3

The Effects of Interventions

3.1 Interventions

The ultimate aim of many statistical studies is to predict the effects of interventions. When we collect data on factors associated with wildfires in the west, we are actually searching for something we can intervene upon in order to decrease wildfire frequency. When we perform a study on a new cancer drug, we are trying to identify how a patient's illness responds when we intervene upon it by medicating the patient. When we research the correlation between violent television and acts of aggression in children, we are trying to determine whether intervening to reduce children's access to violent television will reduce their aggressiveness.

As you have undoubtedly heard many times in statistics classes, "correlation is not causation." A mere association between two variables does not necessarily or even usually mean that one of those variables causes the other. (The famous example of this property is that an increase in ice cream sales is correlated with an increase in violent crime—not because ice cream causes crime, but because both ice cream sales and violent crime are more common in hot weather.) For this reason, the randomized controlled experiment is considered the golden standard of statistics. In a properly randomized controlled experiment, all factors that influence the outcome variable are either static, or vary at random, except for one—so any change in the outcome variable must be due to that one input variable.

Unfortunately, many questions do not lend themselves to randomized controlled experiments. We cannot control the weather, so we can't randomize the variables that affect wildfires. We could conceivably randomize the participants in a study about violent television, but it would be difficult to effectively control how much television each child watches, and nearly impossible to know whether we were controlling them effectively or not. Even randomized drug trials can run into problems when participants drop out, fail to take their medication, or misreport their usage.

In cases where randomized controlled experiments are not practical, researchers instead perform observational studies, in which they merely record data, rather than controlling it. The problem of such studies is that it is difficult to untangle the causal from the merely correlative. Our common sense tells us that intervening on ice cream sales is unlikely to have any effect on crime, but the facts are not always so clear. Consider, for instance, a recent

Causal Inference in Statistics: A Primer, First Edition. Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell.

Companion Website: www.wiley.com/go/Pearl/Causality

University of Winnipeg study that showed that heavy text messaging in teens was correlated with "shallowness." Media outlets jumped on this as proof that texting makes teenagers more shallow. (Or, to use the language of intervention, that intervening to make teens text less would make them less shallow.) The study, however, proved nothing of the sort. It might be the case that shallowness makes teens more drawn to texting. It might be that both shallowness and heavy texting are caused by a common factor—a gene, perhaps—and that intervening on that variable, if possible, would decrease both.

The difference between intervening on a variable and conditioning on that variable should, hopefully, be obvious. When we intervene on a variable in a model, we fix its value. We *change* the system, and the values of other variables often change as a result. When we condition on a variable, we change nothing; we merely narrow our focus to the subset of cases in which the variable takes the value we are interested in. What changes, then, is our perception about the world, not the world itself.

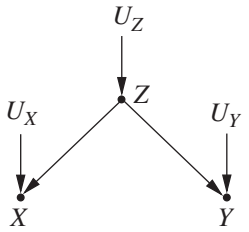

Figure 3.1 A graphical model representing the relationship between temperature (Z), ice cream sales (X), and crime rates (Y)

Consider, for instance, Figure 3.1 that shows a graphical model of our ice cream sales example, with X as ice cream sales, Y as crime rates, and Z as temperature. When we intervene to fix the value of a variable, we curtail the natural tendency of that variable to vary in response to other variables in nature. This amounts to performing a kind of surgery on the graphical model, removing all edges directed into that variable. If we were to intervene to make ice cream sales low (say, by shutting down all ice cream shops), we would have the graphical model shown in Figure 3.2. When we examine correlations in this new graph, we find that crime rates are, of course, totally independent of (i.e., uncorrelated with) ice cream sales since the latter is no longer associated with temperature (Z). In other words, even if we vary the level at which we hold X constant, that variation will not be transmitted to variable Y (crime rates). We see that intervening on a variable results in a totally different pattern of dependencies than conditioning on a variable. Moreover, the latter can be obtained

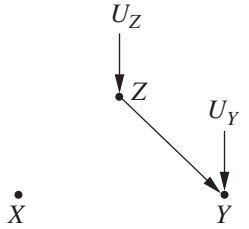

Figure 3.2 A graphical model representing an intervention on the model in Figure 3.1 that lowers ice cream sales

directly from the data set, using the procedures described in Part One, while the former varies depending on the structure of the causal graph. It is the graph that instructs us which arrow should be removed for any given intervention.

In notation, we distinguish between cases where a variable X takes a value x naturally and cases where we fix $X = x$ by denoting the latter $do(X = x)$. So $P(Y = y|X = x)$ is the probability that $Y = y$ conditional on finding $X = x$, while $P(Y = y|do(X = x))$ is the probability that $Y = y$ when we intervene to make $X = x$. In the distributional terminology, $P(Y = y|X = x)$ reflects the population distribution of Y among individuals whose X value is x. On the other hand, $P(Y = y|do(X = x))$ represents the population distribution of Y if *everyone in the population* had their X value fixed at x. We similarly write $P(Y = y|do(X = x), Z = z)$ to denote the conditional probability of $Y = y$, given $Z = z$, in the distribution created by the intervention $do(X = x)$.

Using *do*-expressions and graph surgery, we can begin to untangle the causal relationships from the correlative. In the rest of this chapter, we learn methods that can, astoundingly, tease out causal information from purely observational data, assuming of course that the graph constitutes a valid representation of reality. It is worth noting here that we are making a tacit assumption that the intervention has no "side effects," that is, that *assigning* the value x for the variable X for an individual does not alter subsequent variables in a direct way. For example, being "assigned" a drug might have a different effect on recovery than being forced to take the drug against one's religious objections. When side effects are present, they need to be specified explicitly in the model.

3.2 The Adjustment Formula

The ice cream example represents an extreme case in which the correlation between X and Y was totally spurious from a causal perspective, because there was no causal path from X to Y. Most real-life situations are not so clear-cut. To explore a more realistic situation, let us examine Figure 3.3, in which Y responds to both Z and X. Such a model could represent, for example, the first story we encountered for Simpson's paradox, where X stands for drug usage, Y stands for recovery, and Z stands for gender. To find out how effective the drug is in the population, we imagine a hypothetical intervention by which we administer the drug uniformly to the entire population and compare the recovery rate to what would obtain under the complementary intervention, where we prevent everyone from using the drug. Denoting the first intervention by $do(X = 1)$ and the second by $do(X = 0)$, our task is to estimate the difference

$$P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)) \tag{3.1}$$

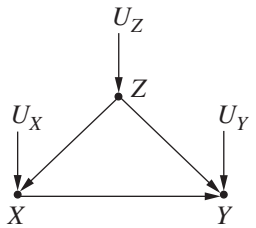

Figure 3.3 A graphical model representing the effects of a new drug, with Z representing gender, X standing for drug usage, and Y standing for recovery

which is known as the "causal effect difference," or "average causal effect" (ACE). In general, however, if X and Y can each take on more than one value, we would wish to predict the general causal effect $P(Y = y|do(X = x))$, where x and y are any two values that X and Y can take on. For example, x may be the dosage of the drug and y the patient's blood pressure.

We know from first principles that causal effects cannot be estimated from the data set itself without a causal story. That was the lesson of Simpson's paradox: The data itself was not sufficient even for determining whether the effect of the drug was positive or negative. But with the aid of the graph in Figure 3.3, we can compute the magnitude of the causal effect from the data. To do so, we simulate the intervention in the form of a graph surgery (Figure 3.4) just as we did in the ice cream example. The causal effect $P(Y = y|do(X = x))$ is equal to the conditional probability $P_m(Y = y|X = x)$ that prevails in the *manipulated* model of Figure 3.4. (This, of course, also resolves the question of whether the correct answer lies in the aggregated or the Z-specific table—when we determine the answer through an intervention, there's only one table to contend with.)

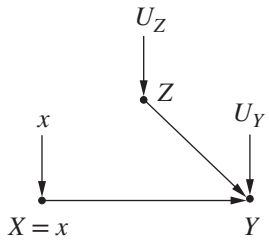

Figure 3.4 A modified graphical model representing an intervention on the model in Figure 3.3 that sets drug usage in the population, and results in the manipulated probability P_m

The key to computing the causal effect lies in the observation that P_m, the manipulated probability, shares two essential properties with P (the original probability function that prevails in the preintervention model of Figure 3.3). First, the marginal probability $P(Z = z)$ is invariant under the intervention, because the process determining Z is not affected by removing the arrow from Z to X. In our example, this means that the proportions of males and females remain the same, before and after the intervention. Second, the conditional probability $P(Y = y|Z = z, X = x)$ is invariant, because the process by which Y responds to X and Z, $Y = f(x, z, u_Y)$, remains the same, regardless of whether X changes spontaneously or by deliberate manipulation. We can therefore write two equations of invariance:

$$P_m(Y = y|Z = z, X = x) = P(Y = y|Z = z, X = x) \quad \text{and} \quad P_m(Z = z) = P(Z = z)$$

We can also use the fact that Z and X are d-separated in the modified model and are, therefore, independent under the intervention distribution. This tells us that $P_m(Z = z|X = x) = P_m(Z = z) = P(Z = z)$, the last equality following from above. Putting these considerations together, we have

$$\begin{aligned} &P(Y = y|do(X = x)) \\ &\qquad = P_m(Y = y|X = x) \qquad \text{(by definition)} \end{aligned} \tag{3.2}$$

$$= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z|X = x) \tag{3.3}$$

$$= \sum_z P_m(Y = y|X = x, Z = z)P_m(Z = z) \tag{3.4}$$

Equation (3.3) is obtained using the Law of Total Probability by conditioning on and summing over all values of $Z = z$ (as in Eq. (1.9)) while Eq. (3.4) makes use of the independence of Z and X in the modified model.

Finally, using the invariance relations, we obtain a formula for the causal effect, in terms of preintervention probabilities:

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \tag{3.5}$$

Equation (3.5) is called the *adjustment formula*, and as you can see, it computes the association between X and Y for each value z of Z, then averages over those values. This procedure is referred to as "adjusting for Z" or "controlling for Z."

This final expression—the right-hand side of Eq. (3.5)—can be estimated directly from the data, since it consists only of conditional probabilities, each of which can be computed by the filtering procedure described in Chapter 1. Note also that no adjustment is needed in a randomized controlled experiment since, in such a setting, the data are generated by a model which already possesses the structure of Figure 3.4, hence, $P_m = P$ regardless of any factors Z that affect Y. Our derivation of the adjustment formula (3.5) constitutes therefore a formal proof that randomization gives us the quantity we seek to estimate, namely $P(Y = y|do(X = x))$. In practice, investigators use adjustments in randomized experiments as well, for the purpose of minimizing sampling variations (Cox 1958).

To demonstrate the working of the adjustment formula, let us apply it numerically to Simpson's story, with $X = 1$ standing for the patient taking the drug, $Z = 1$ standing for the patient being male, and $Y = 1$ standing for the patient recovering. We have

$$P(Y = 1|do(X = 1)) = P(Y = 1|X = 1, Z = 1)P(Z = 1) + P(Y = 1|X = 1, Z = 0)P(Z = 0)$$

Substituting the figures given in Table 1.1 we obtain

$$P(Y = 1|do(X = 1)) = \frac{0.93(87 + 270)}{700} + \frac{0.73(263 + 80)}{700} = 0.832$$

while, similarly,

$$P(Y = 1|do(X = 0)) = \frac{0.87(87 + 270)}{700} + \frac{0.69(263 + 80)}{700} = 0.7818$$

Thus, comparing the effect of drug-taking ($X = 1$) to the effect of nontaking ($X = 0$), we obtain

$$ACE = P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)) = 0.832 - 0.7818 = 0.0502$$

giving a clear positive advantage to drug-taking. A more informal interpretation of ACE here is that it is simply the difference in the fraction of the population that would recover if everyone took the drug compared to when no one takes the drug.

