About this Errata Page

This file contains changes made to the text of *Causal Inference in Statistics: A Primer* by J. Pearl, Madelyn Glymour, and N.P. Jewell.

Changes are pointed to with arrows in the left margins which will make it easy for you you to mark your own personal copy.

If you should discover additional corrections or needed clarification, please let us know (kaoru@cs.ucla.edu).

Similarly, the expected value of any function of X—say, $g(X)$—is obtained by summing $g(x)P(X = x)$ over all values of X.

$$E[g(X)] = \sum_x g(x)P(x) \tag{1.11}$$

For example, if after rolling a die, I receive a cash prize equal to the square of the result, we have $g(X) = X^2$, and the expected prize is

$$E[g(X)] = \left(1^2 \times \frac{1}{6}\right) + \left(2^2 \times \frac{1}{6}\right) + \left(3^2 \times \frac{1}{6}\right) + \left(4^2 \times \frac{1}{6}\right) + \left(5^2 \times \frac{1}{6}\right) + \left(6^2 \times \frac{1}{6}\right) = 15.17 \tag{1.12}$$

We can also calculate the expected value of Y conditional on X, $E(Y|X = x)$, by multiplying each possible value y of Y by $P(Y = y|X = x)$, and summing the products.

$$E(Y|X = x) = \sum_y y\, P(Y = y|X = x) \tag{1.13}$$

$E(X)$ is one way to make a "best guess" of X's value. Specifically, out of all the guesses g that we can make, the choice "$g = E(X)$" minimizes the expected square error $E(g - X)^2$. Similarly, $E(Y|X = x)$ represents a best guess of Y, given that we observe $X = x$. If $g = E(Y|X = x)$, then g minimizes the expected square error $E[(g - Y)^2|X = x)]$.

For example, the expected age of a 2012 voter, as demonstrated by Table 1.3, is

$$E(\textit{Voter's Age}) = 23.5 \times 0.16 + 37 \times 0.23 + 54.5 \times 0.39 + 70 \times 0.22 = 48.9$$

(For this calculation, we have assumed that every age within each category is equally likely, e.g., a voter is as likely to be 18 as 25, and as likely to be 30 as 44. We have also assumed that the oldest age of any voter is 75.) This means that if we were asked to guess the age of a randomly chosen voter, with the understanding that if we were off by e years, we would lose e^2 dollars, we would lose the least money, on average, if we guessed 48.9. Similarly, if we were asked to guess the age of a random voter younger than the age of 45, our best bet would be

$$E[\textit{Voter's Age} \mid \textit{Voter's Age} < 45] = 23.5 \times 0.40 + 37 \times 0.60 = 31.6 \tag{1.14}$$

The use of expectations as a basis for predictions or "best guesses" hinges to a great extent on an implicit assumption regarding the distribution of X or $Y|X = x$, namely that such distributions are approximately *symmetric*. If, however, the distribution of interest is highly *skewed*, other methods of prediction may be better. In such cases, for example, we might use the median of the distribution of X as our "best guess"; this estimate minimizes the expected absolute error $E(|g - X|)$. We will not pursue such alternative measures further here.

1.3.9 Variance and Covariance

The *variance* of a variable X, denoted $Var(X)$ or σ_X^2, is a measure of roughly how "spread out" the values of X in a data set or population are from their mean. If the values of X all hover close

which represents an inclined plane through the three-dimensional coordinate system.

We can create a three-dimensional scatter plot, with values of Y on the y-axis, X on the x-axis, and Z on the z-axis. Then, we can cut the scatter plot into slices along the Z-axis. Each slice will constitute a two-dimensional scatter plot of the kind shown in Figure 1.4. Each of those 2-D scatter plots will have a regression line with a slope r_1. Slicing along the X-axis will give the slope r_2.

The slope of Y on X when we hold Z constant is called the *partial regression coefficient* and is denoted by $R_{YX \cdot Z}$. Note that it is possible for R_{YX} to be positive, whereas $R_{YX \cdot Z}$ is negative as shown in Figure 1.1. This is a manifestation of Simpson's Paradox: positive association between Y and X overall, that becomes negative when we condition on the third variable Z.

