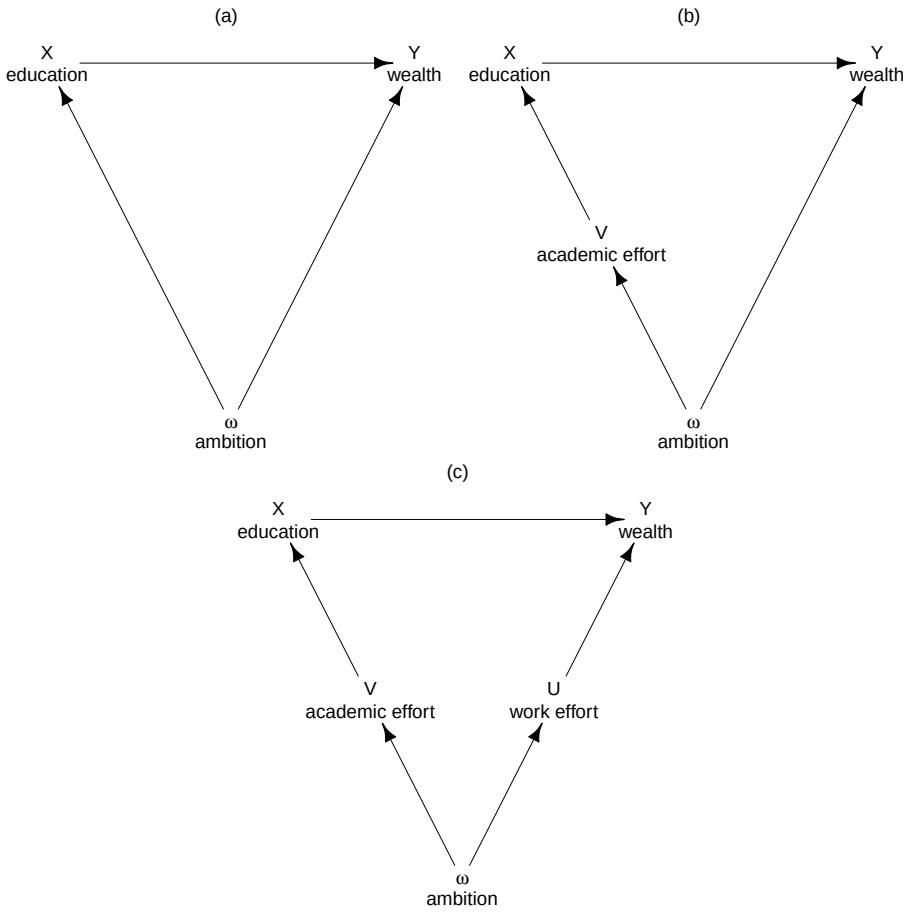

Figure 26.11 DAGs with an unobserved variable: (a) an unobserved confounder ω of the relationship between X and Y; (b) the effect of ω on X mediated by the observed variable V; (c) the effect of ω on X mediated by V and the effect of ω on Y mediated by U.

may help us decide whether the situation is hopeless or whether we can proceed in a principled manner to estimate the effect of an observed cause X on Y.

Figure 26.11 displays three simple DAGs, each including an *unobserved variable* (also called a *latent variable*) ω—adopting the general convention in the text of using Greek letters to denote unobserved quantities, including unobserved random variables. Figure 26.11(a) is very much like Figure 26.2(a) (page 5), except that the observed confounder W is replaced by the unobserved confounder ω; I assume here that the variable ω (ambition) either isn't present in our data or that we don't know how to measure it.

As I have explained, to estimate the effect of X on Y in Figure 26.2(a) (page 5), we should control statistically for W, regressing Y on X and W. We can't do that for the DAG in Figure 26.11(a) because ω isn't available to us. This is an unhappy situation, but at least the DAG makes it clear *why* we can't legitimately estimate the effect of X on Y, and may suggest what we need to do—perhaps by collecting additional data, but not necessarily observing ω—to obtain an unbiased estimator.

Examine, for example, the DAG in Figure 26.11(b). The observed variable V wholly mediates the effect of the confounder ω on X, and thus controlling for V blocks the back-door path in the DAG and is sufficient to obtain an unbiased estimator of the effect of X on Y. If we believe the DAG, then we don't have to observe ω to discount its confounding influence.

A similar, if slightly more elaborate, example appears in Figure 26.11(c), where V wholly mediates the effect of ω on X and U wholly mediates the effect of ω on Y. To block the back-door path between X and Y through ω, we can control for V, or for U, or for both V and U. Any of these choices would produce an unbiased estimator of the effect of X on Y, but the analysis in the preceding section suggests that we would obtain the most precise estimator by controlling for U alone.

> An important contribution of DAGs is that they can help us to understand the role of unobserved (latent) variables in causal inference. In certain cases, we may be able to close back-door paths that include unobserved confounders by controlling for observed variables along these paths.

26.6.1 Instrumental Variables*

Recall that if we wish to construe a least-squares regression causally (see Sections 6.3 and 9.7), the explanatory variables in the regression must be independent of (or, in a linear regression, at least uncorrelated with) the regression error. As described in Section 9.8, instrumental-variables estimation is a method for estimating a linear regression consistently when one or more explanatory variables are related to the error. DAGs can assist in the identification of potential

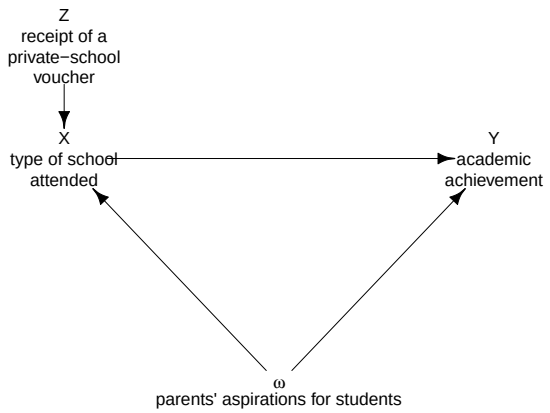

Figure 26.12 DAG with an instrumental variable Z to estimate the effect of X on Y when there's an unobserved confounder ω.

instrumental variables in applications. Although I won't pursue the point here, the use of instrumental variables to obtain unbiased estimators of effects is more general than linear regression.

I previously employed the following example to illustrate instrumental-variables estimation:[25] Imagine an experiment to determine the effect of private-school versus public-school attendance on students' academic achievement, in which students are assigned at random to receive vouchers to attend private schools. If all students who receive vouchers use them to attend private schools and all who don't receive vouchers attend public schools, then type of school attended would be unrelated, at least in expectation, to all potential confounding causes of academic achievement. Suppose, however, that compliance is less than complete, and that some students without vouchers are sent by their families to private schools, and some who receive vouchers nevertheless attend public schools. Then it's possible, and even likely, that self-selection induces a relationship between type of school attended and the omitted causes of academic achievement.

This situation is represented in the DAG in Figure 26.12, where one potential latent confounder is identified: parents' aspirations for their children, The DAG makes it clear that parental aspirations (ω) is an unobserved spurious source of association between type of school attended (X) and academic achievement (Y); moreover, receipt of a voucher (Z) is a cause of, and hence related to, type of school attended, while, because vouchers are randomly assigned, receipt of a voucher is unrelated to parental aspirations—satisfying the two criteria for an instrumental variable.