We see that the adjustment formula instructs us to condition on gender, find the benefit of the drug separately for males and females, and only then average the result using the percentage of males and females in the population. It also thus instructs us to ignore the aggregated

population data $P(Y = 1|X = 1)$ and $P(Y = 1|X = 0)$, from which we might (falsely) conclude that the drug has a negative effect overall.

These simple examples might give readers the impression that whenever we face the dilemma of whether to condition on a third variable Z, the adjustment formula prefers the Z-specific analysis over the nonspecific analysis. But we know this is not so, recalling the blood pressure example of Simpson's paradox given in Table 1.2. There we argued that the more sensible method would be not to condition on blood pressure, but to examine the unconditional population table directly. How would the adjustment formula cope with situations like that?

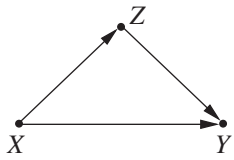

Figure 3.5 A graphical model representing the effects of a new drug, with X representing drug usage, Y representing recovery, and Z representing blood pressure (measured at the end of the study). Exogenous variables are not shown in the graph, implying that they are mutually independent

The graph in Figure 3.5 represents the causal story in the blood pressure example. It is the same as Figure 3.4, but with the arrow between X and Z reversed, reflecting the fact that the treatment has an effect on blood pressure and not the other way around. Let us try now to evaluate the causal effect $P(Y = 1|do(X = 1))$ associated with this model as we did with the gender example. First, we simulate an intervention and then examine the adjustment formula that emanates from the simulated intervention. In graphical models, an intervention is simulated by severing all arrows that enter the manipulated variable X. In our case, however, the graph of Figure 3.5 shows no arrow entering X, since X has no parents. This means that no surgery is required; the conditions under which data were obtained were such that treatment was assigned "as if randomized." If there was a factor that would make subjects prefer or reject treatment, such a factor should show up in the model; the absence of such a factor gives us the license to treat X as a randomized treatment.

Under such conditions, the intervention graph is equal to the original graph—no arrow need be removed—and the adjustment formula reduces to

$$P(Y = y|do(X = x)) = P(Y = y|X = x),$$

which can be obtained from our adjustment formula by letting the empty set be the element adjusted for. Obviously, if we were to adjust for blood pressure, we would obtain an incorrect assessment—one corresponding to a model in which blood pressure causes people to seek treatment.

3.2.1 To Adjust or not to Adjust?

We are now in a position to understand what variable, or set of variables, Z can legitimately be included in the adjustment formula. The intervention procedure, which led to the adjustment formula, dictates that Z should coincide with the parents of X, because it is the influence of

these parents that we neutralize when we fix X by external manipulation. Denoting the parents of X by $PA(X)$, we can therefore write a general adjustment formula and summarize it in a rule:

Rule 1 (The Causal Effect Rule) *Given a graph G in which a set of variables PA are designated as the parents of X, the causal effect of X on Y is given by*

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, PA = z)P(PA = z) \tag{3.6}$$

where z ranges over all the combinations of values that the variables in PA can take.

If we multiply and divide the summand in (3.6) by the probability $P(X = x|PA = z)$, we get a more convenient form:

$$P(y|do(x)) = \sum_z \frac{P(X = x, Y = y, PA = z)}{P(X = x|PA = z)} \tag{3.7}$$

which explicitly displays the role played by the parents of X in predicting the results of interventions. The factor $P(X = x|PA = z)$ is known as the "propensity score" and the advantages of expressing $P(y|do(x))$ in this form will be discussed in Section 3.5.

We can appreciate now what role the causal graph plays in resolving Simpson's paradox, and, more generally, what aspects of the graph allow us to predict causal effects from purely statistical data. We need the graph in order to determine the identity of X's parents—the set of factors that, under nonexperimental conditions, would be sufficient for determining the value of X, or the probability of that value.

This result alone is astounding; using graphs and their underlying assumptions, we were able to identify causal relationships in purely observational data. But, from this discussion, readers may be tempted to conclude that the role of graphs is fairly limited; once we identify the parents of X, the rest of the graph can be discarded, and the causal effect can be evaluated mechanically from the adjustment formula. The next section shows that things may not be so simple. In most practical cases, the set of X's parents will contain unobserved variables that would prevent us from calculating the conditional probabilities in the adjustment formula. Luckily, as we will see in future sections, we can adjust for other variables in the model to substitute for the unmeasured elements of $PA(X)$.

Study questions

Study questions 3.2.1

Referring to Study question 1.5.2 (Figure 1.10) and the parameters listed therein,

(a) Compute $P(y|do(x))$ for all values of x and y, by simulating the intervention $do(x)$ on the model.
(b) Compute $P(y|do(x))$ for all values of x and y, using the adjustment formula (3.5)
(c) Compute the ACE

$$ACE = P(y_1|do(x_1)) - P(y_1|do(x_0))$$

and compare it to the Risk Difference

$$RD = P(y_1|x_1) - P(y_1|x_0)$$

What is the difference between ACE and the RD? What values of the parameters would minimize the difference?

(d) *Find a combination of parameters that exhibit Simpson's reversal (as in Study question 1.5.2(c)) and show explicitly that the overall causal effect of the drug is obtained from the desegregated data.*

3.2.2 Multiple Interventions and the Truncated Product Rule

In deriving the adjustment formula, we assumed an intervention on a single variable, X, whose parents were disconnected, so as to simulate the absence of their influence after intervention. However, social and medical policies occasionally involve multiple interventions, such as those that dictate the value of several variables simultaneously, or those that control a variable over time. To represent multiple interventions, it is convenient to resort to the product decomposition that a graphical model imposes on joint distributions, as we have discussed in Section 1.5.2. According to the Rule of Product Decomposition, the preintervention distribution in the model of Figure 3.3 is given by the product

$$P(x, y, z) = P(z)P(x|z)P(y|x, z) \tag{3.8}$$

whereas the postintervention distribution, governed by the model of Figure 3.4 is given by the product

$$P(z, y|do(x)) = P_m(z)P_m(y|x, z) = P(z)P(y|x, z) \tag{3.9}$$

with the factor $P(x|z)$ purged from the product, since X becomes parentless as it is fixed at $X = x$. This coincides with the adjustment formula, because to evaluate $P(y|do(x))$ we need to marginalize (or sum) over z, which gives

$$P(y|do(x)) = \sum_z P(z)P(y|x, z)$$

in agreement with (3.5).

This consideration also allows us to generalize the adjustment formula to multiple interventions, that is, interventions that fix the values of a set of variables X to constants. We simply write down the product decomposition of the preintervention distribution, and strike out all factors that correspond to variables in the intervention set X. Formally, we write

$$P(x_1, x_2, \ldots, x_n|do(x)) = \prod_i P(x_i|pa_i) \qquad \text{for all } i \text{ with } X_i \text{ not in } X.$$

This came to be known as the *truncated product formula* or g-formula. To illustrate, assume that we intervene on the model of Figure 2.9 and set X to x and Z_3 to z_3. The postintervention distribution of the other variables in the model will be

$$P(z_1, z_2, w, y|do(X = x, Z_3 = z_3)) = P(z_1)P(z_2)P(w|x)P(y|w, z_3, z_2)$$

where we have deleted the factors $P(x|z_1, z_3)$ and $P(z_3|z_1, z_2)$ from the product.

It is interesting to note that combining (3.8) and (3.9), we get a simple relation between the pre- and postintervention distributions:

$$P(z, y|do(x)) = \frac{P(x, y, z)}{P(x|z)} \tag{3.10}$$

It tells us that the conditional probability $P(x|z)$ is all we need to know in order to predict the effect of an intervention $do(x)$ from nonexperimental data governed by the distribution $P(x, y, z)$.

3.3 The Backdoor Criterion

In the previous section, we came to the conclusion that we should adjust for a variable's parents, when trying to determine its effect on another variable. But often, we know, or believe, that the variables have unmeasured parents that, though represented in the graph, may be inaccessible for measurement. In those cases, we need to find an alternative set of variables to adjust for.

This dilemma unlocks a deeper statistical question: Under what conditions does a causal story permit us to compute the causal effect of one variable on another, from data obtained by passive observations, with no interventions? Since we have decided to represent causal stories with graphs, the question becomes a graph-theoretical problem: Under what conditions is the structure of the causal graph sufficient for computing a causal effect from a given data set?

The answer to that question is long enough—and important enough—that we will spend the rest of the chapter addressing it. But one of the most important tools we use to determine whether we can compute a causal effect is a simple test called the *backdoor criterion*. Using it, we can determine, for any two variables X and Y in a causal model represented by a DAG, which set of variables Z in that model should be conditioned on when searching for the causal relationship between X and Y.

Definition 3.3.1 (The Backdoor Criterion) *Given an ordered pair of variables (X, Y) in a directed acyclic graph G, a set of variables Z satisfies the* backdoor criterion *relative to (X, Y) if no node in Z is a descendant of X, and Z blocks every path between X and Y that contains an arrow into X.*

If a set of variables Z satisfies the backdoor criterion for X and Y, then the causal effect of X on Y is given by the formula

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

just as when we adjust for $PA(X)$. (Note that $PA(X)$ always satisfies the backdoor criterion.)

The logic behind the backdoor criterion is fairly straightforward. In general, we would like to condition on a set of nodes Z such that

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

When trying to find the causal effect of X on Y, we want the nodes we condition on to block any "backdoor" path in which one end has an arrow into X, because such paths may make X and Y dependent, but are obviously not transmitting causal influences from X, and if we do not block them, they will confound the effect that X has on Y. We condition on backdoor paths so as to fulfill our first requirement. However, we don't want to condition on any nodes that are descendants of X. Descendants of X would be affected by an intervention on X and might themselves affect Y; conditioning on them would block those pathways. Therefore, we don't condition on descendants of X so as to fulfill our second requirement. Finally, to comply with the third requirement, we should refrain from conditioning on any collider that would unblock a new path between X and Y. The requirement of excluding descendants of X also protects us from conditioning on children of intermediate nodes between X and Y (e.g., the collision node W in Figure 2.4.) Such conditioning would distort the passage of causal association between X and Y, similar to the way conditioning on their parents would.

To see what this means in practice, let's look at a concrete example, shown in Figure 3.6.

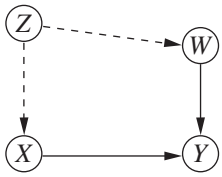

Figure 3.6 A graphical model representing the relationship between a new drug (X), recovery (Y), weight (W), and an unmeasured variable Z (socioeconomic status)

Here we are trying to gauge the effect of a drug (X) on recovery (Y). We have also measured weight (W), which has an effect on recovery. Further, we know that socioeconomic status (Z) affects both weight and the choice to receive treatment—but the study we are consulting did not record socioeconomic status.