The computation of partial regression coefficients (e.g., r_1 and r_2 in (1.23)) is greatly facilitated by a theorem that is one of the most fundamental results in regression analysis. It states that if we write Y as a linear combination of variables $X_1, X_2, \ldots, X_k$ plus a noise term ϵ,

$$Y = r_0 + r_1 X_1 + r_2 X_2 + \cdots + r_k X_k + \epsilon \tag{1.24}$$

then, regardless of the underlying distribution of $Y, X_1, X_2, \ldots, X_k$, the best least-square coefficients are obtained when ϵ is uncorrelated with each of the regressors $X_1, X_2, \ldots, X_k$. That is,

$$Cov(\epsilon, X_i) = 0 \quad \text{for} \quad i = 1,2, \ldots, k$$

To see how this *orthogonality principle* is used to our advantage, assume we wish to compute the best estimate of $X = Die\ 1$ given the sum

$$Y = Die\ 1 + Die\ 2$$

Writing

$$X = \alpha + \beta Y + \epsilon$$

our goal is to find α and β in terms of estimable statistical measures. Assuming without loss of generality $E[\epsilon] = 0$, and taking expectation on both sides of the equation, we obtain

$$E[X] = \alpha + \beta E[Y] \tag{1.25}$$

Further multiplying both sides of the equation by Y and taking the expectation gives

$$E[XY] = \alpha E[Y] + \beta E[Y^2] + E[Y\epsilon] \tag{1.26}$$

The orthogonality principle dictates $E[Y\epsilon] = 0$, and (1.25) and (1.26) yield two equations with two unknowns, α and β. Solving for α and β, we obtain

$$\alpha = E(X) - E(Y)\frac{\sigma_{XY}}{\sigma_Y^2}$$

$$\beta = \frac{\sigma_{XY}}{\sigma_Y^2}$$

which completes the derivation. The slope β could have been obtained from Eq. (1.22), by simply reversing X and Y, but the derivation above demonstrates a general method of computing slopes, in two or more dimensions.

SCM 2.2.3 (Work Hours, Training, and Race Time)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$
$$f_X : X = U_X$$
$$f_Y : Y = 84 - x + U_Y$$
$$f_Z : Z = \frac{100}{y} + U_Z$$

SCMs 2.2.1–2.2.3 share the graphical model shown in Figure 2.1.

SCMs 2.2.1 and 2.2.3 deal with continuous variables; SCM 2.2.2 deals with categorical variables. The relationships between the variables in 2.1.1 are all positive (i.e., the higher the value of the parent variable, the higher the value of the child variable); the correlations between the variables in 2.2.3 are all negative (i.e., the higher the value of the parent variable, the lower the value of the child variable); the correlations between the variables in 2.2.2 are not linear at all, but logical. No two of the SCMs share any functions in common. But because they share a common graphical structure, the data sets generated by all three SCMs must share certain independencies—and we can predict those independencies simply by examining the graphical model in Figure 2.1. The independencies shared by data sets generated by these three SCMs, and the dependencies that are likely shared by all such SCMs, are these:

1. ***Z and Y are likely dependent***
   For some z, y, $P(Z = z|Y = y) \neq P(Z = z)$
2. ***Y and X are likely dependent***
   For some y, x, $P(Y = y|X = x) \neq P(Y = y)$
3. ***Z and X are likely dependent***
   For some z, x, $P(Z = z|X = x) \neq P(Z = z)$
4. ***Z and X are independent, conditional on Y***
   For all x, y, z, $P(Z = z|X = x, Y = y) = P(Z = z|Y = y)$

To understand why these independencies and dependencies hold, let's examine the graphical model. First, we will verify that any two variables with an edge between them are likely dependent. Remember that an arrow from one variable to another indicates that the first variable causes the second—that is, the value of the first variable is part of the function that determines the value of the second. Therefore, the second variable *depends* on the first for

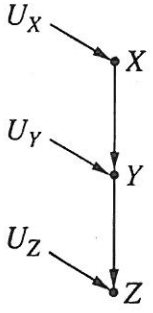

Figure 2.1 The graphical model of SCMs 2.2.1–2.2.3

its value; there is some case in which changing the value of the first variable changes the value of the second. That means that when we examine those variables in the data set, the probability that one variable takes a given value will change, given that we know the value of the other variable. So in any causal model, regardless of the specific functions, two variables connected by an edge are dependent. By this reasoning, we can see that in SCMs 2.2.1–2.2.3, Z and Y are dependent, and Y and X are dependent.

From these two facts, we can conclude that Z and X are *likely* dependent. If Z depends on Y for its value, and Y depends on X for its value, then Z likely depends on X for its value. There are pathological cases in which this is not true. Consider, for example, the following SCM, which also has the graph in Figure 2.1.