[25] Earlier in this chapter (Section 26.4.3), I adapted this example to illustrate the use of DAGs to explain selection bias.

DAGs may help us to identify instrumental variables, as causally prior variables that affect X but not Y directly and that are reasonably construed as unrelated to the omitted (i.e. latent) causes of Y. Such instrumental variables make it possible to estimate the effect of X on Y even when X is related to the regression error in Y, such as when there are unobserved confounders creating back-door paths that can't be blocked by controlling for observed antecedent variables.

26.7 An Example: Blau and Duncan's Basic Stratification Model

As part of an extensive study of social and economic inequality in the United States, Blau and Duncan (1967) developed what they termed "a basic stratification model," depicted in the *path diagram* in Figure 26.13(a).[26] Blau and Duncan fit their model to data from a 1962 U. S. sample survey of more than 20,000 men between the ages of 20 and 64.[27]

The model includes five observed variables:

- The respondent's father's education, which is scaled as follows: (0) no school; (1) 1 to 4 years of elementary school; (2) 5 to 7 years of elementary school; (3) 8 years of elementary school; (4) 1 to 3 years of high school; (5) 4 years of high school; (6) 1 to 3 years of college; (7) 4 years of college; and (8) 1 or more years of post-graduate study. As Blau and Duncan note, this scaling is nearly a linear function of years of education.[28]
- The respondent's father's socioeconomic status ("SES") when the respondent was 16 years old, which is a property of the father's occupation. SES scores ranged from 0 to 96.[29]
- The respondent's education, using the same 0–8 scale as for father's education.
- The SES of the respondent's first job after his education was complete.
- The respondent's SES at the time of the survey.

[26] The path diagram in Figure 26.13(a) differs only trivially from Blau and Duncan's Figure 5.1: For example, Blau and Duncan don't give names to the error variables, which I designate $\varepsilon_1, \varepsilon_2$, and ε_3, and they show the estimated standardized regression coefficient for each of the directed arrows in the model, along with the correlation between father's education and father's SES, corresponding to the double-headed arrow in the path diagram.

[27] It's telling that a book entitled *The American Occupational Structure* dealt entirely with men. See Exercise 26.6 for more on Blau and Duncan's data.

[28] For this last point, also see Exercise 26.6(a).

[29] Duncan (1961) pioneered the construction of socioeconomic-status scales. The SES scores for occupations are either observed or predicted percentages of high ratings in a national survey of occupational prestige; they therefore have a theoretical range of 0 to 100.

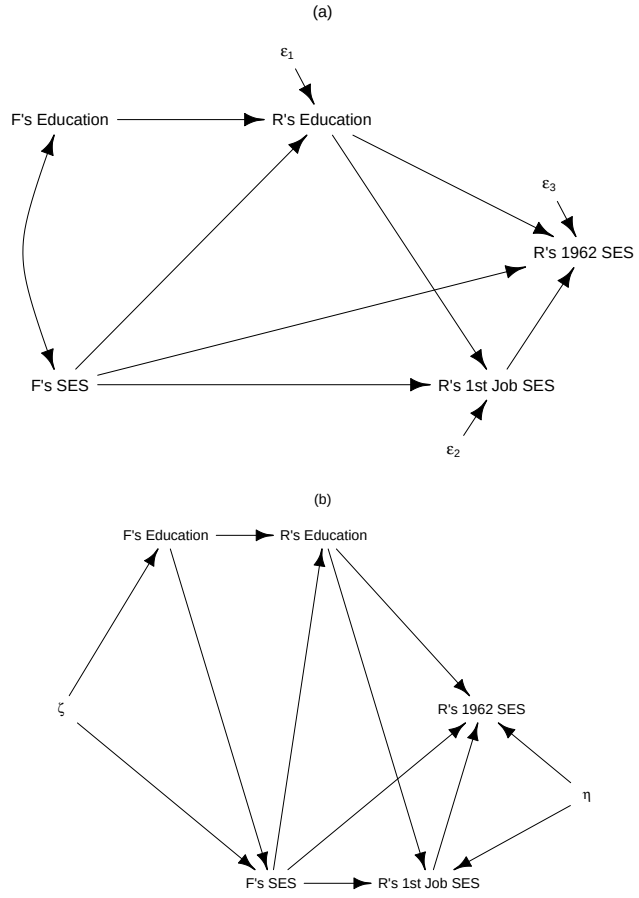

Figure 26.13 Blau and Duncan's basic stratification model: (a) as a path diagram; (b) in slightly modified form as a DAG

Because it is acyclic,[30] Blau and Duncan's path diagram could be construed as a DAG (once we adopt a convention for the double-headed arrow, as explained below), but it is given a slightly different, and more specific, interpretation:

1. There is an implied linear regression equation for each endogenous variable in the model—that is, each variable in the path diagram to which one or more directed arrows point. For example, for respondent's education,

$$\text{R's Education} = \beta_0 + \beta_1 \text{F's Education} + \beta_2 \text{F's SES} + \varepsilon_1$$

   After standardizing the variables in the model, Blau and Duncan estimated this and the other implied regression equations by linear least squares, and so the intercept β_0 was set to 0. Such standardized regression coefficients in a path model are called *path coefficients*.[31]

2. Father's education and father's SES are treated as exogenous variables in the model—that is, variables determined outside of the model. As a consequence, no arrows point towards the exogenous variables, and the exogenous variables are independent of the error variables in the model (here, $\varepsilon_1, \varepsilon_2$, and ε_3)

   The double-headed arrow in the path diagram linking the two exogenous variables is interpreted non-causally as a statistical association. That is, the model is agnostic about whether father's education causes father's SES, father's SES causes father's education, father's SES and education have one or more common prior causes, or any or all of these possibilities.

   In contrast, a double-headed arrow in a DAG is conventionally taken to imply a latent common prior cause (or causes), and so, for example, $U \longleftrightarrow V$ is equivalent to $U \longleftarrow \omega \longrightarrow V$, where ω represents an unobserved prior cause (or the aggregated unobserved prior causes) of U and V

Figure 26.13(b) redraws Blau and Duncan's path diagram as a DAG. The error variables in the model, $\varepsilon_1, \varepsilon_2$, and ε_3, are independent of one-another and affect only one observed variable each; as a consequence, the errors have no implications for associations among the observed variables, and it's unnecessary to show them explicitly in the DAG, although it would be harmless to do so.[32] As explained, I replaced the double-headed arrow linking father's education and father's SES with an unobserved confounder ζ affecting both variables, but (as Blau and Duncan would agree makes sense) I also specified a direct arrow from father's education to father's SES.[33] In contrast to Blau and Duncan's path

[30] An acyclic path diagram is said to be *recursive*, which also implies that each error variable in the diagram points only to one observed variable and that the errors are independent of each other.

[31] One of Sewall Wright's goals, from which the term "path analysis" derives, was to express observed correlations as functions of path coefficients.

[32] Pearl et al. (2016), for example, routinely show latent error variables in DAGs.