Instead, we search for an observed variable that fits the backdoor criterion from X to Y. A brief examination of the graph shows that W, which is not a descendant of X, also blocks the backdoor path $X \leftarrow Z \rightarrow W \rightarrow Y$. Therefore, W meets the backdoor criterion. So long as the causal story conforms to the graph in Figure 3.6, adjusting for W will give us the causal effect of X on Y. Using the adjustment formula, we find

$$P(Y = y|do(X = x)) = \sum_{w} P(Y = y|X = x, W = w)P(W = w)$$

This sum can be estimated from our observational data, so long as W is observed.

With the help of the backdoor criterion, you can easily and algorithmically come to a conclusion about a pressing policy concern, even in complicated graphs. Consider the model in Figure 2.8, and assume again that we wish to evaluate the effect of X on Y. What variables should we condition on to obtain the correct effect? The question boils down to finding a set of variables that satisfy the backdoor criterion, but since there are no backdoor paths from X to Y, the answer is trivial: The empty set satisfies the criterion, hence no adjustment is needed. The answer is

$$P(y|do(x)) = P(y|x)$$

Suppose, however, that we were to adjust for W. Would we get the correct result for the effect of X on Y? Since W is a collider, conditioning on W would open the path $X \rightarrow W \leftarrow Z \leftarrow$

$T \rightarrow Y$. This path is spurious since it lies outside the causal pathway from X to Y. Opening this path will create bias and yield an erroneous answer. This means that computing the association between X and Y for each value of W separately will not yield the correct effect of X on Y, and it might even give the wrong effect for each value of W.

How then do we compute the causal effect of X on Y for a specific value w of W? In Figure 2.8, W may represent, for example, the level of posttreatment pain of a patient, and we might be interested in assessing the effect of X on Y for only those patients who did not suffer any pain. Specifying the value of W amounts to conditioning on $W = w$, and this, as we have realized, opens a spurious path from X to Y by virtue of the fact that W is a collider.

The answer is that we still have the option of blocking that path using other variables. For example, if we condition on T, we would block the spurious path $X \rightarrow W \leftarrow Z \leftarrow T \rightarrow Y$, even if W is part of the conditioning set. Thus to compute the w-specific causal effect, written $P(y|do(x), w)$, we adjust for T, and obtain

$$P(Y = y|do(X = x), W = w) = \sum_t P(Y = y|X = x, W = w, T = t)P(T = t|X = x, W = w) \quad (3.11)$$

Computing such W-specific causal effects is an essential step in examining *effect modification* or *moderation*, that is, the degree to which the causal effect of X on Y is modified by different values of W. Consider, again, the model in Figure 3.6, and suppose we wish to test whether the causal effect for units at level $W = w$ is the same as for units at level $W = w'$ (W may represent any pretreatment variable, such as age, sex, or ethnicity). This question calls for comparing two causal effects,

$$P(Y = y|do(X = x), W = w) \quad \text{and} \quad P(Y = y|do(X = x), W = w')$$

In the specific example of Figure 3.6, the answer is simple, because W satisfies the backdoor criterion. So, all we need to compare are the conditional probabilities $P(Y = y|X = x, W = w)$ and $P(Y = y|X = x, W = w')$; no summation is required. In the more general case, where W alone does not satisfy the backdoor criterion, yet a larger set, $T \cup W$, does, we need to adjust for members of T, which yields Eq. (3.11). We will return to this topic in Section 3.5.

From the examples seen thus far, readers may get the impression that one should refrain from adjusting for colliders. Such adjustment is sometimes unavoidable, as seen in Figure 3.7. Here, there are four backdoor paths from X to Y, all traversing variable Z, which is a collider on the path $X \leftarrow E \rightarrow Z \leftarrow A \rightarrow Y$. Conditioning on Z will unblock this path and will violate the backdoor criterion. To block all backdoor paths, we need to condition on one of the following sets: $\{E, Z\}$, $\{A, Z\}$, or $\{E, Z, A\}$. Each of these contains Z. We see, therefore, that Z, a collider, must be adjusted for in any set that yields an unbiased estimate of the effect of X on Y.

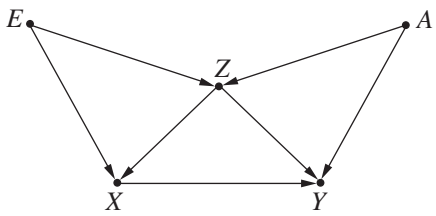

Figure 3.7 A graphical model in which the backdoor criterion requires that we condition on a collider (Z) in order to ascertain the effect of X on Y

The backdoor criterion has some further possible benefits. Consider the fact that $P(Y = y|do(X = x))$ is an empirical fact of nature, not a byproduct of our analysis. That means that any suitable variable or set of variables that we adjust on—whether it be $PA(X)$ or any other set that conforms to the backdoor criterion—must return the same result for $P(Y = y|do(X = x))$. In the case we looked at in Figure 3.6, this means that

$$\sum_w P(Y = y|X = x, W = w)P(W = w) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

This equality is useful in two ways. First, in the cases where we have multiple observed sets of variables suitable for adjustment (e.g., in Figure 3.6, if both W and Z had been observed), it provides us with a choice of which variables to adjust for. This could be useful for any number of practical reasons—perhaps one set of variables is more expensive to measure than the other, or more prone to human error, or simply has more variables and is therefore more difficult to calculate.

Second, the equality constitutes a testable constraint on the data when all the adjustment variables are observed, much like the rules of d-separation. If we are attempting to fit a model that leads to such an equality on a data set that violates it, we can discard that model.

Study questions

Study question 3.3.1

Consider the graph in Figure 3.8:

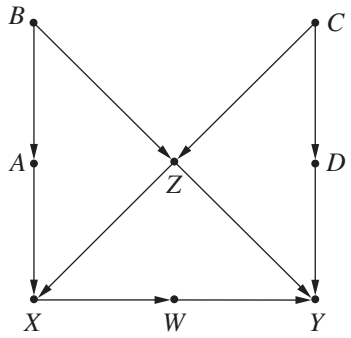

Figure 3.8 Causal graph used to illustrate the backdoor criterion in the following study questions

(a) List all of the sets of variables that satisfy the backdoor criterion to determine the causal effect of X on Y.
(b) List all of the minimal sets of variables that satisfy the backdoor criterion to determine the causal effect of X on Y (i.e., any set of variables such that, if you removed any one of the variables from the set, it would no longer meet the criterion).
(c) List all minimal sets of variables that need be measured in order to identify the effect of D on Y. Repeat, for the effect of $\{W, D\}$ on Y.

Study question 3.3.2 (Lord's paradox)

At the beginning of the year, a boarding school offers its students a choice between two meal plans for the year: Plan A and Plan B. The students' weights are recorded at the beginning and the end of the year. To determine how each plan affects students' weight gain, the school hired two statisticians who, oddly, reached different conclusions. The first statistician calculated the difference between each student's weight in June (W_F) and in September (W_I) and found that the average weight gain in each plan was zero.

The second statistician divided the students into several subgroups, one for each initial weight, W_I. He finds that for each initial weight, the final weight for Plan B is higher than the final weight for Plan A.

So, the first statistician concluded that there was no effect of diet on weight gain and the second concluded there was.

Figure 3.9 illustrates data sets that can cause the two statisticians to reach conflicting conclusions. Statistician-1 examined the weight gain $W_F - W_I$, which, for each student, is represented by the shortest distance to the 45° line. Indeed, the average gain for each diet plan is zero; the two groups are each situated symmetrically relative to the zero-gain line, $W_F = W_I$. Statistician-2, on the other hand, compared the final weights of plan A students to those of plan B students who entered school with the same initial weight W_0 and, as the vertical line in the figure indicates, plan B students are situated above plan A students along this vertical line. The same will be the case for any other vertical line, regardless of W_0.

(a) Draw a causal graph representing the situation.
(b) Determine which statistician is correct.
(c) How is this example related to Simpson's paradox?

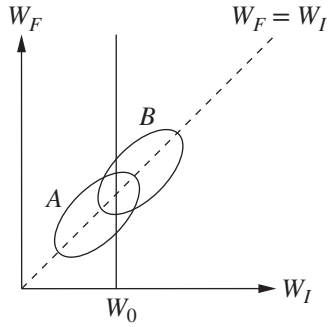

Figure 3.9 Scatter plot with students' initial weights on the *x*-axis and final weights on the *y*-axis. The vertical line indicates students whose initial weights are the same, and whose final weights are higher (on average) for plan B compared with plan A

Study questions 3.3.3

Revisit the lollipop story of Study question 1.2.4 and answer the following questions:

(a) Draw a graph that captures the story.
(b) Determine which variables must be adjusted for by applying the backdoor criterion.

(c) Write the adjustment formula for the effect of the drug on recovery.
(d) Repeat questions (a)–(c) assuming that the nurse gave lollipops a day after the study, still preferring patients who received treatment over those who received placebo.

3.4 The Front-Door Criterion

The backdoor criterion provides us with a simple method of identifying sets of covariates that should be adjusted for when we seek to estimate causal effects from nonexperimental data. It does not, however, exhaust *all* ways of estimating such effects. The *do*-operator can be applied to graphical patterns that do not satisfy the backdoor criterion to identify effects that on first sight seem to be beyond one's reach. One such pattern, called front-door, is discussed in this section.

Consider the century-old debate on the relation between smoking and lung cancer. In the years preceding 1970, the tobacco industry managed to prevent antismoking legislation by promoting the theory that the observed correlation between smoking and lung cancer could be explained by some sort of carcinogenic genotype that also induces an inborn craving for nicotine.

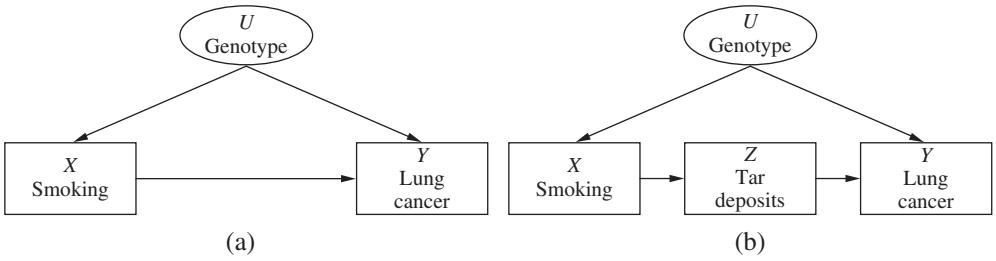

Figure 3.10 A graphical model representing the relationships between smoking (X) and lung cancer (Y), with unobserved confounder (U) and a mediating variable Z

A graph depicting this example is shown in Figure 3.10(a) This graph does not satisfy the backdoor condition because the variable U is unobserved and hence cannot be used to block the backdoor path from X to Y. The causal effect of smoking on lung cancer is not identifiable in this model; one can never ascertain which portion of the observed correlation between X and Y is spurious, attributable to their common effect, U, and what portion is genuinely causative. (We note, however, that even in these circumstances, much compelling work has been done to quantify how strong the (unobserved) associates between both U and X, and U and Y, must be in order to entirely explain the observed association between X and Y.)

However, we can go much further by considering the model in Figure 3.10(b), where an additional measurement is available: the amount of tar deposits in patients' lungs. This model does not satisfy the backdoor criterion, because there is still no variable capable of blocking the spurious path $X \leftarrow U \rightarrow Y$. We see, however, that the causal effect $P(Y = y|do(X = x))$ is nevertheless identifiable in this model, through two consecutive applications of the backdoor criterion.