SCM 2.2.4 (Pathological Case of Intransitive Dependence)

$$V = \{X, Y, Z\}, U = \{U_X, U_Y, U_Z\}, F = \{f_X, f_Y, f_Z\}$$

$$f_X : X = U_X$$

$$f_Y : Y = \begin{cases} a & \text{IF } X = 1 \text{ AND } U_Y = 1 \\ b & \text{IF } X = 2 \text{ AND } U_Y = 1 \\ c & \text{IF } U_Y = 2 \end{cases}$$

$$f_Y : Z = \begin{cases} i & \text{IF } Y = c \text{ OR } U_Z = 1 \\ j & \text{IF } U_Z = 2 \end{cases}$$

In this case, no matter what value U_Y and U_Z take, X will have no effect on the value that Z takes; changes in X account for variation in Y between a and b, but Y doesn't affect Z unless it takes the value c. Therefore, X and Z vary independently in this model. We will call cases such as these *intransitive cases*.

However, intransitive cases form only a small number of the cases we will encounter. In most cases, the values of X and Z vary together just as X and Y do, and Y and Z. Therefore, they are likely dependent in the data set.

Now, let's consider point 4: Z and X are independent conditional on Y. Remember that when we condition on Y, we filter the data into groups based on the value of Y. So we compare all the cases where $Y = a$, all the cases where $Y = b$, and so on. Let's assume that we're looking at the cases where $Y = a$. We want to know whether, *in these cases only*, the value of Z is independent of the value of X. Previously, we determined that X and Z are likely dependent, because when the value of X changes, the value of Y likely changes, and when the value of Y changes, the value of Z is likely to change. Now, however, examining *only the cases where* $Y = a$, when we select cases with different values of X, the value of U_Y changes so as to keep Y at $Y = a$, but since Z depends only on Y and U_Z, not on U_Y, the value of Z remains unaltered. So selecting a different value of X doesn't change the value of Z. So, in the case where $Y = a$, X is independent of Z. This is of course true no matter which specific value of Y we condition on. So X is independent of Z, conditional on Y.

This configuration of variables—three nodes and two edges, with one edge directed into and one edge directed out of the middle variable—is called a *chain*. Analogous reasoning to the above tells us that in any graphical model, given any two variables X and Y, if the only path between X and Y is composed entirely of chains, then X and Y are independent conditional on any intermediate variable on that path. This independence relation holds regardless of the functions that connect the variables. This gives us a rule:

[#] This occurs for example when X and U_Y are fair coins and $Y = 1$ if and only $X=U_Y$. In this case $P(Y=1|X=1) = P(Y=1|X=0) = P(Y=1)=1/2$. Such pathological cases require precise numerical probabilities to achieve independence ($P(X=1)=P(U_X)=1/2$); they are rare, and can be ignored for all practical purposes.

If we assume that the error terms U_X, U_Y, and U_Z are independent, then by examining the graphical model in Figure 2.2, we can determine that SCMs 2.2.5 and 2.2.6 share the following dependencies and independencies:

1. ***X and Y are dependent.***
   For some x, y, $P(X = x|Y = y) \neq P(X = x)$
2. ***X and Z are dependent.***
   For some x, z, $P(X = x|Z = z) \neq P(X = x)$
3. ***Z and Y are likely dependent.***
   For some z, y, $P(Z = z|Y = y) \neq P(Z = z)$
4. ***Y and Z are independent, conditional on X.***
   For all $4x, y, z$, $P(Y = y|Z = z, X = x) = P(Y = y|X = x)$

Points 1 and 2 follow, once again, from the fact that Y and Z are both directly connected to X by an arrow, so when the value of X changes, the values of both Y and Z change. This tells us something further, however: If Y changes when X changes, and Z changes when X changes, then it is likely (though not certain) that Y changes together with Z, and vice versa. Therefore, since a change in the value of Y gives us information about an associated change in the value of Z, Y, and Z are likely dependent variables.

Why, then, are Y and Z independent conditional on X? Well, what happens when we condition on X? We filter the data based on the value of X. So now, we're only comparing cases where the value of X is constant. Since X does not change, the values of Y and Z do not change in accordance with it—they change only in response to U_Y and U_Z, which we have assumed to be independent. Therefore, any additional changes in the values of Y and Z must be independent of each other.