[33] Because of the unobserved confounder, however, the effect of father's education on father's SES isn't estimable. For their reasoning in specifying the basic stratification model, see Blau and Duncan (1967, Chap. 5).

model, I also added an unobserved confounder η of respondent's first-job SES and 1962 SES, because it's implausible that these two very similar variables would not share unobserved causes.[34]

Imagine now that, guided by the DAG in Figure 26.13(b), we wish to estimate the effect of father's SES on respondent's SES in 1962 (that is, intergenerational transmission of SES—the inverse of social mobility), and of respondent's education on respondent's SES (that is, the contribution of education to status attainment). How should we proceed?

- All of the back-door paths connecting father's SES to respondent's 1962 SES go through father's education, and so it suffices to regress respondent's SES on father's education and father's SES to estimate the effect of the latter. This would not be the case had I omitted the arrow from father's education to father's SES: In the absence of that arrow, the effect of father's SES on respondent's SES wouldn't be estimable because of inability to block back-door paths through the unobserved confounder ζ.
- Similarly, all of the back-door paths connecting respondent's education to respondent's 1962 SES go through father's SES, and so to estimate the effect of respondent's education we can regress respondent's SES on father's SES and respondent's education. That it is sufficient to control for father's SES *doesn't* in this case depend on the additional arrow from father's education to father's SES.[35]

It is also interesting to examine the conditional statistical independencies among the observed variables implied by the DAG for Blau and Duncan's model. There are two:

1. (respondent's first-job SES $\perp\!\!\!\perp$ father's education) | (father's SES and respondent's education)
2. (respondent's 1962 SES $\perp\!\!\!\perp$ father's education) | (father's SES and respondent's education)

These conditional independencies correspond to two arrows that could be added to the DAG without violating its acyclic structure: arrows from father's education to respondent's first job and to 1962 SES. The conditional independencies also follow from the rules for conditional d-separation given in Section 26.5.

This observation suggests that the DAG for Blau and Duncan's model constrains the data in a testable manner. If we regress respondent's first-job and 1962 SES on father's education, father's SES, and respondent's education, then the coefficient of father's education in each regression should be 0 within sampling error.[36] That's not to say, however, that satisfying these constraints *proves*

[34] See Exercise 26.6(b).

[35] See Exercise 26.6(c) for the results of the regression of respondent's 1962 SES on father's education and father's SES, and the regression of respondent's 1962 SES on father's SES and respondent's education.

[36] See Exercise 26.6(d) for these regressions.

that the causal structure of the DAG is correct. As I previously pointed out in connection with Figures 26.2(a) (page 5) and 26.4(a) (page 8), it's generally the case that different, causally distinct, DAGs can be *observationally equivalent* in that they imply the same conditional independencies. This would be true for Blau and Duncan's DAG, for example, were we nonsensically to reverse the arrow from father's education to father's SES.

26.8 Potential Outcomes, Causal Inference, and DAGs

Consider a simple randomized experiment in which half of n subjects are assigned to a treatment condition and half to a control condition, with the random assignment encoded by the dummy variable $X_i = 1$ if subject i is in the treatment group and $X_i = 0$ if subject i is in the control group. A response variable Y_i is subsequently measured for each subject, and we focus either on the conditional distribution of the response given experimental condition, $p[Y|(X = x)]$, or on some property of the conditional distribution such as its mean $\mu|(X = x)$.

The *potential outcomes* (or *counterfactual*) approach to causal inference, developed by Donald Rubin and his colleagues, initially in Rubin (1974),[37] is a framework for conceptualizing statistical causation in experimental and observational data. We imagine that each subject i in our simple experiment is associated with two potential outcomes: $y_i^{(1)} \equiv Y_i|(X_i = 1)$ if the subject is assigned to the experimental condition; and $y_i^{(0)} \equiv Y_i|(X_i = 0)$ if the subject is assigned to the control condition. Here, as is common in explanations of Rubin's causal model, I've treated the potential outcomes for subject i as fixed values, $y_i^{(1)}$ and $y_i^{(0)}$, but, perhaps more realistically in most instances,[38] we can also think of them as random variables, $Y_i^{(1)} \equiv Y_i|(X_i = 1)$ and $Y_i^{(0)} \equiv Y_i|(X_i = 0)$. In either case, the observed response Y_i is random because of the random assignment of X_i. For simplicity, I'll stick with the fixed values $y_i^{(1)}$ and $y_i^{(0)}$ for the potential outcomes.

What Holland (1986) calls the "fundamental problem of causal inference" is that we can't observe *both* $y_i^{(1)}$ *and* $y_i^{(0)}$ for subject i, which prevents us from directly measuring the effect of X on Y for subject i, defined as $y_i^{(1)} - y_i^{(0)}$.[39] If we observe $Y_i = y_i^{(1)}$ for subject i, we *don't* observe $y_i^{(0)}$, and vice-versa—hence the use of the term "counterfactual" to characterize the potential-outcomes approach. Half of the data for computing the individual-subject effects are missing.

[37] Essentially the same approach was proposed much earlier by Jerzy Neyman, an important contributor to the theory of statistical inference in the first half of the 20th Century; see Neyman (1990) [1923].

[38] For example, $Y_i^{(x)}$ might include measurement error or an inherently random component.

[39] That is, in general we can't observe both $y_i^{(1)}$ and $y_i^{(0)}$ at the same moment of time, and so to observe both we have to make additional assumptions—for example, that if we observe $y_i^{(1)}$ and $y_i^{(0)}$ sequentially, the first measurement doesn't influence the second.

The potential-outcomes approach requires that, prior to collecting data, all subjects can be observed under both treatment and control, which is typically interpreted to mean that the cause X is amenable to experimental manipulation, at least in principle, even if the data are observational. Immutable or intrinsic characteristics of subjects are consequently ruled out as causes.[40]

Because the individual effects $y_i^{(1)} - y_i^{(0)}$ can't be observed, let us focus instead on their distribution over the n subjects in the study (or, alternatively, a population of subjects from whom the n subjects in the study were drawn), and, in particular on the average value of this difference, $E_i\left(y_i^{(1)} - y_i^{(0)}\right)$, where $E_i(\cdot)$ denotes expectation over all subjects. This average is estimable because

$$\begin{aligned} E_i\left(y_i^{(1)} - y_i^{(0)}\right) &= E_i\left(y_i^{(1)}\right) - E_i\left(y_i^{(0)}\right) \\ &= E_i[Y_i|(X_i = 1)] - E_i[Y_i|(X_i = 0)] \end{aligned} \tag{26.2}$$

$E_i[Y_i|(X_i = 1)]$ can be estimated by the average Y-value for subjects in the treatment condition, and $E_i[Y_i|(X_i = 0)]$ by the average Y-value for subjects in the control condition. The second line of Equation 26.2 is justified because the subjects in each treatment are a simple random sample of all of the subjects. That would not generally be the case in the absence of randomization—for example, if the subjects in the treatment condition were self-selected.