How can the intermediate variable Z help us to assess the effect of X on Y? The answer is not at all trivial: as the following quantitative example shows, it may lead to heated debate.

Assume that a careful study was undertaken, in which the following factors were measured simultaneously on a randomly selected sample of 800,000 subjects considered to be at very high risk of cancer (because of environmental exposures such as smoking, asbestos, radon, and the like).

1. Whether the subject smoked
2. Amount of tar in the subject's lungs
3. Whether lung cancer has been detected in the patient.

The data from this study are presented in Table 3.1, where, for simplicity, all three variables are assumed to be binary. All numbers are given in thousands.

Table 3.1 A hypothetical data set of randomly selected samples showing the percentage of cancer cases for smokers and nonsmokers in each tar category (numbers in thousands)

	Tar 400		No tar 400		All subjects 800	
	Smokers	Nonsmokers	Smokers	Nonsmokers	Smokers	Nonsmokers
	380	*20*	*20*	*380*	*400*	*400*
No cancer	323	1	18	38	341	39
	(85%)	(5%)	(90%)	(10%)	(85%)	(9.75%)
Cancer	57	19	2	342	59	361
	(15%)	(95%)	(10%)	(90%)	(15%)	(90.25%)

Two opposing interpretations can be offered for these data. The tobacco industry argues that the table proves the beneficial effect of smoking. They point to the fact that only 15% of the smokers have developed lung cancer, compared to 90.25% of the nonsmokers. Moreover, within each of two subgroups, tar and no tar, smokers show a much lower percentage of cancer than nonsmokers. (These numbers are obviously contrary to empirical observations but well illustrate our point that observations are not to be trusted.)

However, the antismoking lobbyists argue that the table tells an entirely different story—that smoking would actually increase, not decrease, one's risk of lung cancer. Their argument goes as follows: If you choose to smoke, then your chances of building up tar deposits are 95%, compared to 5% if you choose not to smoke (380/400 vs 20/400). To evaluate the effect of tar deposits, we look separately at two groups, smokers and nonsmokers, as done in Table 3.2. All numbers are given in thousands.

Table 3.2 Reorganization of the data set of Table 3.1 showing the percentage of cancer cases in each smoking-tar category (numbers in thousands)

	Smokers 400		Nonsmokers 400		All subjects 800	
	Tar	No tar	Tar	No tar	Tar	No tar
	380	*20*	*20*	*380*	*400*	*400*
No cancer	323	18	1	38	324	56
	(85%)	(90%)	(5%)	(10%)	(81%)	(19%)
Cancer	57	2	19	342	76	344
	(15%)	(10%)	(95%)	(90%)	(19%)	(81%)

It appears that tar deposits have a harmful effect in both groups; in smokers it increases cancer rates from 10% to 15%, and in nonsmokers it increases cancer rates from 90% to 95%. Thus, regardless of whether I have a natural craving for nicotine, I should avoid the harmful effect of tar deposits, and no-smoking offers very effective means of avoiding them.

The graph of Figure 3.10(b) enables us to decide between these two groups of statisticians. First, we note that the effect of X on Z is identifiable, since there is no backdoor path from X to Z. Thus, we can immediately write

$$P(Z = z|do(X = x)) = P(Z = z|X = x) \tag{3.12}$$

Next we note that the effect of Z on Y is also identifiable, since the backdoor path from Z to Y, namely $Z \leftarrow X \leftarrow U \rightarrow Y$, can be blocked by conditioning on X. Thus, we can write

$$P(Y = y|do(Z = z)) = \sum_x P(Y = y|Z = z, X = x)P(X = x) \tag{3.13}$$

Both (3.12) and (3.13) are obtained through the adjustment formula, the first by conditioning on the null set, and the second by adjusting for X.

We are now going to chain together the two partial effects to obtain the overall effect of X on Y. The reasoning goes as follows: If nature chooses to assign Z the value z, then the probability of Y would be $P(Y = y|do(Z = z))$. But the probability that nature would choose to do that, given that we choose to set X at x, is $P(Z = z|do(X = x))$. Therefore, summing over all states z of Z, we have

$$P(Y = y|do(X = x)) = \sum_z P(Y = y|do(Z = z))P(Z = z|do(X = x)) \tag{3.14}$$

The terms on the right-hand side of (3.14) were evaluated in (3.12) and (3.13), and we can substitute them to obtain a *do*-free expression for $P(Y = y|do(X = x))$. We also distinguish between the x that appears in (3.12) and the one that appears in (3.13), the latter of which is merely an index of summation and might as well be denoted x'. The final expression we have is

$$P(Y = y|do(X = x)) = \sum_z \sum_{x'} P(Y = y|Z = z, X = x')P(X = x')P(Z = z|X = x) \tag{3.15}$$

Equation (3.15) is known as the *front-door formula*.

Applying this formula to the data in Table 3.1, we see that the tobacco industry was wrong; tar deposits have a harmful effect in that they make lung cancer more likely and smoking, by increasing tar deposits, increases the chances of causing this harm.

The data in Table 3.1 are obviously unrealistic and were deliberately crafted so as to surprise readers with counterintuitive conclusions that may emerge from naive analysis of observational data. In reality, we would expect observational studies to show positive correlation between smoking and lung cancer. The estimand of (3.15) could then be used for confirming and quantifying the harmful effect of smoking on cancer.

The preceding analysis can be generalized to structures where multiple paths lead from X to Y.

Definition 3.4.1 (Front-Door) *A set of variables Z is said to satisfy the front-door criterion relative to an ordered pair of variables* (X, Y) *if*

1. *Z intercepts all directed paths from X to Y.*
2. *There is no backdoor path from X to Z.*
3. *All backdoor paths from Z to Y are blocked by X.*

Theorem 3.4.1 (Front-Door Adjustment) *If Z satisfies the front-door criterion relative to* (X, Y) *and if* $P(x, z) > 0$*, then the causal effect of X on Y is identifiable and is given by the formula*

$$P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \tag{3.16}$$

The conditions stated in Definition 3.4.1 are overly conservative; some of the paths excluded by conditions (2) and (3) can actually be allowed provided they are blocked by some variables. There is a powerful symbolic machinery, called the *do-calculus*, that allows analysis of such intricate structures. In fact, the *do*-calculus uncovers *all* causal effects that can be identified from a given graph. Unfortunately, it is beyond the scope of this book (see Tian and Pearl 2002, Shpitser and Pearl 2008, Pearl 2009, and Bareinboim and Pearl 2012 for details). But the combination of the adjustment formula, the backdoor criterion, and the front-door criterion covers numerous scenarios. It proves the enormous, even revelatory, power that causal graphs have in not merely representing, but actually discovering causal information.

Study questions

Study question 3.4.1

Assume that in Figure 3.8, only X, Y*, and one additional variable can be measured. Which variable would allow the identification of the effect of X on Y? What would that effect be?*

Study question 3.4.2

I went to a pharmacy to buy a certain drug, and I found that it was available in two different bottles: one priced at $1, the other at $10. I asked the druggist, "What's the difference?" and he told me, "The $10 bottle is fresh, whereas the $1 bottle one has been on the shelf for 3 years. But, you know, data shows that the percentage of recovery is much higher among those who bought the cheap stuff. Amazing isn't it?" I asked if the aged drug was ever tested. He said, "Yes, and this is even more amazing; 95% of the aged drug and only 5% of the fresh drug has lost the active ingredient, yet the percentage of recovery among those who got bad bottles, with none of the active ingredient, is still much higher than among those who got good bottles, with the active ingredient."

Before ordering a cheap bottle, it occurred to me to have a good look at the data. The data were, for each previous customer, the type of bottle purchased (aged or fresh), the concentration of the active ingredient in the bottle (high or low), and whether the customer recovered from the illness. The data perfectly confirmed the druggist's story. However, after making some additional calculations, I decided to buy the expensive bottle after all; even without testing its

content, I could determine that a fresh bottle would offer the average patient a greater chance of recovery.

Based on two very reasonable assumptions, the data show clearly that the fresh drug is more effective. The assumptions are as follows:

(i) Customers had no information about the chemical content (high or low) of the specific bottle of the drug that they were buying; their choices were influenced by price and shelf-age alone.
(ii) The effect of the drug on any given individual depends only on its chemical content, not on its shelf age (fresh or aged).

(a) Determine the relevant variables for the problem, and describe this scenario in a causal graph.
(b) Construct a data set compatible with the story and the decision to buy the expensive bottle.
(c) Determine the effect of choosing the fresh versus the aged drug by using assumptions (i) and (ii), and the data given in (b).

3.5 Conditional Interventions and Covariate-Specific Effects

The interventions considered thus far have been limited to actions that merely force a variable or a group of variables X to take on some specified value x. In general, interventions may involve dynamic policies in which a variable X is made to respond in a specified way to some set Z of other variables—say, through a functional relationship $x = g(z)$ or through a stochastic relationship, whereby X is set to x with probability $P^*(x|z)$. For example, suppose a doctor decides to administer a drug only to patients whose temperature Z exceeds a certain level, $Z = z$. In this case, the action will be *conditional* upon the value of Z and can be written $do(X = g(Z))$, where $g(Z)$ is equal to one when $Z > z$ and zero otherwise (where $X = 0$ represents no drug). Since Z is a random variable, the value of X chosen by the action will similarly be a random variable, tracking variations in Z. The result of implementing such a policy is a probability distribution written $P(Y = y|do(X = g(Z)))$, which depends only on the function g and the set Z of variables that drive X.

In order to estimate the effect of such a policy, let us take a closer look at another concept, the "z-specific effect" of X, which we encountered briefly in Section 3.3 (Eq. (3.11)). This effect, written $P(Y = y|do(X = x), Z = z)$, measures the distribution of Y in a subset of the population for which Z achieves the value z after the intervention. For example, we may be interested in how a treatment affects a specific age group, $Z = z$, or people with a specific feature, $Z = z$, which may be measured after the treatment.

The z-specific effect can be identified by a procedure similar to the backdoor adjustment. The reasoning goes as follows: When we aim to estimate $P(Y = y|do(X = x))$, an adjustment for a set S is justified if S blocks all backdoor paths from X to Y. Now that we wish to identify $P(Y = y|do(X = x), Z = z)$, we need to ensure that those paths remain blocked when we add one more variable, Z, to the conditioning set. This yields a simple criterion for the identification of the z-specific effect:

Rule 2 *The z-specific effect $P(Y = y|do(X = x), Z = z)$ is identified whenever we can measure a set S of variables such that $S \cup Z$ satisfies the backdoor criterion. Moreover, the z-specific*

effect is given by the following adjustment formula

$$P(Y = y|do(X = x), Z = z) = \sum_s P(Y = y|X = x, S = s, Z = z)P(S = s|Z = z)$$

This modified adjustment formula is similar to Eq. (3.5) with two exceptions. First, the adjustment set is $S \cup Z$, not just S and, second, the summation goes only over S, not including Z. The $\cup$ symbol in the expression $S \cup Z$ stands for set addition (or union), which means that, if Z is a subset of S, we have $S \cup Z = S$, and S alone need satisfy the backdoor criterion.

Note that the identifiability criterion for z-specific effects is somewhat stricter than that for nonspecific effect. Adding Z to the conditioning set might create dependencies that would prevent the blocking of all backdoor paths. A simple example occurs when Z is a collider; conditioning on Z will create a new dependency between Z's parents and may thus violate the backdoor requirement.