This configuration of variables—three nodes, with two arrows emanating from the middle variable—is called a *fork*. The middle variable in a fork is the *common cause* of the other two variables, and of any of their descendants. If two variables share a common cause, and if that common cause is part of the only path between them, then analogous reasoning to the above tells us that these dependencies and conditional independencies are true of those variables. Therefore, we come by another rule:

Rule 2 (Conditional Independence in Forks) *If a variable X is a common cause of variables Y and Z, and there is only one path between Y and Z, then Y and Z are independent conditional on X.*

2.3 Colliders

So far we have looked at two simple configurations of edges and nodes that can occur on a path between two variables: chains and forks. There is a third such configuration that we speak of separately, because it carries with it unique considerations and challenges. The third configuration contains a *collider* node, and it occurs when one node receives edges from two other nodes. The simplest graphical causal model containing a collider is illustrated in Figure 2.3, representing a common effect, Z, of two causes X and Y.

As is the case with every graphical causal model, all SCMs that have Figure 2.3 as their graph share a set of dependencies and independencies that we can determine from the graphical

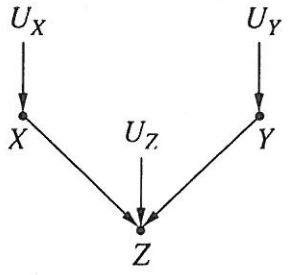

Figure 2.3 A simple collider

model alone. In the case of the model in Figure 2.3, assuming independence of U_X, U_Y, and U_Z, these independencies are as follows:

1. ***X and Z are dependent.***
   For some x, z, $P(X = x|Z = z) \neq P(X = x)$
2. ***Y and Z are dependent.***
   For some y, z, $P(Y = y|Z = z) \neq P(Y = y)$
3. ***X and Y are independent.***
   For all x, y, $P(X = x|Y = y) = P(X = x)$
4. ***X and Y are dependent conditional on Z.***
   For some x, y, z, $P(X = x|Y = y, Z = z) \neq P(X = x|Z = z)$

The truth of the first two points was established in Section 2.2. Point 3 is self-evident; neither X nor Y is a descendant or an ancestor of the other, nor do they depend for their value on the same variable. They respond only to U_X and U_Y, which are assumed independent, so there is no causal mechanism by which variations in the value of X should be associated with variations in the value of Y. This independence also reflects our understanding of how causation operates in time; events that are independent in the present do not become dependent merely because they may have common effects in the future.

Why, then, does point 4 hold? Why would two independent variables suddenly become dependent when we condition on their common effect? To answer this question, we return again to the definition of conditioning as filtering by the value of the conditioning variable. When we condition on Z, we limit our comparisons to cases in which Z takes the same value. But remember that Z depends, for its value, on X and Y. So, when comparing cases where Z takes, for example, the value, any change in value of X must be compensated for by a change in the value of Y—otherwise, the value of Z would change as well.

The reasoning behind this attribute of colliders—that conditioning on a collision node produces a dependence between the node's parents—can be difficult to grasp at first. In the most basic situation where $Z = X + Y$, and X and Y are independent variables, we have the following logic: If I tell you that $X = 3$, you learn nothing about the potential value of Y, because the two numbers are independent. On the other hand, if I start by telling you that $Z = 10$, then telling you that $X = 3$ immediately tells you that Y must be 7. Thus, X and Y are dependent, *given that* $Z = 10$.

This phenomenon can be further clarified through a real-life example. For instance, suppose a certain college gives scholarships to two types of students: those with unusual musical talents and those with extraordinary grade point averages. Ordinarily, musical talent and scholastic achievement are independent traits, so, in the population at large, finding a person with musical

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

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

using the Law of Total Probability

Equation (3.3) is obtained from Bayes' rule 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

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

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

The second line was licensed by Theorem 4.3.1, whereas the third line was licensed by the consistency rule (4.6).

The fact that we obtained the familiar adjustment formula in Eq. (4.16) is not really surprising, because this same formula was derived in Section 3.2 (Eq. (3.4)), for $P(Y = y|do(x))$, and we know that $P(Y_x = y)$ is just another way of writing $P(Y = y|do(x))$. Interestingly, this derivation invokes only algebraic steps; it makes no reference to the model once we ensure that Z satisfies the backdoor criterion. Equation (4.15), which converts this graphical reality into algebraic notation, and allows us to derive (4.16), is sometimes called "conditional ignorability"; Theorem 4.3.1 gives this notion a scientific interpretation and permits us to test whether it holds in any given model.