In the absence of random assignment, we must either justify Equation 26.2 by arguing that the mechanism assigning subjects to treatments is effectively random (a *natural experiment*), or produce an unbiased estimate of $E_i\left(y_i^{(1)} - y_i^{(0)}\right)$ by conditioning on an antecedent variable (or variables), say Z, for which

$$E_i\left(y_i^{(1)} - y_i^{(0)}\right) = E_i[Y_i|(X_i = 1, Z_i = z)] - E_i[Y_i|(X_i = 0, Z_i = z)]$$

We can estimate the average effect of X on Y by averaging the values of Y for $X = 1$ and $X = 0$ over all values z of Z and taking the difference in the two averages—or by the regression of Y on X and Z.[41] There is a large literature in the potential outcomes tradition that addresses how to go about selecting control variables and how to control for them.[42]

[40] What counts as an intrinsic characteristic of subjects is not as obvious as it may seem. For example, Holland (1986, p. 946) cites race and gender as two intrinsic characteristics of individuals that are not subject to experimental manipulation, but we can certainly imagine performing experiments in which these characteristics *are* manipulated, as in photos attached to job applications, or musical auditions with or without the ability of judges to see the performers.

[41] If X interacts with Z in determining Y, we might prefer to estimate the interaction (i.e., the varying effect of X at specific values z of Z) rather than the main effect of X (its effect averaged over values z of Z).

[42] In addition to the direct comparison of conditional means and multiple regression, other strategies for controlling for confounders include *matching* and *propensity scores*: See the recommended readings at the end of this chapter.

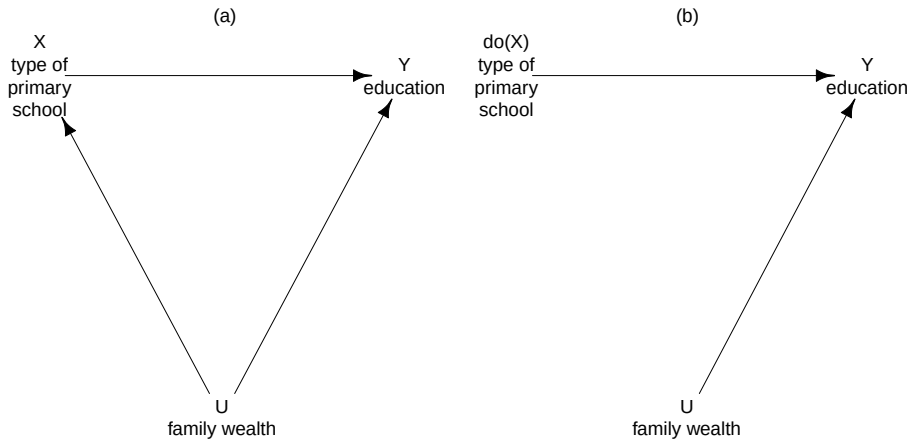

Figure 26.14 (a) In observational data, the variable U is a confounder of the causal relationship between X and Y; (b) manipulating X directly, denoted by $\text{do}(X)$, removes all arrows pointing to X, in this case $U \longrightarrow X$, and so it is no longer necessary to control for U to estimate the effect of X on Y.

The potential-outcomes (or counterfactual) framework for causal inference requires that, prior to data collection, the value of the response variable Y for every subject in a study can be observed with the explanatory variable X set to each of its possible values. This requirement is equivalent to asserting that X can be subject to experimental control, at least in principle, even if the data at hand are observational.

When data are collected, however, only one value x_i of the explanatory variable and the associated value $y_i^{(x_i)} = Y_i|(X = x_i)$ of the response are realized for each subject i. The effect of X on Y for an individual subject, defined as differences in the response Y_i with X set to its distinct values (the potential outcomes $y_i^{(x)}$ for all x), is therefore unobservable—the fundamental problem of causal inference.

Attention consequently shifts to the distribution of the individual effects, or to characteristics of this distribution—for example, the individual effects averaged over subjects. Average effects are estimable in experimental data and may be estimable in observational data if confounders can be controlled statistically.

How does all this relate to DAGs? Experimental control of X in a DAG can be represented by the *do operator*, where we understand $\text{do}(X = x)$ to mean that X is *set* to the value x. More generally $\text{do}(X)$ implies that X is under

experimental control: do(X), therefore, erases all arrows in an observational DAG that point to X because, in an experiment, X has no causes other than direct manipulation.

An example appears in Figure 26.14, where panel (a) is an observational DAG in which U confounds the effect of X on Y. I imagine here that X represents the type of primary school an individual attends—public or private; Y represents level of eventual completed education; and U represents family wealth. To estimate the effect of X on Y from observational data, we must therefore control for U to block the back-door path through U. In panel (b) (admittedly unrealistically for this example), we exert experimental control over X [i.e., do(X)], erasing the arrow $U \longrightarrow X$. We can now obtain an unbiased estimate of the effect of X on Y by regressing Y on X alone.[43]

It is more common, however, to reason in reverse—that is, to start with a DAG in which X is directly manipulated, at least hypothetically, and so has no arrows pointing to it, and then to ask whether we can estimate the (same) effect of X on Y from observational data by controlling for antecedent variables, some of which point directly to X. That leads to the methods for closing back-door paths described earlier in this chapter, and more generally to Pearl's *do-calculus*.[44]

> In a DAG, direct experimental manipulation of X is represented by the do operator, do(X), which has the effect of removing all arrows that point directly to X in the corresponding observational DAG. The observational DAG can help us to decide whether and how to obtain an unbiased estimate equivalent to the effect of do(X) by controlling for antecedent variables to close back-door paths linking X and Y.

26.9 DAGs and Missing Data

Missing data in regression models are discussed in Chapter 20, and three types of missing data (introduced by Donald Rubin, 1976[45]) are described in Section 20.1: data that are missing completely at random (MCAR), missing at

[43] It may still be advantageous to include U in the regression equation as a covariate, to reduce the error variance and thus obtain a more precise estimate of the coefficient of X, but the key point is that ignoring U doesn't bias the estimate of the effect of X.

[44] The details of do-calculus, which is a coherent set of rules for estimating causal effects in observational DAGs, are too complex to develop here, but see the recommended reading at the end of the chapter, in particular Pearl (2009), Pearl and Mackenzie (2018), and Morgan and Winship (2014).

[45] It is interesting, and probably not coincidental, that Rubin made fundamental contributions to two topics discussed in this chapter: the potential-outcomes approach to causal inference (discussed in Section 26.8) and the treatment of missing data.

random (MAR), or missing not at random (MNAR). To recapitulate briefly, missing data are MCAR if the observed data are equivalent to a simple random sample from the complete data set; MAR if missingness (i.e., the probability that a value is missing) is unrelated to the missing values themselves given the observed data; and MNAR if missingness depends on the missing values after accounting for the information in the observed data. These distinctions are important for how data with missing values should be analyzed to obtain consistent estimators of regression coefficients and other parameters.

Though a centrally important advance in principled methods for dealing with missing data, Rubin's categorization is admittedly opaque. Perhaps more importantly, it is difficult to know in an application whether reasonable assumptions about how missing data may have been generated correspond to the various types of missing data—for example, whether it is justified on the basis of these assumptions to treat missing data as MAR. With a small adjustment to Rubin's definition of MAR, DAGs can help to determine whether data are MCAR, MAR, or MNAR, and, in some circumstances, even how to obtain consistent estimators of parameters of interest when missing data are MNAR.