We are now ready to tackle our original task of estimating conditional interventions. Suppose a policy maker contemplates an age-dependent policy whereby an amount x of drug is to be administered to patients, depending on their age Z. We write it as $do(X = g(Z))$. To find out the distribution of outcome Y that results from this policy, we seek to estimate $P(Y = y|do(X = g(Z)))$.

We now show that identifying the effect of such policies is equivalent to identifying the expression for the z-specific effect $P(Y = y|do(X = x), Z = z)$.

To compute $P(Y = y|do(X = g(Z)))$, we condition on $Z = z$ and write

$$\begin{aligned} P(Y = y|do(X = g(Z))) &= \sum_z P(Y = y|do(X = g(Z)), Z = z)P(Z = z|do(X = g(Z))) \\ &= \sum_z P(Y = y|do(X = g(z)), Z = z)P(Z = z) \end{aligned} \tag{3.17}$$

The equality

$$P(Z = z|do(X = g(Z))) = P(Z = z)$$

stems, of course, from the fact that Z occurs before X; hence, any control exerted on X can have no effect on the distribution of Z. Equation (3.17) can also be written as

$$\sum_z P(Y = y|do(X = x), Z = z)|_{x=g(z)}P(Z = z)$$

which tells us that the causal effect of a conditional policy $do(X = g(Z))$ can be evaluated directly from the expression of $P(Y = y|do(X = x), Z = z)$ simply by substituting $g(z)$ for x and taking the expectation over Z (using the observed distribution $P(Z = z)$).

Study question 3.5.1

Consider the causal model of Figure 3.8.

(a) Find an expression for the c-specific effect of X on Y.

(b) Identify a set of four variables that need to be measured in order to estimate the z-specific effect of X on Y, and find an expression for the size of that effect.

(c) Using your answer to part (b), determine the expected value of Y under a Z-dependent strategy, where X is set to 0 when Z is smaller or equal to 2 and X is set to 1 when Z is larger than 2. (Assume Z takes on integer values from 1 to 5.)

3.6 Inverse Probability Weighing

By now, the astute reader may have noticed a problem with our intervention procedures. The backdoor and front-door criteria tell us whether it is possible to predict the results of hypothetical interventions from data obtained in an observational study. Moreover, they tell us that we can make this prediction without simulating the intervention and without even thinking about it. All we need to do is identify a set Z of covariates satisfying one of the criteria, plug this set into the adjustment formula, and we're done: the resulting expression is guaranteed to provide a valid prediction of how the intervention will affect the outcome.

This is lovely in theory, but in practice, adjusting for Z may prove problematic. It entails looking at each value or combination of values of Z separately, estimating the conditional probability of Y given X in that stratum and then averaging the results. As the number of strata increases, adjusting for Z will encounter both computational and estimational difficulties. Since the set Z can be comprised of dozens of variables, each spanning dozens of discrete values, the summation required by the adjustment formula may be formidable, and the number of data samples falling within each $Z = z$ cell may be too small to provide reliable estimates of the conditional probabilities involved.

All of our work in this chapter has not been for naught, however. The adjustment procedure is straightforward, and, therefore, easy to use in the explanation of intervention criteria. But there is another, more subtle procedure that overcomes the practical difficulties of adjustment.

In this section, we discuss one way of circumventing this problem, provided only that we can obtain a reliable estimate of the function $g(x, z) = P(X = x|Z = z)$, often called the "propensity score," for each x and z. Such an estimate can be obtained by fitting the parameters of a flexible function $g(x, z)$ to the data at hand, in much the same way that we fitted the coefficients of a linear regression function, so as to minimize the mean square error with respect to a set of samples (Figure 1.4). The method used will depend on the nature of the random variable X, whether it is continuous, discrete or binary, for example.

Assuming that the function $P(X = x|Z = z)$ is available to us, we can use it to generate artificial samples that act as though they were drawn from the postintervention probability P_m, rather than $P(x, y, z)$. Once we obtain such fictitious samples, we can evaluate $P(Y = y|do(x))$ by simply counting the frequency of the event $Y = y$, for each stratum $X = x$ in the sample. In this way, we skip the labor associated with summing over all strata $Z = z$; we essentially let nature do the summation for us.

The idea of estimating probabilities using fictitious samples is not new to us; it was used all along, though implicitly, whenever we estimated conditional probabilities from finite samples.

In Chapter 1, we characterized conditioning as a process of filtering—that is, ignoring all cases for which the condition $X = x$ does not hold, and normalizing the surviving cases, so that their total probabilities would add up to one. The net result of this operation is that the probability of each surviving case is boosted by a factor $1/P(X = x)$. This can be seen directly

from Bayes' rule, which tells us that

$$P(Y = y, Z = z|X = x) = \frac{P(Y = y, Z = z, X = x)}{P(X = x)}$$

In other words, to find the probability of each row in the surviving table, we multiply the unconditional probability, $P(Y = y, Z = z, X = x)$ by the constant $1/P(X = x)$.

Let us now examine the population created by the $do(X = x)$ operation and ask how the probability of each case changes as a result of this operation. The answer is given to us by the adjustment formula, which reads

$$P(y|do(x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

Multiplying and dividing the expression inside the sum by the propensity score $P(X = x|Z = z)$, we get

$$P(y|do(x)) = \sum_z \frac{P(Y = y|X = x, Z = z)P(X = x|Z = z)P(Z = z)}{P(X = x|Z = z)}$$

Upon realizing the numerator is none other but the pretreatment distribution of (X, Y, Z), we can write

$$P(y|do(x)) = \sum_z \frac{P(Y = y, X = x, Z = z)}{P(X = x|Z = z)}$$

and the answer becomes clear: each case $(Y = y, X = x, Z = z)$ in the population should boost its probability by a factor equal to $1/P(X = x|Z = z)$. (Hence the name "inverse probability weighing.")

This provides us with a simple procedure of estimating $P(Y = y|do(X = x))$ when we have finite samples. If we weigh each available sample by a factor $= 1/P(X = x|Z = z)$, we can then treat the reweighted samples as if they were generated from P_m, not P, and proceed to estimate $P(Y = y|do(x))$ accordingly.

This is best demonstrated in an example.

Table 3.3 returns to our Simpson's paradox example of the drug that seems to help men and women but to hurt the general population. We'll use the same data we used before but presented

Table 3.3 Joint probability distribution $P(X, Y, Z)$ for the drug-gender-recovery story of Chapter 1 (Table 1.1)

X	Y	Z	% of population
Yes	Yes	Male	0.116
Yes	Yes	Female	0.274
Yes	No	Male	0.009
Yes	No	Female	0.101
No	Yes	Male	0.334
No	Yes	Female	0.079
No	No	Male	0.051
No	No	Female	0.036

Table 3.4 Conditional probability distribution $P(Y, Z|X)$ for drug users ($X = yes$) in the population of Table 3.3

X	Y	Z	% of population
Yes	Yes	Male	0.231
Yes	Yes	Female	0.549
Yes	No	Male	0.017
Yes	No	Female	0.203

this time as a weighted table. In this case, X represents whether or not the patient took the drug, Y represents whether the patient recovered, and Z represents the patient's gender.

If we condition on "$X = Yes$," we get the data set shown in Table 3.4, which was formed in two steps. First, all rows with $X = No$ were excluded. Second, the weights given to the remaining rows were "renormalized," that is, multiplied by a constant so as to make them sum to one. This constant, according to Bayes' rule, is $1/P(X = yes)$, and $P(X = yes)$ in our example, is the combined weight of the first four rows of Table 3.3, which amounts to

$$P(X = yes) = 0.116 + 0.274 + 0.01 + 0.101 = 0.501$$

The result is the weight distribution in the four rows of Table 3.4; the weight of each row has been boosted by a factor $1/0.501 = 2.00$.

Let us now examine the population created by the $do(X = yes)$ operation, representing a deliberate decision to administer the drug to the same population.

To calculate the distribution of weights in this population, we need to compute the factor $P(X = yes|Z = z)$ for each z, which, according to Table 3.3, is given by

$$P(X = yes|Z = Male) = \frac{(0.116 + 0.01)}{(0.116 + 0.01 + 0.334 + 0.051)} = 0.247$$

$$P(X = yes|Z = Female) = \frac{(0.274 + 0.101)}{(0.274 + 0.101 + 0.079 + 0.036)} = 0.765$$

Multiplying the gender-matching rows by $1/0.247$ and $1/0.765$, respectively, we obtain Table 3.5, which represents the postintervention distribution of the population of Table 3.3. The probability of recovery in this distribution can now be computed directly from the data, by summing the first two rows:

$$P(Y = yes|do(X = yes)) = 0.476 + 0.357 = 0.833$$

Table 3.5 Probability distribution for the population of Table 3.3 under the intervention $do(X = Yes)$, determined via the inverse probability method

X	Y	Z	% of population
Yes	Yes	Male	0.475
Yes	Yes	Female	0.358
Yes	No	Male	0.035
Yes	No	Female	0.132

Three points are worth noting about this procedure. First, the redistribution of weight is no longer proportional but quite discriminatory. Row #1, for instance, boosted its weight from 0.116 to 0.476, a factor of 4.1, whereas Row #2 is boosted from 0.274 to 0.357, a factor of only 1.3. This redistribution renders X independent of Z, as in a randomized trial (Figure 3.4).

Second, an astute reader would notice that in this example no computational savings were realized; to estimate $P(Y = yes|do(X = yes))$ we still needed to sum over all values of Z, males and females. Indeed, the savings become significant when the number of Z values is in the thousands or millions, and the sample size is in the hundreds. In such cases, the number of Z values that the inverse probability method would encounter is equal to the number of samples available, not to the number of possible Z values, which is prohibitive.

Finally, an important word of caution. The method of inverse probability weighing is only valid when the set Z entering the factor $1/P(X = x|Z = z)$ satisfies the backdoor criterion. Lacking this assurance, the method may actually introduce more bias than the one obtained through naive conditioning, which produces Table 3.4 and the absurdities of Simpson's paradox.

Up to this point, and in the following, we focus on unbiased estimation of causal effects. In other words, we focus on estimates that will converge to the true causal effects as the number of samples increases indefinitely.

This is obviously important, but it is not the *only* issue relevant to estimation. In addition, we must also address *precision*. Precision refers to the variability of our causal estimates if the number of samples is finite, and, in particular, how much our estimate would vary from experiment to experiment. Clearly, all other things being equal, we prefer estimation procedures with high precision in addition to their possessing little or no bias. Practically, high-precision estimates lead to shorter confidence intervals that quantify our level of certainty as to how our sample estimates describe the causal effect of interest. Most of our discussion does not address the "best," or most precise, way to estimate relevant causal means and effects but focuses on whether it is possible to estimate such quantities from observed data distributions, when the number of samples goes to infinity.

For example, suppose we wish to estimate the causal effect of X on Y (in a causal graph as above), where X and Y both reflect continuous variables. Suppose the effect of Z is to make both high and low values of X most commonly observed, with values close to the middle of the range of X much less common. Then, inverse probability weighting down-weights the extreme values of X on both ends of its range (since these are observed most frequently due to Z) and essentially focuses entirely on the "middle" values of X. If we then use a regression model to estimate the causal effect of X on Y (see Section 3.8, for example) using the reweighed observations to account for the role of Z, the resulting estimates will be very imprecise. In such cases, we usually seek for alternative estimation strategies that are more precise. While we do not pursue these alternatives in this book, it is important to emphasize that, in addition to seeing that causal effects be identified from the data, we must also devise effective strategies of using finite data to estimate effect sizes.

