is randomized. We recall that this probability can be calculated from a causal model M either directly, by simulating the intervention $do(X = x)$, or (if $P(x, s) > 0$) via the adjustment formula (equation (3.19))

$$P(y|do(x)) = \sum_s P(y|x, s)P(s),$$

where S stands for any set of variables, observed as well as unobserved, that satisfy the back-door criterion (Definition 3.3.1). Equivalently, $P(y|do(x))$ can be written $P(Y(x) = y)$, where $Y(x)$ is the potential-outcome variable as defined in (3.51) or in Rubin (1974). We bear in mind that the operator $do(\cdot)$, and hence also effect estimates and confounding, must be defined relative to a specific causal or data-generating model M because these notions are not statistical in character and cannot be defined in terms of joint distributions.

Definition 6.2.2 (No-Confounding; Associational Criterion)
Let T be the set of variables in a problem that are not affected by X. We say that X and Y are not confounded *in the presence of T if each member Z of T satisfies at least one of the following conditions:*

(U_1) *Z is not associated with X (i.e., $P(x|z) = P(x)$);*

(U_2) *Z is not associated with Y, conditional on X (i.e., $P(y|z, x) = P(y|x)$).*

Conversely, X and Y are said to be confounded *if any member Z of T violates both (U_1) and (U_2).*

Note that the associational criterion in Definition 6.2.2 is not purely statistical in that it invokes the predicate "affected by" which is not discernible from probabilities but rests instead on causal information. This exclusion of variables that are affected by treatments (or exposures) is unavoidable and has long been recognized as a necessary judgmental input to every analysis of treatment effect in observational and experimental studies alike (Cox 1958, p. 48; Greenland and Neutra 1980). We shall assume throughout that investigators possess the knowledge required for distinguishing variables that are affected by the treatment X from those that are not. We shall then explore what additional causal knowledge is needed, if any, for establishing a test of confounding.

6.3 How the Associational Criterion Fails

We will say that a criterion for no-confounding is *sufficient* if it never errs when it classifies a case as no-confounding and *necessary* if it never errs when it classifies a case as confounding. There are several ways that the associational criterion of Definition 6.2.2 fails to match the causal criterion of Definition 6.2.1. Failures with respect to sufficiency and necessity will be addressed in turn.

6.3.1 Failing Sufficiency via Marginality

The criterion in Definition 6.2.2 is based on testing each element of T individually. A situation may well be present where two factors, Z_1 and Z_2, jointly confound X and Y (in the sense of Definition 6.2.2) and yet each factor separately satisfies (U_1) or (U_2). This may occur because statistical independence between X and individual members of T does not guarantee the independence of X and groups of variables taken from T. For example, let Z_1 and Z_2 be the outcomes of two independent fair coins, each affecting both X and Y. Assume that X occurs when Z_1 and Z_2 are equal and that Y occurs whenever Z_1 and Z_2 are unequal. Clearly, X and Y are highly confounded by the pair $T = (Z_1,\ Z_2)$; they are, in fact, perfectly correlated (negatively) without causally affecting each other. Yet, neither Z_1 nor Z_2 is associated with either X or Y; discovering the outcome of any one coin does not change the probability of X (or of Y) from its initial value of $\frac{1}{2}$.

An attempt to remedy Definition 6.2.2 by replacing Z with arbitrary subsets of T in (U_1) and (U_2) would be much too restrictive, because the set of *all* causes of X and Y, when treated as a group, would almost surely fail the tests of (U_1) and (U_2). In Section 6.5.2 we identify the subsets that should replace Z in (U_1) and (U_2) if sufficiency is to be restored.

6.3.2 Failing Sufficiency via Closed-World Assumptions

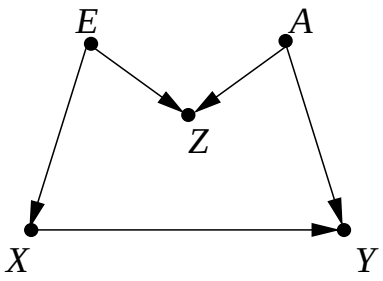
By "closed-world" assumption I mean the assumption that our model accounts for all relevant variables and, specifically to Definition 6.2.2, that the set T of variables consists of *all* potential confounders in a problem. In order to correctly classify every case of no-confounding, the associational criterion requires that condition (U_1) or (U_2) be satisfied for every potential confounder Z in a problem. In practice, since investigators can never be sure whether a given set T of potential confounders is complete, the associational criterion will falsely classify certain confounded cases as unconfounded.

