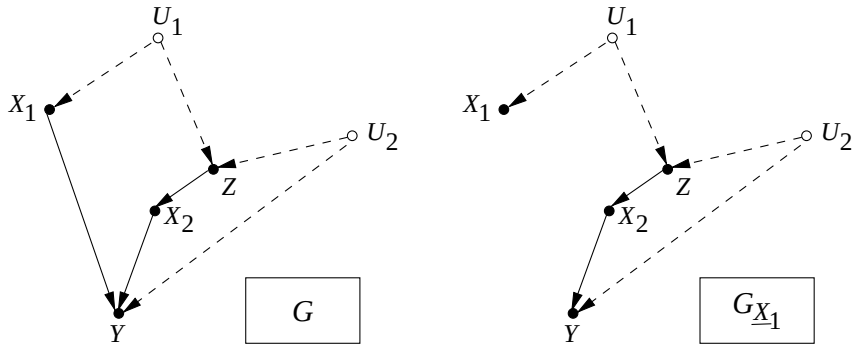

Figure 4.8: Causal diagram G in which proper ordering of the control variables X_1 and X_2 is important.

4.5 Direct Effects and Their Identification

4.5.1 Direct versus Total Effects:

The causal effect we have analyzed so far, $P(y|\hat{x})$, measures the *total* effect of a variable (or a set of variables) X on a response variable Y. In many cases, this quantity does not adequately represent the target of investigation and attention is focused instead on the direct effect of X on Y. The term "direct effect" is meant to quantify an effect that is not mediated by other variables in the model or, more accurately, the sensitivity of Y to changes in X while all other factors in the analysis are held fixed. Naturally, holding those factors fixed would sever all causal paths from X to Y with the exception of the direct link $X \rightarrow Y$, which is not intercepted by any intermediaries.

A classical example of the ubiquity of direct effects (see Hesslow 1976; Cartwright 1989) tells the story of a birth-control pill that is suspect of producing thrombosis in women and, at the same time, has a negative indirect effect on thrombosis by reducing the rate of pregnancies (pregnancy is known to encourage thrombosis). In this example,

interest is focused on the direct effect of the pill because it represents a stable biological relationship that, unlike the total effect, is invariant to marital status and other social factors that may affect women's chances of getting pregnant or of sustaining pregnancy.

Another class of examples involves legal disputes over race or sex discrimination in hiring. Here, neither the effect of sex or race on applicants' qualification nor the effect of qualification on hiring are targets of litigation. Rather, defendants must prove that sex and race do not *directly* influence hiring decisions, whatever indirect effects they might have on hiring by way of applicant qualification.

In all these examples, the requirement of holding the mediating variables fixed must be interpreted as (hypothetically) setting these variables to constants by physical intervention, not by analytical means such as selection, conditioning, or adjustment. For example, it will not be sufficient to measure the association between the birth-control pill and thrombosis separately among pregnant and nonpregnant women and then aggregate the results. Instead, we must perform the study among women who became pregnant before the use of the pill and among women who prevented pregnancy by means other than the drug. The reason is that, by conditioning on an intermediate variable (pregnancy in the example), we may create spurious associations between X and Y even when there is no direct effect of X on Y. This can easily be illustrated in the model $X \rightarrow Z \leftarrow U \rightarrow Y$, where X has no direct effect on Y. Physically holding Z constant would permit no association between X and Y, as can be seen by deleting all arrows entering Z. But if we were to condition on Z, a spurious association would be created through U (unobserved) that might be construed as a direct effect of X on Y.

4.5.2 Direct Effects, Definition, and Identification

Controlling all variables in a problem is obviously a major undertaking, if not an impossibility. The analysis of identification tells us under what conditions direct effects can be estimated from nonexperimental data even without such control. Using our $do(x)$ notation (or $\hat{x}$ for short),

we can express the direct effect as follows.

Definition 4.5.1 (Direct Effect)
The direct effect of X on Y is given by $P(y|\hat{x}, \hat{s}_{XY})$, where S_{XY} is the set of all endogenous variables except X and Y in the system.

We see that the measurement of direct effects is ascribed to an ideal laboratory; the scientist controls for all possible conditions S_{XY} and need not be aware of the structure of the diagram or of which variables are truly intermediaries between X and Y. Much of the experimental control can be eliminated, however, if we know the structure of the diagram. For one thing, there is no need to actually hold *all* other variables constant; holding constant the direct parents of Y (excluding X) should suffice. Thus, we obtain the following equivalent definition of a direct effect.