The discussion of DAGs for missing data in this section relies heavily on a review paper by Mohan and Pearl (2021). As is typical in this chapter, the treatment here is abbreviated and simplified, and is meant principally to clarify fundamental ideas.

With the exception of panel (a), where data are complete, the DAGs in Figure 26.15 represent examples of different mechanisms for generating missing data, corresponding (roughly) to the distinction among MCAR, MAR, and MNAR. The following conventions are used in these DAGs, termed *missingness graphs* (or *m-graphs*) by Mohan and Pearl (2021):[46]

- The completely observed response variable is denoted Y.
- Some of the values of Y may not be available to the researcher. In that case, the observed response Y^* includes missing values. In panels (b), (c), and (d), only the response variable is subject to missing data, and the explanatory variables (X and W) are completely observed. M-graphs can also be used to analyze situations in which several—indeed, all—variables are subject to missing data, but examples in which missing data are restricted to the response are particularly simple, which is the reason for using them here.
- The *missingness indicator variable* M is coded 1 for case i if the value of Y_i is missing and 0 if it is observed. Thus, $Y_i^* = Y_i$ if $M_i = 0$, and $Y_i^* =$ missing if $M_i = 1$. The observed response Y^* has Y and M as its direct "causes." If more than one variable were subject to missing data, then there would be a distinct M indicator variable for each (e.g., M_X for X, M_W for W, etc.).

[46] My conventions for labeling variables in m-graphs differ from those in Mohan and Pearl (2021).

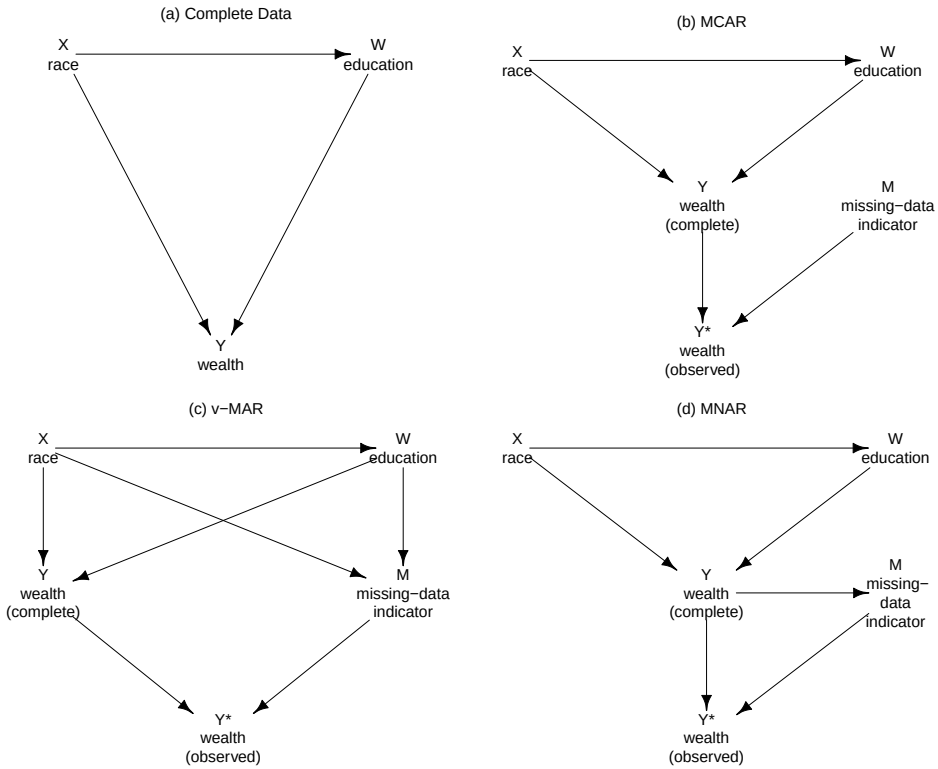

Figure 26.15 Illustrative missingness graphs (m-graphs) representing various missing-data patterns: (a) complete data set; (b) missing data that are missing completely at random (MCAR); missing data that are missing at random (v-MAR); (d) missing data that are missing not at random (MNAR).

- Other variables (X and W in these examples) are interpreted in the usual manner for DAGs.

As mentioned, the DAG in Figure 26.15(a) represents complete data on X, W, and V. From earlier material in this chapter, we know that we can estimate the effect of X on Y (for which W is a mediator) from the regression of Y on X alone, and the effect of W on Y (for which X is a confounder) from the regression of Y on both W and X.

The DAG in panel (b) of Figure 26.15 is similar to that in panel (a), except that the observed response Y^* is incomplete. We can see from the structure of the DAG, however, that missingness (i.e., M) is unrelated to all of X, W, and Y (i.e., the completely observed response), and so missing data in Y^* are MCAR. We could, for example, estimate the effects of X and W on Y by complete-case analysis (see Section 20.2), performing regressions as for the DAG in panel (a).

More generally, imagine that some of the variables $\mathbf{V}_c$ in a DAG are completely observed, that others $\mathbf{V}_m$ contain some missing values in the observed data (but are conceptualized as including the unobserved missing values), and that the missingness indicators are in $\mathbf{M}$. There may also be latent variables $\boldsymbol{\Omega}$, which, of course, aren't observed, and $\mathbf{V}^*$ contains the observed data, including missing values, corresponding to the variables in $\mathbf{V}_m$. Mohan and Pearl (2021) show that missing data are MCAR if (in my notation) $\{\mathbf{V}_c, \mathbf{V}_m, \boldsymbol{\Omega}\} \perp\!\!\!\perp \mathbf{M}$. In the DAG in Figure 26.15(b), $\mathbf{V}_c = \{X, W\}$, $\mathbf{V}_m = \{Y\}$, $\mathbf{M} = \{M\}$, $\mathbf{V}^* = \{Y^*\}$, and $\boldsymbol{\Omega}$ is empty (i.e., there are no latent variables). As I explained, it's apparent from the DAG that $\{X, W, Y\} \perp\!\!\!\perp M$, and so missing data are MCAR.

Mohan and Pearl (2021) define a slightly stronger condition than MAR, which they term *v-MAR* (missing at random defined in terms of variables): Missing data are v-MAR if $(\{\mathbf{V}_m, \boldsymbol{\Omega}\} \perp\!\!\!\perp \mathbf{M}) | \mathbf{V}_c$, that is if the variables with missing data and the latent variables in the DAG are independent of missingness given the fully observed variables. V-MAR implies MAR, but not vice-versa. Nevertheless, it's much easier to reason substantively about v-MAR than more abstractly about MAR, and so v-MAR is more likely to be helpful in applications.

In the DAG in Figure 26.15(c), missingness M depends on the fully observed variables X and W, but given X and W, the complete response Y is independent of M (that is, $(Y \perp\!\!\!\perp M) | \{X, W\}$). In this example, therefore, missing data are v-MAR. Because there are missing data only in the response, we can again consistently estimate the effects of interest by complete-case regressions. More generally, we could use methods, such as multiple imputation, that are appropriate for missing data that are MAR.

The DAG in Figure 26.15(d) represents a situation in which missing data are MNAR, because missingness M in the observed response Y^* depends directly on the complete response variable Y, even if we condition on the completely observed variables X and W. In this case, we can't consistently estimate the effects of X and W on Y without introducing additional information.