This limitation actually implies that any statistical test whatsoever is destined to be insufficient. Since practical tests always involve proper subsets of T, the most we can hope to achieve by statistical means is *necessity*—that is, a test that would correctly label cases as confounding when criteria such as (U_1) and (U_2) are violated by an arbitrary subset of T. This prospect, too, is not fulfilled by Definition 6.2.2, as we now demonstrate.

6.3.3 Failing Necessity via Barren Proxies

Example 6.3.1 Imagine a situation where exposure (X) is influenced by a person's education (E), disease (Y) is influenced by both exposure and age (A), and car type (Z) is influenced by both age (A) and education (E). These relationships are shown schematically in Figure 6.3.

The car-type variable (Z) violates the two conditions in Definition 6.2.2 because: (1) car type is indicative of education and hence is associated with the exposure variable; and (2) car type is indicative of age and hence is associated with the disease among the exposed and the nonexposed. However, in this example the effect of X on Y



X - exposure
Y - disease
Z = type of car owned by patient
E = education
A = age

Figure 6.3: X and Y are not confounded,though Z is associated with both confounder

is not confounded; the type of car owned by a person has no effect on either exposure or disease and is merely one among many irrelevant properties that are associated with both via intermediaries. The analysis of Chapter 3 establishes that, indeed, (6.10) is satisfied in this model[12] *and that, moreover, adjustment for Z would generally yield a biased result:*

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

Thus we see that the traditional criterion based on statistical association fails to identify an unconfounded effect and would tempt one to adjust for the wrong variable. This failure occurs whenever we apply (U_1) and (U_2) to a variable Z that is a *barren proxy*—that is, a variable that has no influence on X or Y but is a proxy for factors that do have such influence.

Readers may not consider this failure to be too serious, because experienced epidemiologists would rarely regard a variable as confounder unless it is suspect of having some influence on either X or Y. Nevertheless, adjustment for proxies is a prevailing practice in epidemiology and should be done with great caution (Greenland and Neutra 1980; Weinberg 1993). To regiment this caution, the associational criterion must be modified to exclude barren proxies from the test set T. This yields the following modified criterion in which T consists only of variables that (causally) influence Y (possibly through X).

Definition 6.3.2 (No-Confounding; Modified Associational Criterion)
Let T be the set of variables in a problem that are not affected by X but may potentially affect Y. We say that X and Y are unconfounded *by the presence of T if and only if every member Z of T satisfies either (U_1) or (U_2) of Definition 6.2.2.*

Stone (1993) and Robins (1997) proposed alternative modifications of Definition 6.2.2 that avoid the problems created by barren proxies without requiring one to judge whether a variable has an effect on Y. Instead of restricting the set T to potential

[12] Because the (back-door) path $X \leftarrow E \rightarrow Z \leftarrow A \rightarrow Y$ is blocked by the colliding arrows at Z (see Definition 3.3.1).

causes of Y, we let T remain the set of *all* variables unaffected by X,[13] requiring instead that T be composed of two disjoint subsets, T_1 and T_2, such that

(U_1^*) T_1 is unassociated with X *and*

(U_2^*) T_2 is unassociated with Y given X and T_1.

In the model of Figure 6.3, for instance, conditions (U_1^*) and (U_2^*) are satisfied by the choice $T_1 = A$ and $T_2 = \{Z, E\}$, because (using the d-separation test) A is independent of X, and E is independent of Y, given $\{X, A\}$.

This modification of the associational criterion further rectifies the problem associated with marginality (see Section 6.3.1) because (U_1^*) and (U_2^*) treat T_1 and T_2 as compound variables. However, the modification falls short of restoring necessity. Because the set $T = (T_1, T_2)$ must include *all* variables unaffected by X (see note 13) and because practical tests are limited to proper subsets of T, we cannot conclude that confounding is present solely upon the failure of (U_1^*) and (U_2^*), as specified in Section 6.3.2. This criterion, too, is thus inadequate as a basis for practical detection of confounding.