3.7 Mediation

Often, when one variable causes another, it does so both directly and indirectly, through a set of mediating variables. For instance, in our blood pressure/treatment/recovery example of Simpson's paradox, treatment is both a direct (negative) cause of recovery, and an indirect

(positive) cause, through the mediator of blood pressure—treatment decreases blood pressure, which increases recovery. In many cases, it is useful to know how much of variable X's effect on variable Y is direct and how much is mediated. In practice, however, separating these two avenues of causation has proved difficult.

Suppose, for example, we want to know whether and to what degree a company discriminates by gender (X) in its hiring practices (Y). Such discrimination would constitute a direct effect of gender on hiring, which is illegal in many cases. However, gender also affects hiring practices in other ways; often, for instance, women are more or less likely to go into a particular field than men, or to have achieved advanced degrees in that field. So gender may also have an indirect effect on hiring through the mediating variable of qualifications (Z).

In order to find the direct effect of gender on hiring, we need to somehow hold qualifications steady, and measure the remaining relationship between gender and hiring; with qualifications unchanging, any change in hiring would have to be due to gender alone. Traditionally, this has been done by conditioning on the mediating variable. So if $P(Hired|Female, Highly\ Qualified)$ is different from $P(Hired|Male, Highly\ Qualified)$, the reasoning goes, then there is a direct effect of gender on hiring.

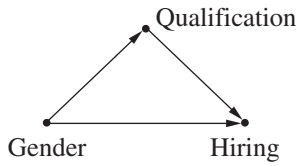

Figure 3.11 A graphical model representing the relationship between gender, qualifications, and hiring

In the example in Figure 3.11, this is correct. But consider what happens if there are confounders of the mediating variable and the outcome variable. For instance, income: People from higher income backgrounds are more likely to have gone to college and more likely to have connections that would help them get hired.

Now, if we condition on qualifications, we are conditioning on a collider. So if we don't condition on qualifications, indirect dependence can pass from gender to hiring through the path $Gender \rightarrow Qualifications \rightarrow Hiring$. But if we do condition on qualifications, indirect dependence can pass from gender to hiring through the path $Gender \rightarrow Qualifications \leftarrow Income \rightarrow Hiring$. (To understand the problem intuitively, note that by conditioning on qualification, we will be comparing men and women at different levels of income, because income must change to keep qualification constant.) No matter how you look at it, we're not getting the true direct effect of gender on hiring. Traditionally, therefore, statistics has had to abandon a huge class of potential mediation problems, where the concept of "direct effect" could not be defined, let alone estimated.

Luckily, we now have a conceptual way of holding the mediating variable steady without conditioning on it: We can intervene on it. If, instead of conditioning, we fix the qualifications, the arrow between gender and qualifications (and the one between income and qualifications) disappears, and no spurious dependence can pass through it. (Of course, it would be impossible for us to literally change the qualifications of applicants, but recall, this is a theoretical intervention of the kind discussed in the previous section, accomplished by choosing a proper adjustment.) So for any three variables X, Y, and Z, where Z is a mediator between X and Y,

the *controlled direct effect* (CDE) on Y of changing the value of X from x to x' is defined as

$$CDE = P(Y = y|do(X = x), do(Z = z)) - P(Y = y|do(X = x'), do(Z = z)) \tag{3.18}$$

The obvious advantage of this definition over the one based on conditioning is its generality; it captures the intent of "keeping Z constant" even in cases where the $Z \to Y$ relationship is confounded (the same goes for the $X \to Z$ and $X \to Y$ relationships). Practically, this definition assures us that in any case where the intervened probabilities are identifiable from the observed probabilities, we can estimate the direct effect of X on Y. Note that the direct effect may differ for different values of Z; for instance, it may be that hiring practices discriminate against women in jobs with high qualification requirements, but they discriminate against men in jobs with low qualifications. Therefore, to get the full picture of the direct effect, we'll have to perform the calculation for every relevant value z of Z. (In linear models, this will not be necessary; for more information, see Section 3.8.)

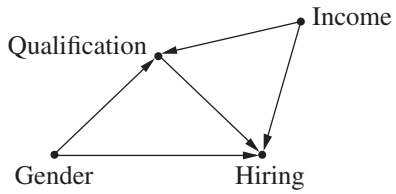

Figure 3.12 A graphical model showing qualification (Z) as a mediator between gender (X) and hiring (Y), and income (I) as a confounder between qualification and hiring.

How do we estimate the direct effect when its expression contains two *do*-operators? The technique is more or less the same as the one employed in Section 3.2, where we dealt with a single *do*-operator by adjustment. In our example of Figure 3.12, we first notice that there is no backdoor path from X to Y in the model, hence we can replace $do(x)$ with simply conditioning on x (this essentially amounts to adjusting for all confounders). This results in

$$P(Y = y|X = x, do(Z = z)) - P(Y = y|X = x', do(Z = z))$$

Next, we attempt to remove the $do(z)$ term and notice that two backdoor paths exist from Z to Y, one through X and one through I. The first is blocked (since X is conditioned on) and the second can be blocked if we adjust for I. This gives

$$\sum_i [P(Y = y|X = x, Z = z, I = i) - P(Y = y|X = x', Z = z, I = i)]P(I = i)$$

The last formula is *do*-free, which means it can be estimated from nonexperimental data.

In general, the CDE of X on Y, mediated by Z, is identifiable if the following two properties hold:

1. There exists a set S_1 of variables that blocks all backdoor paths from Z to Y.
2. There exists a set S_2 of variables that blocks all backdoor paths from X to Y, after deleting all arrows entering Z.

If these two properties hold in a model M, then we can determine $P(Y = y|do(X = x), do(Z = z))$ from the data set by adjusting for the appropriate variables, and estimating the conditional probabilities that ensue. Note that condition 2 is not necessary in randomized trials, because randomizing X renders X parentless. The same is true in cases where X is judged to be exogenous (i.e., "as if" randomized), as in the aforementioned gender discrimination example.

It is even trickier to determine the indirect effect than the direct effect, because there is simply no way to condition away the direct effect of X on Y. It's easy enough to find the total effect and the direct effect, so some may argue that the indirect effect should just be the difference between those two. This may be true in linear systems, but in nonlinear systems, differences don't mean much; the change in Y might, for instance, depend on some interaction between X and Z—if, as we posited above, women are discriminated against in high-qualification jobs and men in low-qualification jobs, subtracting the direct effect from the total effect would tell us very little about the effect of gender on hiring as mediated by qualifications. Clearly, we need a definition of indirect effect that does not depend on the total or direct effects.

We will show in Chapter 4 that these difficulties can be overcome through the use of *counterfactuals*, a more refined type of intervention that applies at the individual level and can be computed from structural models.

3.8 Causal Inference in Linear Systems

One of the advantages of the causal methods we have introduced in this book is that they work regardless of the type of equations that make up the model in question. d-separation and the backdoor criterion make no assumptions about the form of the relationship between two variables—only that the relationship exists.

However, showcasing and explaining causal methods from a nonparametric standpoint has limited our ability to present the full power of these methods as they play out in linear systems—the arena where traditional causal analysis has primarily been conducted in the social and behavioral sciences. This is unfortunate, as many statisticians work extensively in linear systems, and nearly all statisticians are very familiar with them.

In this section, we examine in depth what causal assumptions and implications look like in systems of linear equations and how graphical methods can help us answer causal questions posed in those systems. This will serve as both a reinforcement of the methods we applied in nonparametric models and as a useful aid for those hoping to apply causal inference specifically in the context of linear systems.

For instance, we might want to know the effect of birth control use on blood pressure after adjusting for confounders; the total effect of an after-school study program on test scores; the direct effect, unmediated by other variables, of the program on test scores; or the effect of enrollment in an optional work training program on future earnings, when enrollment and earnings are confounded by a common cause (e.g., motivation). Such questions, invoking continuous variables, have traditionally been formulated as linear equation models with only minor attention to the unique causal character of those equations; we make this character unambiguous.

In all models used in this section, we make the strong assumption that the relationships between variables are linear, and that all error terms have Gaussian (or "normal") distributions (in some cases, we only need to assume symmetric distributions). This assumption provides an

enormous simplification of the procedure needed for causal analysis. We are all familiar with the bell-shaped curve that characterizes the normal distribution of one variable. The reason it is so popular in statistics is that it occurs so frequently in nature whenever a phenomenon is a byproduct of many noisy microprocesses that add up to produce macroscopic measurements such as height, weight, income, or mortality. Our interest in the normal distribution, however, stems primarily from the way several normally distributed variables combine to shape their joint distribution. The assumption of normality gives rise to four properties that are of enormous use when working with linear systems:

1. Efficient representation
2. Substitutability of expectations for probabilities
3. Linearity of expectations
4. Invariance of regression coefficients.

Starting with two normal variables, X and Y, we know that their joint density forms a three-dimensional cusp (like a mountain rising above the X–Y plane) and that the planes of equal height on that cusp are ellipses like those shown in Figure 1.2. Each such ellipse is characterized by five parameters: $\mu_X, \mu_Y, \sigma_X, \sigma_Y$, and ρ_{XY}, as defined in Sections 1.3.8 and 1.3.9. The parameters μ_X and μ_Y specify the location (or the center of gravity) of the ellipse in the X–Y plane, the standard deviations σ_X and σ_Y specify the spread of the ellipse along the X and Y dimensions, respectively, and the correlation coefficient ρ_{XY} specifies its orientation. In three dimensions, the best way to depict the joint distribution is to imagine an oval football sus-pended in the X–Y–Z space (Figure 1.2); every plane of constant Z would then cut the football in a two-dimensional ellipse like the ones shown in Figure 1.1.

As we go to higher dimensions, and consider a set of N normally distributed variables $X_1, X_2, \ldots, X_N$, we need not concern ourselves with additional parameters; it is sufficient to specify those that characterize the $N(N-1)/2$ pairs of variables, (X_i, X_j). In other words, the joint density of $(X_1, X_2, \ldots, X_N)$ is fully specified once we specify the bivariate density of (X_i, X_j), with i and j $(i \neq j)$ ranging from 1 to N. This is an enormously useful property, as it offers an extremely parsimonious way of specifying the N-variable joint distribution. Moreover, since the joint distribution of each pair is specified by five parameters, we conclude that the joint distribution requires at most $5 \times N(N-1)/2$ parameters (means, variances, and covariances), each defined by expectation. In fact, the total number of parameters is even smaller than this, namely $2N + N(N-1)/2$; the first term gives the number of mean and variance parameters, and the second the number of correlations.

This brings us to another useful feature of multivariate normal distributions: they are fully defined by expectations, so we need not concern ourselves with probability tables as we did when dealing with discrete variables. Conditional probabilities can be expressed as conditional expectations, and notions such as conditional independence that define the structure of graphical models can be expressed in terms of equality relationships among conditional expectations. For instance, to express the conditional independence of Y and X, given Z,

$$P(Y|X,Z) = P(Y|Z)$$

we can write

$$E[Y|X,Z] = E[Y|Z]$$

(where Z is a set of variables).