More generally, however, Mohan and Pearl (2021) explain how in certain circumstances it's possible to obtain consistent estimators when missing data

are MNAR. Even when missing data are v-MAR, an m-graph may suggest specific consistent estimators of effects of interest that are more efficient than generic estimators such as those obtained using multiple imputation. Mohan and Pearl also describe how to test for MCAR and v-MAR in m-graphs that imply these conditions.

The variables in a missingness DAG (or m-graph) are divided into several sets: completely observed variables $\mathbf{V}_c$; variables $\mathbf{V}_m$ some of whose values are missing in the observed data; latent variables $\boldsymbol{\Omega}$; missingness indicator variables $\mathbf{M}$, one for each variable in $\mathbf{V}_m$; and observed variables $\mathbf{V}^*$ including missing data and corresponding to the variables in $\mathbf{V}_m$.

Missing data are MCAR if $\{\mathbf{V}_c, \mathbf{V}_m, \boldsymbol{\Omega}\} \perp\!\!\!\perp \mathbf{M}$, and are v-MAR if $(\{\mathbf{V}_m, \boldsymbol{\Omega}\} \perp\!\!\!\perp \mathbf{M})|\mathbf{V}_c$. V-MAR is a slightly stronger condition than Rubin's MAR (so v-MAR implies MAR), but, because of its m-graph interpretation, v-MAR is easier to justify in applications.

26.10 Concluding Remarks about DAGs and Causal Inference

Most research in the social sciences is based on observational data, and I believe that most social-science researchers who use observational data are interested in drawing causal inferences—whether or not they explicitly acknowledge it—and not merely in discovering statistical associations or predicting future observations. I've argued in this chapter that DAGs are a potentially useful conceptual device for deciding whether it's possible to obtain estimates that can be given a causal interpretation, and, if so, how to go about deciding which antecedent variables to control.

DAGs are also useful for clarifying which variables *shouldn't* be controlled statistically, such as mediating variables and consequences of the response, which can be colliders. Both mediators and consequences may be very useful in pure prediction problems, but, as DAGs make clear, they can wreak havoc with causal inferences,[47] as can using "statistical significance" as a criterion to eliminate weakly predictive antecedent variables as controls: Such antecedent variables may have small or even absent direct effects (their effects may be entirely indirect through X), yet failing to control them can still open back-door paths that bias the estimate of the effect of X on Y.

Perhaps it's stating the obvious to say that using a DAG to support causal inferences from observational data rests on the truth of the causal structure

[47] Thus, model-selection methods that focus on prediction (such as those described in Section 22.1) are nearly guaranteed to produce causally misleading models.

represented in the DAG. If the DAG is wrong, then causal assertions based on it may be wrong as well. Although it's important to recognize this fact, the conditional nature of causal inference based on DAGs merely reflects the familiar limitation of observational data: There can be unobserved confounders of which we are unaware, and so the antecedent variables that we control statistically may be insufficient for blocking unanticipated back-door paths. This is why it's important to think about potential confounders, even if they are unobserved or unobservable.

Despite their utility, one can easily get in trouble with DAGs when they are adjusted to permit the causal inferences that the researcher wants to make. That is, there is a temptation to redraw a DAG if, on a first attempt, it turns out that there are unobserved confounders creating back-door paths that can't be blocked by controlling for observed antecedent variables, or where an initial graph isn't acyclic. In my experience, this kind of respecification is common in applications of structural-equation models, where researchers adjust their models to ensure that they are identified (i.e., estimable). This kind of respecification may not represent deliberate dishonesty but it nevertheless is self-deceptive and self-defeating.

DAGs help us reason about how to estimate causal effects in observational data, including in situations where some confounders are unmeasured or even unmeasurable. The validity of the conclusions and procedures for causal inference that we derive from a DAG depend on the validity of its causal structure. Although some causal assumptions in a DAG may be testable, it is always true that causal conclusions also depend on untestable—that is, extra-statistical—assumptions, which must therefore be justified on substantive grounds.

Summary

- A graph is a labeled set of nodes (points) connected by edges (line segments). Graphs may be undirected or directed, in which case the edges are represented as single-headed arrows. The node at the head of each arrow in a directed graph is the parent node and that at the tail is the child node.

  A path through a graph between two nodes is a sequence of consecutive edges connecting the nodes, and a directed path is a path all of whose arrows point in the same direction. The initial node of a directed path is an ancestor of the terminal node, which is a descendant of the initial node. A directed graph is acyclic if it has no reciprocal paths or loops.

- Directed acyclic graphs (or DAGs) represent causal relationships among variables, where the variables are the nodes of the graph, and arrows

connecting the variables represent direct effects, with the direct cause at the tail of an arrow and the effect at the tip. To say that DAGs are acyclic implies that causation is unidirectional, with no reciprocal arrows or feedback loops.

- A confounder creates a back-door path connecting a cause X and effect Y, which in turn generates spurious (i.e., non-causal) association between these two variables. We can estimate the effect of X on Y by controlling statistically for the confounder. The path $X \longleftarrow W \longrightarrow Y$ is called a causal fork.

- A mediator is a variable that intervenes between a cause X and effect Y. We should not control statistically for a mediator if we want to estimate the effect of X on Y. The path $X \longrightarrow W \longrightarrow Y$ is called a causal chain. Confounders and mediators can't be distinguished solely on statistical grounds.

- To obtain an unbiased estimator of the effect of X on Y we must close (i.e., block) all of the back-door paths connecting the two variables in the DAG (the back-door criterion). Closing all back-door paths does not in general require that we control for all variables in the DAG that are causally prior to X and Y. It's generally advantageous to control for the antecedent variable or variables that are sufficient to close all back-door paths and that, in doing so, produce the most precise estimate of the effect of X on Y. When we have a choice, we therefore prefer to control for antecedent variables that are close to Y and remote from X.

- Colliders are variables that block sources of non-causal association between X and Y. Controlling for a collider opens a non-causal path between the two focal variables, biasing the estimated effect of X and Y.

- Controlling for descendants of both X and Y or of Y alone biases the estimate of the effect of X on Y, while controlling for descendants of X alone makes the estimate of the effect of X less precise. We should therefore avoid controlling for descendants of the focal causal variables.

- Selection bias in estimating the effect of X on Y can be understood as controlling for a collider, in which we examine the partial relationship between X and Y within one category of the collider—that is, for self-selected subjects.

- An important contribution of DAGs is that they can help us to understand the role of unobserved (latent) variables in causal inference. In certain cases, we may be able to close back-door paths that include unobserved confounders by controlling for observed variables along these paths.

- DAGs may help us to identify instrumental variables, as causally prior variables that affect X but not Y directly and that are reasonably construed as unrelated to the omitted (i.e. latent) causes of Y. Such instrumental

variables make it possible to estimate the effect of X on Y even when X is related to the regression error in Y, such as when there are unobserved confounders creating back-door paths that can't be blocked by controlling for observed antecedent variables.

- The potential-outcomes (or counterfactual) framework for causal inference requires that, prior to data collection, the value of the response variable Y for every subject in a study can be observed with the explanatory variable X set to each of its possible values. This requirement is equivalent to asserting that X can be subject to experimental control, at least in principle, even if the data at hand are observational.