Corollary 4.5.2 *The direct effect of X on Y is given by $P(y|\hat{x}, \widehat{pa}_{Y\setminus X})$, where $pa_{Y\setminus X}$ stands for any realization of the parents of Y, excluding X.*

Clearly, if X does not appear in the equation for Y (equivalently, if X is not a parent of Y), then $P(y|\hat{x}, \widehat{pa}_{Y\setminus X})$ defines a constant distribution on Y that is independent of x, thus matching our understanding of "having no direct effect." In general, assuming that X is a parent of Y, Corollary 4.5.2 implies that the direct effect of X on Y is identifiable whenever $P(y|\widehat{pa}_Y)$ is identifiable. Moreover, since the conditioning part of this expression corresponds to a plan in which the parents of Y are the control variables, we conclude that a direct effect is identifiable whenever the effect of the corresponding parents' plan is identifiable. We can now use the analysis of Section 4.4 and apply the graphical criteria of Theorems 4.4.1 and 4.4.6 to the analysis of direct effects. In particular, we can state our next theorem.

Theorem 4.5.3 *Let $PA_Y = \{X_1, \ldots, X_k, \ldots, X_m\}$. The direct effect of any X_k on Y is identifiable whenever the conditions of Corollary 4.4.5 hold for the plan $(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m)$ in some admissible ordering of the variables. The direct effect is then given by* (4.9).

Theorem 4.5.3 implies that if the effect of one parent of Y is identifiable then the effect of every parent of Y is identifiable as well. Of course, the magnitude of the effect would differ from parent to parent, as seen in (4.9).

The following corollary is immediate.

Corollary 4.5.4 *Let X_j be a parent of Y. The direct effect of X_j on Y is, in general, nonidentifiable if there exists a confounding arc that embraces any link $X_k \rightarrow Y$.*

4.5.3 Example: Sex Discrimination in College Admission

To illustrate the use of this result, consider the study of Berkeley's alleged sex bias in graduate admission (Bickel et al. 1975), where data showed a higher rate of admission for male applicants overall but, when broken down by departments, a slight bias toward female applicants. The explanation was that female applicants tend to apply to the more competitive departments, where rejection rates are high; based on this finding, Berkeley was exonerated from charges of discrimination. The philosophical aspects of such reversals, known as Simpson's paradox, will be discussed more fully in Chapter 6. Here we focus on the question of whether adjustment for department is appropriate for assessing sex discrimination in college admission. Conventional wisdom has it that such adjustment is appropriate because "We know that applying to a popular department (one with considerably more applicants than positions) is just the kind of thing that causes rejection" [Cartwright, 1983, p. 38], but we will soon see that additional factors should be considered.

Let us assume that the relevant factors in the Berkeley example are configured as in Figure 4.9, with the following interpretation of the variables:

$X_1 =$ applicant's gender;

$X_2 =$ applicant's choice of department;

$Z =$ applicant's career objectives;

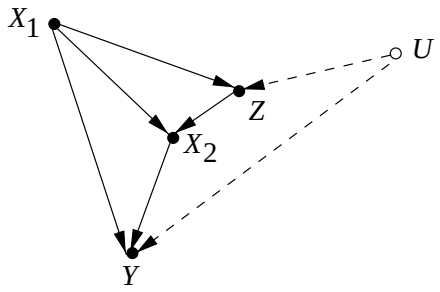

Figure 4.9: Causal relationships relevant to Berkeley's sex-discrimination study; adjusting for department choice (X_2) or career objective (Z) (or both) would be inappropriate in estimating the direct effect of gender on admission. The appropriate adjustment is given in (4.11).

$Y =$ admission outcome (accept/reject);

$U =$ applicant's aptitude (unrecorded).

Note that U affects applicant's career objective and also the admission outcome (say, through verbal skills (unrecorded)). Adjusting for department choice amounts to computing the following expression:

$$E_{x_2} P(y|\hat{x}_1, x_2) = \sum_{x_2} P(y|x_1, x_2)P(x_2). \tag{4.10}$$

In contrast, the direct effect of X_1 on Y, as given by (4.8), reads

$$P(y|\hat{x}_1, \hat{x}_2) = \sum_{z} P(y|z, x_1, x_2)P(z|x_1). \tag{4.11}$$

It is clear that the two expressions may differ substantially. The first measures the (average) effect of sex on admission among applicants to a given department, a quantitiy that is sensitive to the fact that some gender-department combinations may be associated with high admission rates merely because such combinations are indicative of certain aptitude (U) that was left unrecorded. The second expression eliminates such spurious associations by separately adjusting for career objectives (Z) in each of the two genders.