Having a graphical representation for counterfactuals, we can resolve the dilemma we faced in Section 4.3.1 (Figure 4.3), and explain graphically why a stronger education (X) would have had an effect on the salary (Y) of people who are currently at skill level $Z = z$, despite the fact that, according to the model, salary is determined by skill only. Formally, to determine if the effect of education on salary (Y_x) is statistically independent of the level of education, we need to locate Y_x in the graph and see if it is d-separated from X given Z. Referring to Figure 4.3, we see that Y_x can be identified with U_2, the only parent of nodes on the causal path from X to Y (and therefore, the only variable that produces variations in Y_x while X is held constant). A quick inspection of Figure 4.3 tells us that Z acts as a collider between X and U_2, and, therefore, X and U_2 (and similarly X and Y_x) are not d-separated given Z. We conclude therefore

$$E[Y_x|X, Z] \neq E[Y_x|Z]$$

despite the fact that

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

In Study question 4.3.1, we evaluate these counterfactual expectations explicitly, assuming a linear Gaussian model. The graphical representation established in this section permits us to determine independencies among counterfactuals by graphical means, without assuming linearity or any specific parametric form. This is one of the tools that modern causal analysis has introduced to statistics, and, as we have seen in the analysis of the education–skill–salary story, it takes a task that is extremely hard to solve by unaided intuition and reduces it to simple operations on graphs. Additional methods of visualizing counterfactual dependencies, called "twin networks," are discussed in (Pearl 2000, pp. 213–215).

4.3.3 *Counterfactuals in Experimental Settings*

Having convinced ourselves that every counterfactual question can be answered from a fully specified structural model, we next move to the experimental setting, where a model is not available, and the experimenter must answer interventional questions on the basis of a finite sample of observed individuals. Let us refer back to the "encouragement design" model of Figure 4.1, in which we analyzed the behavior of an individual named Joe, and assume that the experimenter observes a set of 10 individuals, with Joe being participant 1. Each individual is characterized by a distinct vector $U_i = (U_X, U_H, U_Y)$, as shown in the first three columns of Table 4.3.

We note that, in general, the total effect can be decomposed as

$$TE = NDE - NIE_r \tag{4.48}$$

where NIE_r stands for the NIE under the reverse transition, from $T = 1$ to $T = 0$. This implies that NIE is identifiable whenever NDE and TE are identifiable. In linear systems, where reversal of transitions amounts to negating the signs of their effects, we have the standard additive formula, $TE = NDE + NIE$.

We further note that TE and $CDE(m)$ are *do*-expressions and can, therefore, be estimated from experimental data or in observational studies using the backdoor or front-door adjustments. Not so for the NDE and NIE; a new set of assumptions is needed for their identification.

Conditions for identifying natural effects

The following set of conditions, marked A-1 to A-4, are sufficient for identifying both direct and indirect natural effects.

We can identify the NDE and NIE provided that there exists a set W of measured covariates such that

A-1 No member of W is a descendant of T.
A-2 W blocks all backdoor paths from M to Y (after removing $T \rightarrow M$ and $T \rightarrow Y$).
A-3 The W-specific effect of T on M is identifiable (possibly using experiments or adjustments).
A-4 The W-specific joint effect of $\{T, M\}$ on Y is identifiable (possibly using experiments or adjustments).

Theorem 4.5.2 (Identification of the *NDE*) *When conditions A-1 and A-2 hold, the natural direct effect is experimentally identifiable and is given by*

$$NDE = \sum_m \sum_w [E[Y|do(T = 1, M = m), W = w] - E[Y|do(T = 0, M = m), W = w]] \times P(M = m|do(T = 0), W = w)P(W = w) \tag{4.49}$$

The identifiability of the do-expressions in Eq. (4.49) is guaranteed by conditions A-3 and A-4 and can be determined using the backdoor or front-door criteria.

Corollary 4.5.1 *If conditions A-1 and A-2 are satisfied by a set W that also deconfounds the relationships in A-3 and A-4, then the do-expressions in Eq. (4.49) are reducible to conditional expectations, and the natural direct effect becomes*

$$NDE = \sum_m \sum_w [E[Y|T = 1, M = m, W = w] - E[Y|T = 0, M = m, W = w]] \times P(M = m|T = 0, W = w)P(W = w) \tag{4.50}$$

In the nonconfounding case (Figure 4.6(a)), NDE reduces to

$$NDE = \sum_m [E[Y \mid T = 1, M = m] - E[Y \mid T = 0, M = m]]P(M = m \mid T = 0). \tag{4.51}$$