  When data are collected, however, only one value x_i of the explanatory variable and the associated value $y_i^{(x_i)} = Y_i|(X = x_i)$ of the response are realized for each subject i. The effect of X on Y for an individual subject, defined as differences in the response Y_i with X set to its distinct values (the potential outcomes $y_i^{(x)}$ for all x), is therefore unobservable—the fundamental problem of causal inference.

  Attention consequently shifts to the distribution of the individual effects, or to characteristics of this distribution—for example, the individual effects averaged over subjects. Average effects are estimable in experimental data and may be estimable in observational data if confounders can be controlled statistically.

- In a DAG, direct experimental manipulation of X is represented by the do operator, do(X), which has the effect of removing all arrows that point directly to X in the corresponding observational DAG. The observational DAG can help us to decide whether and how to obtain an unbiased estimate equivalent to the effect of do(X) by controlling for antecedent variables to close back-door paths linking X and Y.

- The variables in a missingness DAG (or m-graph) are divided into several sets: completely observed variables $\mathbf{V}_c$; variables $\mathbf{V}_m$ some of whose values are missing in the observed data; latent variables $\boldsymbol{\Omega}$; missingness indicator variables $\mathbf{M}$, one for each variable in $\mathbf{V}_m$; and observed variables $\mathbf{V}^*$ including missing data and corresponding to the variables in $\mathbf{V}_m$.

  Missing data are MCAR if $\{\mathbf{V}_c, \mathbf{V}_m, \boldsymbol{\Omega}\} \perp\!\!\!\perp \mathbf{M}$, and are v-MAR if $(\{\mathbf{V}_m, \boldsymbol{\Omega}\} \perp\!\!\!\perp \mathbf{M})|\mathbf{V}_c$. V-MAR is a slightly stronger condition than Rubin's MAR (so v-MAR implies MAR), but, because of its m-graph interpretation, v-MAR is easier to justify in applications.

- DAGs help us reason about how to estimate causal effects in observational data, including in situations where some confounders are unmeasured or even unmeasurable. The validity of the conclusions and procedures for causal inference that we derive from a DAG depend on the validity of its causal structure. Although some causal assumptions in a

DAG may be testable, it is always true that causal conclusions also depend on untestable—that is, extra-statistical—assumptions, which must therefore be justified on substantive grounds.

Exercises

Exercise 26.1. How would you go about blocking the back-door paths to estimate the effect of X on Y in the DAG in Figure 26.5(b) (on page 10)? Explain your reasoning.[48]

Exercise 26.2. Add an arrow from race (V) to family wealth (W) to the DAG in Figure 26.6(b) (page 13). How, if at all, does that modify the status of U as a collider? What is the consequence of controlling statistically for V in estimating the effect of X on Y?

Exercise 26.3.

(a) Reproduce the illustrative collider simulation reported in Table 26.1 (on page 14), adding the model that regresses Y on X and W. Is anything learned from this additional regression beyond the conclusions already drawn from the models in Table 26.1?

(b) Focusing now on models 1, 2, 3, and 5 in the simulation, draw the added-variable plot (AV plot) for X in each model. Do these plots help you visualize the properties of the estimators B_X of $\beta_X = 1$ for these models?

Exercise 26.4.

(a) Referring to the DAG in Figure 26.8 (on page 17), generate $n = 10,000$ independent observations on X, Y, the collider U, and the various descendants of these variables, W, V, and T, according to the following scheme:

$$
\begin{aligned}
X &\sim N(0, 1) \\
Y &\sim N(X, 1) \\
U &\sim N(X + Y, 1) \\
W &\sim N(U, 1) \\
V &\sim N(X, 1) \\
T &\sim N(Y, 1)
\end{aligned}
$$

Then regress Y on each of the follow (sets of) variables: (1) X; (2) X and U; (3) X and W; (4) X and V; and (5) X and T. Examine the

[48] Optionally, for this and other exercises, use daggity to check you work. You can access daggity via its web interface or as an R package: See http://www.dagitty.net/.

coefficient of X, its standard error, and the residual standard error from each of these regressions. What do you conclude about the bias or unbias and relative efficiency of the various estimates of the effect of X on Y? Do your conclusions square with those stated in the text.

(b) Now reverse the causal arrow between W and U, obtaining

$$W' \sim N(0, 1)$$
$$U' \sim N(X + Y + W', 1)$$

Fit additional regressions of Y on X and U' and on X and W'. What do you conclude?

Exercise 26.5. Refer to the DAG in Figure 26.10 (page 21):

(a) Are there any pairs of variables other than C_1 and C_2 that are d-separated? If so, which one(s)?

(b) Are there any sets of variables other than $\{C_1, C_2, M_1, M_2\}$ that conditionally d-separate X and Y? If so, which set(s)?

(c) There are many other pairs of variables in this DAG that can be rendered conditionally independent by controlling for particular sets of variables. Can you identify three such cases? Alternatively, use appropriate software to identify *all* conditional independencies that can be derived from the DAG.

(d) As I explained, the variables C_1 and C_2 are d-connected conditioning on X, even though they are unconditionally d-separated. Can you find another variable or other variables that render C_1 and C_2 conditionally d-connected? Recalling part (a) of this exercise, can you find more pairs of d-separated variables that can be rendered d-connected by conditioning on another variable or other variables?

Exercise 26.6. The data for Blau and Duncan's basic stratification model are in the file `BlauDuncan.txt`, which is available on the website for the text. The data file includes the five variables in Blau and Duncan's model, along with the respondents' age and race. There are some missing values in the Blau and Duncan data, denoted by `NA`. See Blau et al. (1983, 1994) for details about the original data set from which the data used here were drawn.

(a) As mentioned, Blau and Duncan used a 0–8 coding for nine levels of education. Recode education so that its values reflect the mid-points, in years, of the education categories. You'll have to pick a reasonable value for the last, open-ended category of one or more years of post-graduate

education. Then check the correlation between education and education in years, both for respondents and for their fathers. Will it matter which version of education is used?

(b) What are the consequences for causal inference, if any, of adding the unobserved variable η to the DAG for Blau and Duncan's basic stratification model in Figure 26.13(b) (on page 26)? (*Hint*: How might one go about estimating the effect of first-job SES on 1962 SES?)

(c) To estimate the effects of father's SES and of respondent's education on respondent's SES, perform the regression of respondent's 1962 SES on father's education and father's SES, and the regression of respondent's 1962 SES on father's SES and respondent's education. What do you conclude from these regressions? Are there other ways to obtain unbiased estimates of the effects of father's SES and respondent's education? If so, which estimates would you prefer?

(d) Test the implied independencies in the DAG representing Blau and Duncan's basic stratification model by regressing each of respondent's first-job and 1962 SES on father's education, father's SES, and respondent's education.