This feature of normal systems gives us an incredibly useful ability: Substituting expectations for probabilities allows us to use regression (a predictive method) to determine causal information. The next useful feature of normal distributions is their linearity: every conditional expectation $E[Y|X_1, X_2, \ldots, X_n]$ is given by a linear combination of the conditioning variables. Formally,

$$E[Y|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = r_0 + r_1x_1 + r_2x_2 + \cdots + r_nx_n$$

where each of the slopes r_1,r_2,...,r_n is a partial regression coefficient as defined in Sections 1.3.10 and 1.3.11.

The magnitudes of these slopes do not depend on the values $x_1, x_2, \ldots, x_n$ of the conditioning variables, called *regressors*; they depend only on which variables are chosen as regressors. In other words, the sensitivity of Y to the measurement $X_i = x_i$ does not depend on the measured values of the other variables in the regression; it depends only on which variables we choose to measure. It doesn't matter whether X_i =1, X_i =2, or X_i =312.3; as long as we regress Y on X_1, X_2,..., X_n all slopes will remain the same.

This unique and useful feature of normal distributions is illustrated in Figures 1.1 and 1.2 of Chapter 1. Figure 1.1 shows that regardless of what level of age we choose, the slope of Y on X at that level is the same. If, however, we do not hold age constant (i.e., we do not regress on it), the slope becomes vastly different, as is shown in Figure 1.2.

The linearity assumption also permits us to fully specify the functions in the model by annotating the causal graph with a *path coefficient* (or structural coefficient) along each edge. The path coefficient β along the edge $X \rightarrow Y$ quantifies the contribution of X in the function that defines Y in the model. For instance, if the function defines $Y = 3X + U$, the path coefficient of $X \rightarrow Y$ will be 3. The path coefficients $\beta_1, \beta_2, \ldots, \beta_n$ are fundamentally different from the regression coefficients $r_1, r_2, \ldots, r_n$ that we discussed in Section 1.3. The former are "structural" or "causal," whereas the latter are statistical. The difference is explained in the next section.

Many of the regression methods we discuss are far more general, applying in situations where the variables $X_1, \ldots, X_k$ follow distribution far from multivariate Normal; for example, when some of the X_i's are categorical or even binary. Such generalizations also therefore allow the conditional mean $E(Y|X_1 = x_1, \ldots, X_k = x_k)$ to include nonlinear combinations of the X_i's, including such terms as X_1X_2, for example, to allow for effect modification, or interaction. Since we are conditioning on the values of the X_i's, it is usually not necessary to enforce a distributional assumption for such variables. Nevertheless, the full multivariate Normal scenario provides considerable insight into structural causal models.

3.8.1 Structural versus Regression Coefficients

As we are now about to deal with linear models, and thus, as a matter of course, with regression-like equations, it is of paramount importance to define the difference between regression equations and the structural equations we have used in SCMs throughout the book. A regression equation is descriptive; it makes no assumptions about causation. When we write $y = r_1x + r_2z + \epsilon$, as a regression equation, we are not saying that X and Z cause Y. We merely confess our need to know which values of r_1 and r_2 would make the equation $y = r_1x + r_2z$

the best linear approximation to the data, or, equivalently, the best linear approximation of $E(y|x, z)$.

Because of this fundamental difference between structural and regression equations, some books distinguish them by writing an arrow, instead of equality sign, in structural equations, and some distinguish the coefficients by using a different font. We distinguish them by denoting structural coefficients as α, β, and so on, and regression coefficients as r_1, r_2, and so on. In addition, we distinguish between the stochastic "error terms" that appear in these equations. Errors in regression equations are denoted ϵ_1, ϵ_2, and so on, as in Eq. (1.24), and those in structural equations by U_1, U_2, and so on, as in SCM 1.5.2. The former denote the residual errors in observation, after fitting the equation $y = r_1x + r_2z$ to data, whereas the latter represent latent factors (sometimes called "disturbances" or "omitted variables") that influence Y and are not themselves affected by X. The former are human-made (due to imperfect fitting); the latter are nature-made.

Though they are not causally binding themselves, regression equations are of significant use in the study of causality as it pertains to linear systems. Consider: In Section 3.2, we were able to express the effects of interventions in terms of conditional probabilities, as, for example, in the adjustment formula of Eq. (3.5). In linear systems, the role of conditional probabilities will be taken over by regression coefficients, since these coefficients represent the dependencies induced by the model and, in addition, they are easily estimable using least square analyses. Similarly, whereas the testable implications of nonparametric models are expressed in the form of conditional independencies, these independencies are signified in linear models by vanishing regression coefficients, like those discussed in Section 1.3.11. Specifically, given the regression equation

$$y = r_0 + r_1x_1 + r_2x_2 + \cdots + r_nx_n + \epsilon$$

if $r_i = 0$, then Y is independent of X_i conditional on all the other regression variables.

3.8.2 The Causal Interpretation of Structural Coefficients

In a linear system, every path coefficient stands for the direct effect of the independent variable, X, on the dependent variable, Y. To see why this is so, we refer to the interventional definition of direct effect given in Section 3.7 (Eq. (3.18)), which calls for computing the change in Y as X increases by one unit whereas all other parents of Y are held constant. When we apply this definition to any linear system, regardless of whether the disturbances are correlated or not, the result will be the path coefficient on the arrow $X \rightarrow Y$.

Consider, for example, the model in Figure 3.13, and assume we wish to estimate the direct effect of Z on Y. The structural equations in the fully specified model read:

$$\begin{aligned} X &= U_X \\ Z &= aX + U_Z \\ W &= bX + cZ + U_W \\ Y &= dZ + eW + U_Y \end{aligned}$$

Writing Eq. (3.18) in expectation form, we obtain

$$DE = E[Y|do(Z = z + 1), do(W = w)] - E[Y|do(Z = z), do(W = w)]$$

since W is the only other parent of Y in the graph. Applying the *do* operators by deleting the appropriate equations from the model, the postincrease term in DE becomes $d(z + 1) + ew$ and the preincrease term becomes $dz + ew$. As expected, the difference between the two is d—the path coefficient between Z and Y. Note that the license to reduce the equation in this way comes directly from the definition of the *do*-operator (Eq. (3.18)) making no assumption about correlations among the U factors; the equality $DE = d$ would be valid even if the error term U_Y were correlated with U_Z, though this would have made d nonidentifiable. The same goes for the other direct effects; every structural coefficient represents a direct effect, regardless of how the error terms are distributed. Note also that variable X, as well as the coefficients a, b, and c, do not enter into this computation, because the "surgeries" required by the *do* operators remove them from the model.

That is all well and good for the direct effect. Suppose, however, we wish to calculate the *total* effect of Z on Y.

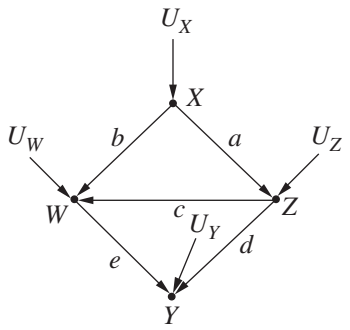

Figure 3.13 A graphical model illustrating the relationship between path coefficients and total effects

In a linear system, the total effect of X on Y is simply the sum of the products of the coefficients of the edges on every nonbackdoor path from X to Y.

That's a bit of a mouthful, so think of it as a process: To find the total effect of X on Y, first find every nonbackdoor path from X to Y; then, for each path, multiply all coefficients on the path together; then add up all the products.

The reason for this identity lies in the nature of SCMs. Consider again the graph of Figure 3.13. Since we want to find the total effect of Z on Y, we should first intervene on Z, removing all arrows going into Z, then express Y in terms of Z in the remaining model. This we can do with a little algebra:

$$\begin{aligned} Y &= dZ + eW + U_Y \\ &= dZ + e(bX + cZ) + U_Y + eU_W \\ &= (d + ec)Z + ebX + U_Y + eU_W \end{aligned}$$

The final expression is in the form $Y = \tau Z + U$, where $\tau = d + ec$ and U contains only terms that do not depend on Z in the modified model. An increase of a single unit in Z, therefore, will increase Y by τ—the definition of the total effect. A quick examination will show that τ

is the sum of the products of the coefficients on the two nonbackdoor paths from Z to Y. This will be the case in all linear models; algebra demands it. Moreover, the sum of product rule will be valid regardless of the distributions of the U variables and regardless of whether they are dependent or independent.

3.8.3 Identifying Structural Coefficients and Causal Effect

Thus far, we have expressed the total and direct effects in terms of path coefficients, assuming that the latter are either known to us a priori or estimated from interventional experiments. We now tackle a much harder problem; estimating total and direct effects from nonexperimental data. This problem is known as "identifiability" and, mathematically, it amounts to expressing the path coefficients associated with the total and direct effects in terms of the covariances σ_{XY} or regression coefficients $R_{YX \cdot Z}$, where X and Y are any two variables in the model, and Z a set of variables in the model (Eqs. (1.27) and (1.28) and Section 1.3.11).

In many cases, however, it turns out that to identify direct and total effects, we do not need to identify each and every structural parameter in the model. Let us first demonstrate with the total effect, τ. The backdoor criterion gives us the set Z of variables we need to adjust for in order to determine the causal effect of X on Y. How, though, do we make use of the criterion to determine effects in a linear system? In principle, once we obtain the set, Z, we can estimate the conditional expectation of Y given X and Z and, then, averaging over Z, we can use the resultant dependence between Y and X to measure the effect of X on Y. We need only translate this procedure to the language of regression.

The translation is rather simple. First, we find a set of covariates Z that satisfies the backdoor criterion from X to Y in the model. Then, we regress Y on X and Z. The coefficient of X in the resulting equation represents the true causal effect of X on Y. The reasoning for this is similar to the reasoning we used to justify the backdoor criterion in the first place—regressing on Z adds those variables into the equation, blocking all backdoor paths from X and Y, thus preventing the coefficient of X from absorbing the spurious information those paths contain.

For example, consider a linear model that complies with the graph in Figure 3.14. If we want to find the total causal effect of X on Y, we first determine, using the backdoor criterion, that we must adjust for T. So we regress Y on X and T, using the regression equation $y = r_X X +$

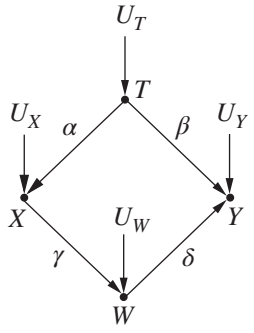

Figure 3.14 A graphical model in which X has no direct effect on Y, but a total effect that is determined by adjusting for T

$r_T T + \epsilon$. The coefficient r_X represents the total effect of X on Y. Note that this identification was possible without identifying any of the model parameters and without measuring variable W; the graph structure in itself gave us the license to ignore W, regress Y on T and X only, and identify the total effect (of X on Y) with the coefficient of X in that regression.

Suppose now that instead of the total causal effect, we want to find X's direct effect on Y. In a linear system, this direct effect is the structural coefficient α in the function $y = \alpha x + \beta z + \cdots + U_Y$ that defines Y in the system. We know from the graph of Figure 3.14 that $\alpha = 0$, because there is no direct arrow from X to Y. So, in this particular case, the answer is trivial: the direct effect is zero. But in general, how do we find the magnitude of α from data, if the model does not determine its value?