We now discuss another fundamental limitation on our ability to detect confounding by statistical means.

6.3.4 Failing Necessity via Incidental Cancellations

Here we present a case that is devoid of barren proxies and in which the effect of X on Y (i) is not confounded in the sense of (6.10) but (ii) is confounded according to the modified associational criterion of Definition 6.3.2.

Example 6.3.3 *Consider a causal model defined by the linear equations*

$$x = \alpha z + \epsilon_1, \tag{6.11}$$

$$y = \beta x + \gamma z + \epsilon_2, \tag{6.12}$$

where ϵ_1 and ϵ_2 are correlated unmeasured variables with $\mathrm{cov}(\epsilon_1, \epsilon_2) = r$ and where Z is an exogenous variable that is uncorrelated with ϵ_1 or ϵ_2. The diagram associated with this model is depicted in Figure 6.4. The effect of X on Y is quantified by the path coefficient β, which gives the rate of change of $E(Y|do(x))$ per unit change in x.[14]

It is not hard to show (assuming standardized variables) that the regression of Y on X gives

$$y = (\beta + r + \alpha\gamma)x + \epsilon,$$

[13] Alternatively, T can be confined to any set S of variables sufficient for control of confounding:

$$P(y|do(x)) = \sum_s P(y|x, s)P(s).$$

Again, however, we can never be sure if the measured variables in the model contain such a set, or which of T's subsets possess this property.

[14] See Sections 3.5–3.6 or (5.24) in Section 5.4.1.

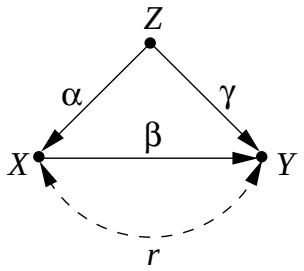


Figure 6.4: Z is associated with both X and Y, yet the effect of X on Y is not confounded (when $r = -\alpha\gamma$).

where $\mathrm{cov}(x, \epsilon) = 0$. Thus, whenever the equality $r = -\alpha\gamma$ holds, the regression coefficient of $r_{YX} = \beta + r + \alpha\gamma$ is an unbiased estimate of β, meaning that the effect of X on Y is unconfounded (no adjustment is necessary). Yet the associational conditions (U_1) and (U_2) are both violated by the variable Z; Z is associated with X (if $\alpha \neq 0$) and conditionally associated with Y, given X (except for special values of γ for which $\rho_{yz \cdot x} = 0$).

This example demonstrates that the condition of unbiasedness (Definition 6.2.1) does not imply the modified criterion of Definition 6.3.2. The associational criterion might falsely classify some unconfounded situations as confounded and, worse yet, adjusting for the false confounder (Z in our example) will introduce bias into the effect estimate.[15]

6.4 Stable versus Incidental Unbiasedness

6.4.1 Motivation

The failure of the associational criterion in the previous example calls for a reexamination of the notion of confounding and unbiasedness as defined in (6.10). The reason that X and Y were classified as unconfounded in Example 6.3.3 was that, by setting $r = -\alpha\gamma$, we were able to make the spurious association represented by r *cancel* the one mediated by Z. In practice, such perfect cancelation would be an incidental event specific to a peculiar combination of study conditions, and it would not persist when the parameters of the problem (i.e., α, γ, and r) undergo slight changes—say, when the study is repeated in a different location or at a different time. In contrast, the condition of no-confounding found in Example 6.3.1 does not exhibit such volatility. In this example, the unbiasedness expressed in (6.10) would continue to hold regardless of the strength of connection between education and exposure and regardless on how education and age influence the type of car that a patient owns. We call this type of unbiasedness *stable*, since it is robust to change in parameters and remains intact as long as the configuration of causal connections in the model remains the same.

[15] Note that the Stone-Robins modifications of Definition 6.3.2 would also fail in this example, unless we can measure the factors responsible for the correlation between ϵ_1 and ϵ_2.