To verify that (4.10) does not properly measure the direct effect of X_1 on Y, we note that the expression depends on the value of X_1 even in cases where the arrow between X_1 and Y is absent. (4.11), on the other hand, becomes insensitive to x_1 in such cases—an exercise that we leave for the reader to verify.[8]

To cast this analysis in a concrete numerical setting, let us imagine a college consisting of two departments, A and B, both admitting students on the basis of qualification, Q, alone. Let us further assume (i) that the applicant pool consists of 100 males and 100 females and (ii) that 50 applicants in each gender have high qualifications (hence are admitted) and 50 have low qualifications (hence are rejected). Clearly, this college cannot be accused of sex discrimination.

A different result would surface, however, if we adjust for departments while ignoring qualifications, which amounts to using (4.10) to estimate the effect of gender on admission. Assume that the nature of the departments is such that *all and only* qualified male applicants apply to department A, while all females apply to department B (see Table 4.1).

	Males		Females		Total	
	Admitted	Applied	Admitted	Applied	Admitted	Applied
Dept. A	50	50	0	0	50	50
Dept. B	0	50	50	100	50	150
Unadjusted	50%		50%		50%	
Adjusted	25%		37.5%			

Table 4.1: Admission rate among males and females in each department.

We see from the table that adjusting for department would falsely indicate a bias of $37.5 : 25$ $(= 3 : 2)$ in favor of female applicants. An unadjusted (sometimes called "crude") analysis happens to give the correct result in this example—50% admission rate for males and females

[8] Hint: Factorize $P(y, u, z|\hat{x_1}, \hat{x_2})$ using the independencies in the graph and eliminate u as in the derivation of (3.29).

alike—thus exonerating the school from charges of sex discrimination.

Our analysis is not meant to imply that the Berkeley study of Bickel et al. (1975) is defective, or that adjustment for department was not justified in that study. The purpose is to emphasize that no adjustment is guaranteed to give an unbiased estimate of causal effects, direct or indirect, absent a careful examination of the causal assumptions that ensure identification. Theorem 4.5.3 provides us with the understanding of those assumptions and with a mathematical means of expressing them. We note that if applicants' qualifications were not recorded in the data, then the direct effect of gender on admission will not be identifiable unless we can measure some proxy variable that stands in the same relation to Q as Z stands to U in Figure 4.9.

4.5.4 Average Direct Effects

Readers versed in structural equation models (SEMs) will note that, in linear systems, the direct effect $P(Y|\hat{x}, \widehat{pa}_{Y\setminus X})$ is fully specified by the path coefficient attached to the link from X to Y (see (5.24) for mathematical definition); therefore, the direct effect is independent of the values $pa_{Y\setminus X}$ at which we hold the other parents of Y. In nonlinear systems, those values would, in general, modify the effect of X on Y and thus should be chosen carefully to represent the target policy under analysis. For example, the direct effect of a pill on thrombosis would most likely be different for pregnant and nonpregnant women. Epidemiologists call such differences "effect modification" and insist on separately reporting the effect in each subpopulation.

Although the direct effect is sensitive to the levels at which we hold the parents of the outcome variable, it is sometimes meaningful to average the direct effect over those levels. For example, if we wish to assess the degree of discrimination in a given school without reference to specific departments, we can compute the difference

$$P(\text{admission}|\widehat{\text{male}}, \widehat{\text{dept}}) - P(\text{admission}|\widehat{\text{female}}, \widehat{\text{dept}})$$

and average this difference over all departments. This average measures the increase in admission rate in a hypothetical experiment in which we instruct all female candidates to retain their department preferences

but change their gender identification (on the application form) from female to male.

In general, the average direct effect can be defined as a set of probabilities

$$\sum_{pa_{Y\setminus X}} P(y|\hat{x}, \widehat{pa}_{Y\setminus X})P(pa_{Y\setminus X}),$$

one for each level of X.

Several variants of this definition may be used when X affects other parents of Y. For example, we may wish to assess the average change in $E(Y)$ induced by changing X from x to x' while keeping the other parents of Y constant at whatever value they obtain under $do(x)$. The appropriate expression for this change is

$$\begin{aligned}\Delta_{x,x'}(Y) \;=\; & \sum_{pa_{Y\setminus X}} [E(Y|do(x'), do(pa_{Y\setminus X})) \\ & -E(Y|do(x), do(pa_{Y\setminus X}))]P(pa_{Y\setminus X}|do(x)).\end{aligned}$$

This expression represents what we actually wish to measure in race or sex discrimination cases, where we are instructed to assess the effect of one factor (X) while keeping "all other factors constant."

Acknowledgment

Sections 4.3 and 4.4 are based, respectively, on collaborative works with David Galles and James Robins.