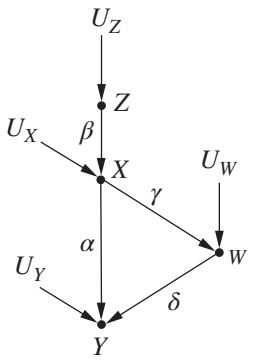

Figure 3.15 A graphical model in which X has direct effect α on Y

We can invoke a procedure similar to backdoor, except that now, we need to block not only backdoor paths but also indirect paths going from X to Y. First, we remove the edge from X to Y (if such an edge exists), and call the resulting graph G_α. If, in G_α, there is a set of variables Z that d-separates X and Y, then we can simply regress Y on X and Z. The coefficient of X in the resulting equation will equal the structural coefficient α.

The procedure above, which we might as well call "The Regression Rule for Identification" provides us with a quick way of determining whether any given parameter (say α) can be identified by ordinary least square (OLS) regression and, if so, what variables should go into the regression equation. For example, in the linear model of Figure 3.15, we can find the direct effect of X on Y by this method. First, we remove the edge between X and Y and get the graph G_α shown in Figure 3.16. It's easy to see that in this new graph, W d-separates X and Y. So we regress Y on X and W, using the regression equation $Y = r_X X + r_W W + \epsilon$. The coefficient r_X is the direct effect of X on Y.

Summarizing our observations thus far, two interesting features emerge. First, we see that, in linear systems, regression serves as the major tool for the identification and estimation of causal effects. To estimate a given effect, all we need to do is to write down a regression equation and specify (1) what variables should be included in the equation and (2) which of the coefficients in that equation represents the effect of interest. The rest is routine least square analysis on the sampled data which, as we remarked before, is facilitated by a variety of extremely efficient software packages. Second, we see that, as long as the U variables are independent of each

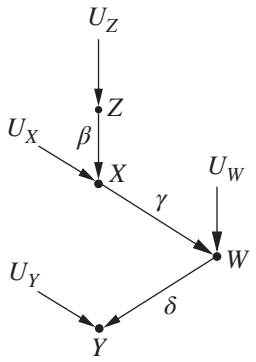

Figure 3.16 By removing the direct edge from X to Y and finding the set of variables $\{W\}$ that d-separate them, we find the variables we need to adjust for to determine the direct effect of X on Y

other, and all variables in the graph are measured, every structural parameter can be identified in this manner, namely, there is at least one identifying regression equation in which one of the coefficients corresponds to the parameter we seek to estimate. One such equation is obviously the structural equation itself, with the parents of Y serving as regressors. But there may be several other identifying equations, with possibly better features for estimation and graphical analysis can reveal them all (see Study question 3.8.1(c)). Moreover, when some variables are not measured, or when some error terms are correlated, the task of finding an identifying regression from the structural equations themselves would normally be insurmountable; the G_α procedure then becomes indispensable (see Study question 3.8.1(d)).

Remarkably, the regression rule procedure has eluded investigators for almost a century, possibly because it is extremely difficult to articulate in algebraic, nongraphical terms.

Suppose, however, there is no set of variables that d-separates X and Y in G_α. For instance, in Figure 3.17, X and Y have an unobserved common cause represented by the dashed

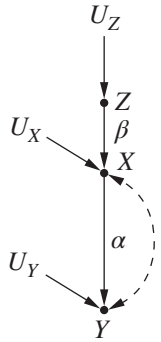

Figure 3.17 A graphical model in which we cannot find the direct effect of X on Y via adjustment, because the dashed double-arrow arc represents the presence of a backdoor path between X and Y, consisting of unmeasured variables. In this case, Z is an instrument with regard to the effect of X on Y that enables the identification of α

double-arrowed arc. Since it hasn't been measured, we can't condition on it, so X and Y will always be dependent through it. In this particular case, we may use an *instrumental variable* to determine the direct effect. A variable is called an "instrument" if it is d-separated from Y in G_α and, it is d-connected to X. To see why such a variable enables us to identify structural coefficients, we take a closer look at Figure 3.17.

In Figure 3.17, Z is an instrument with regard to the effect of X on Y because it is d-connected to X and d-separated from Y in G_α. We regress X and Y on Z separately, yielding the regression equations $y = r_1 z + \epsilon$ and $x = r_2 z + \epsilon$, respectively. Since Z emits no backdoors, r_2 equals β and r_1 equals the total effect of Z on Y, $\beta\alpha$. Therefore, the ratio r_1/r_2 provides the desired coefficient α. This example illustrates how direct effects can be identified from total effects but not the other way around.

Graphical models provide us with a procedure for finding all instrumental variables in a system, though the procedure for enumerating them is beyond the scope of this book. Those interested in learning more can (see Chen and Pearl 2014; Kyono 2010).

Study questions

Study question 3.8.1

Model 3.1

$$
\begin{aligned}
Y &= aW_3 + bZ_3 + cW_2 + U & X &= t_1 W_1 + t_2 Z_3 + U' \\
W_3 &= c_3 X + U'_3 & W_1 &= a'_1 Z_1 + U'_1 \\
Z_3 &= a_3 Z_1 + b_3 Z_2 + U_3 & Z_1 &= U_1 \\
W_2 &= c_2 Z_2 + U'_2 & Z_2 &= U_2
\end{aligned}
$$

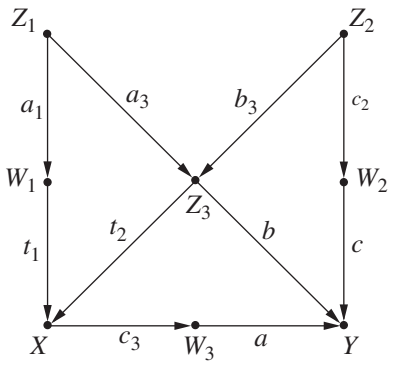

Figure 3.18 Graph corresponding to Model 3.1 in Study question 3.8.1

Given the model depicted above, answer the following questions:

(All answers should be given in terms of regression coefficients in specified regression equations.)

(a) Identify three testable implications of this model.
(b) Identify a testable implication assuming that only X, Y, W_3, and Z_3 are observed.
(c) For each of the parameters in the model, write a regression equation in which one of the coefficients is equal to that parameter. Identify the parameters for which more than one such equation exists.
(d) Suppose X, Y, and W_3 are the only variables observed. Which parameters can be identified from the data? Can the total effect of X on Y be estimated?
(e) If we regress Z_1 on all other variables in the model, which regression coefficient will be zero?
(f) The model in Figure 3.18 implies that certain regression coefficients will remain invariant when an additional variable is added as a regressor. Identify five such coefficients with their added regressors.
(g) Assume that variables Z_2 and W_2 cannot be measured. Find a way to estimate b using regression coefficients. [*Hint: Find a way to turn Z_1 into an instrumental variable for b.*]

3.8.4 Mediation in Linear Systems

When we can assume linear relationships between variables, mediation analysis becomes much simpler than the analysis conducted in nonlinear or nonparametric systems (Section 3.7). Estimating the direct effect of X on Y, for instance, amounts to estimating the path coefficient between the two variables, and this reduces to estimating correlation coefficients, using the techniques introduced in Section 3.8.3. The indirect effect, similarly, is computed via the difference $IE = \tau - DE$, where τ, the total effect, can be estimated by regression in the manner shown in Figure 3.14. In nonlinear systems, on the other hand, the direct effect is defined through expressions such as (3.18), or

$$DE = E[Y|do(x, z)] - E[Y|do(x', z)]$$

where $Z = z$ represents a specific stratum of all other parents of Y (besides X). Even when the identification conditions are satisfied, and we are able to reduce the *do*() operators (by adjustments) to ordinary conditional expectations, the result will still depend on the specific values of x, x', and z. Moreover, the indirect effect cannot be given a definition in terms as *do*-expressions, since we cannot disable the capacity of Y to respond to X by holding variables constant. Nor can the indirect effect be defined as the difference between the total and direct effects, since differences do not faithfully reflect operations in nonlinear systems to X.

Such an operation will be introduced in Chapter 4 (Sections 4.4.5 and 4.5.2) using the language of counterfactuals.

Bibliographical Notes for Chapter 3

Study question 3.3.2 is a version of Lord's paradox (Lord 1967), and is described in Glymour (2006), Hernández-Díaz et al. (2006), Senn (2006), and Wainer (1991). A unifying treatment is given in Pearl (2016). The definition of the *do*-operator and "ACE" in terms of a modified model, has its conceptual origin with the economist Trygve Haavelmo (1943), who was the first

to simulate interventions by modifying equations in the model (see Pearl (2015c) for historical account). Strotz and Wold (1960) later advocated "wiping out" the equation determining X, and Spirtes et al. (1993) gave it a graphical representation in a form of a "manipulated graph." The "adjustment formula" of Eq. (3.5) as well as the "truncated product formula" first appeared in Spirtes et al. (1993), though these are implicit in the G-computation formula of Robins (1986), which was derived using counterfactual assumptions (see Chapter 4). The backdoor criterion of Definition 3.3.1 and its implications for adjustments were introduced in Pearl (1993). The front-door criterion and a general calculus for identifying causal effects (named *do*-calculus) from observations and experimental data were introduced in Pearl (1995) and were further improved in Tian and Pearl (2002), Shpitser and Pearl (2007), and Bareinboim and Pearl (2012). Section 3.7, and the identification of conditional interventions and c-specific effects is based on (Pearl 2009, pp. 113–114). Its extension to dynamic, time-varying policies is described in Pearl and Robins (1995) and (Pearl 2009, pp. 119–126). More recently, the *do*-calculus was used to solve problems of external validity, data-fusion, and meta-analysis (Bareinboim and Pearl 2013, Bareinboim and Pearl 2016, and Pearl and Bareinboim 2014). The role of covariate-specific effects in assessing interaction, moderation or effect modification is described in Morgan and Winship (2014) and Vanderweele (2015), whereas applications of Rule 2 to the detection of latent heterogeneity are described in Pearl (2015b). Additional discussions on the use of inverse probability weighting (Section 3.6) can be found in Hernán and Robins (2006). Our discussion of mediation (Section 3.7) and the identification of CDEs are based on Pearl (2009, pp. 126–130), whereas the fallibility of "conditioning" on a mediator to assess direct effects is demonstrated in Pearl (1998) as well as Cole and Hernán (2002).

The analysis of mediation has become extremely active in the past 15 years, primarily due to the advent of counterfactual logic (see Section 4.4.5); a comprehensive account of this progress is given in Vanderweele (2015). A tutorial survey of causal inference in linear systems (Section 3.8), focusing on parameter identification, is provided by Chen and Pear1 (2014). Additional discussion on the confusion of regression versus structural equations can be found in Bollen and Pearl (2013).

A classic, and still the best textbook on the relationships between structural and regession coefficients is Heise (1975) (available online: http://www.indiana.edu/~socpsy/public_files/CausalAnalysis.zip). Other classics are Duncan (1975), Kenny (1979), and Bollen (1989). Classical texts, however, fall short of providing graphical tools of identification, such as those invoking backdoor and G_α (see Study question 3.8.1). A recent exception is Kline (2016).

Introductions to instrumental variables can be found in Greenland (2000) and in many textbooks of econometrics (e.g., Bowden and Turkington 1984, Wooldridge 2013). Generalized instrumental variables, extending the classical definition of Section 3.8.3 were introduced in Brito and Pearl (2002).

The program DAGitty (which is available online: http://www.dagitty.net/dags.html), permits users to search the graph for generalized instrumental variables, and reports the resulting IV estimators (Textor et al. 2011).