(e) Blau and Duncan estimated their basic stratification model for standardized variables, obtaining the following coefficient estimates (and residual standard errors, S_E):

$$\widehat{\text{R's Education}} = 0.310 \times \text{F's Education} + 0.279 \times \text{F's SES}, \text{S}_\text{E} = 0.859$$

$$\widehat{\text{R's 1st Job SES}} = 0.224 \times \text{F's SES} + 0.440 \times \text{R's Education}, \text{S}_\text{E} = 0.818$$

$$\begin{aligned}\widehat{\text{R's 1962 SES}} = {} & 0.115 \times \text{F's SES} + 0.394 \times \text{R's Education} \\ & + 0.281 \times \text{R's 1st Job SES}, \text{S}_\text{E} = 0.753\end{aligned}$$

In addition, they reported the correlation between father's education and father's SES (the two exogenous variables in the original path model) as $r = .516$.

Blau and Duncan's analysis of the data used sampling weights meant to match the sample to the U. S. population, so your estimates can't be expected to reproduce their results exactly, but are they close?.

(f) * I included the sampling weight variable in the Blau and Duncan data file. Using the sampling weights, try to recover Blau and Duncan's estimates (within rounding error). Are you successful?

Exercise 26.7. Section 20.4.4 develops an example in which, using data reported by the United Nations, national infant mortality rates are regressed on gross domestic product per capita, the average number of years of education

for women, and the percentage of married women practicing contraception in the countries. There are missing values in each of these variables, and multiple imputation is employed to obtain estimated regression coefficients. Recall that multiple imputation of missing values is appropriate when missing data are MCAR or MAR. I also report the results of a complete-case analysis, which is appropriate when missing data are MCAR.[49]

(a) The DAG in Figure 26.16(a) is a proposed m-graph for the variables in the United Nations regression.[50] Do you find the proposed causal structure plausible? If not, how might it be improved? Assuming that the m-graph is reasonable, does it imply that missing data are MCAR, v-MAR, or MNAR? Based on this m-graph, how might you go about estimating the regression of infant mortality on the other variables in the data set?

(b) The DAG in Figure 26.16(b) proposes an alternative structure for the missing data in the infant-morality data set by introducing a latent variable ω that affects all four missingness indicators. Is this structure more plausible that in Figure 26.16(a)? What are the implications for estimation of adopting the alternative m-graph?

Recommended Reading

- Pearl (2009) is a detailed, and at times highly technical, exposition of the role of directed acyclic graphs in causal inference by the principal contributor to the subject. Pearl and Mackenzie (2018) is a much gentler, yet still extensive, introduction to the same topic. Pearl et al. (2016) is a considerably briefer, slightly more sophisticated, yet still accessible treatment of causal DAGs, and is a good place to start reading about the subject. Pearl's approach to causality emphasizes the conditional probability distribution of the response variable rather than common regression models, and so is more general than the approach in this chapter.

- Holland (1986) is a clear and accessible description of the potential-outcomes approach to causal inference, which the author terms "Rubin's causal model."

- The potential-outcomes and DAG approaches to causal inference developed largely separately, but Morgan and Winship (2014) lucidly describe both at length and explain the relationship between the two. I reviewed the somewhat less difficult—but also less complete—first edition of the book in Fox (2008).

[49] The data for the example are on the website for the text, in the file `UnitedNations.txt`.

[50] In the least-squares regressions reported in Section 20.4.4, both infant mortality and GDP are log-transformed. That's not relevant to the DAG, however, which, recall is nonparametric.

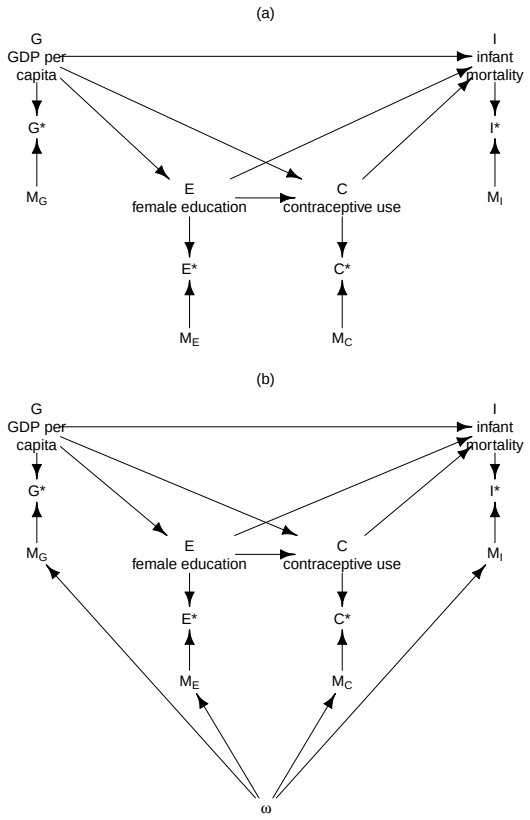

Figure 26.16 Alternative m-graphs for the infant-mortality data set.

- Mohan and Pearl (2021) present an extensive review of the application of DAGs to missing-data problems.

References for Chapter 26

P. M. Blau and O. D. Duncan. *The American Occupational Structure*. Wiley, New York, 1967.

P. M. Blau, O. D. Duncan, D. L. Featherman, and R. M. Hauser. *Occupational Changes in a General, 1962 and 1973 [computer file]*, 1983, 1994. Madison, WI: University of Wisconsin [producer]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor].

K. A. Bollen. *Structural Equations With Latent Variables*. Wiley, New York, 1989.

O. D. Duncan. A socioeconomic index for all occupations. In A. J. Reiss, Jr., editor, *Occupations and social status*, pages 109–138. Free Press, New York, 1961.

O. D. Duncan. Path analysis: sociological examples. *American Journal of Sociology*, 72:1–16, 1966.

O. D. Duncan. *Introduction to Structural Euqation Models*. Academic Press, New York, 1975.

J. Fox. Review of Stephen L. Morgan and Christopher Winship, Counterfactuals and Causal Inference: Methods and Principles for Social Research (New York: Cambridge University Press, 2007). *Canadian Journal of Sociology*, 33:432–435, 2008.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–960, 1986.

Mayo Clinic. Hormone therapy: Is it right for you? `https://www.mayoclinic.org/diseases-conditions/menopause/in-depth/hormone-therapy/ART-20046372`, 2022. Accessed: 2023-01-02.

K. Mohan and J. Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116:1023–1037, 2021.

S. L. Morgan and C. Winship. *Counterfacturals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, Cambridge, second edition, 2014.

J. S. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. section 9. Translated from the Polish original, which appeared in *Roczniki Nauk Rolniczych Tom (Annals of Agricultural Sciences)*, X: 1–51, 1923 [with an introduction and commentary, D. M. Dabrowska and T. P. Speed translators and editors]. *Statistical Science*, 5: 463–480, 1990.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, 2000.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, second edition, 2009.

J. Pearl and D. Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, New York, 2018.

J. Pearl, M. Glymour, and N. P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, Chichester UK, 2016.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

D. B. Rubin. Inference and missing data [with discussion]. *Biomatrika*, pages 581–592, 1976.

P. D. Stolley. When genius errs: R. A. Fisher and the lung cancer controversy [with commentary]. *American Journal of Epidemiology*, 133:416–436, 1991.

J. Textor, B. van der Zander, M. S. Gilthorpe, M. Liśkiewicz, and G. T. Ellison. Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, 2017.

S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20: 557–585, 1921.

Index