Chapter 4

Models of Causal Exposure and Identification Criteria for Conditioning Estimators

In this chapter, we present the basic conditioning strategy for the estimation of causal effects. We first provide an account of the two basic implementations of conditioning – balancing the determinants of the cause of interest and adjusting for other causes of the outcome – using the language of "back-door paths." After explaining the unique role that collider variables play in systems of causal relationships, we present what has become known as the *back-door criterion* for sufficient conditioning to identify a causal effect. To bring the back-door criterion into alignment with related guidance based on the potential outcome model, we then present models of causal exposure, introducing the treatment assignment and treatment selection literature from statistics and econometrics. We conclude with a discussion of the identification and estimation of conditional average causal effects by conditioning.

4.1 Conditioning and Directed Graphs

In Section 1.5, we introduced the three most common approaches for the estimation of causal effects, using language from the directed graph literature: (1) conditioning on variables that block all back-door paths from the causal variable to the outcome variable, (2) using exogenous variation in an appropriate instrumental variable to isolate covariation in the causal variable and the outcome variable, and (3) establishing the exhaustive and isolated mechanism that intercepts the effect of the causal variable on the outcome variable and then calculating the causal effect as it propagates through the mechanism. In this chapter, we consider the first of these strategies, which motivates the basic matching, regression, and weighted regression techniques that we will present in Chapters 5, 6, and 7.

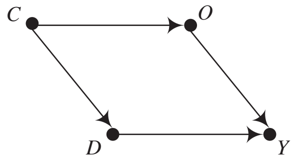


Figure 4.1 A graph in which the causal effect of D on Y is confounded by the back-door path $D \leftarrow C \rightarrow O \rightarrow Y$.

In Chapter 3, we explained the motivation for and execution of very simple conditioning estimators (see Section 3.2.3). In this chapter, we provide a more complete explanation of when, why, and how conditioning estimators will succeed in delivering estimates that can be given causal interpretations. We first reintroduce conditioning in a way that will reorient our perspective away from the "confounder variable" perspective of Chapter 3 to the more complete "back-door path" perspective that we will use in the remainder of this book.

4.1.1 From Confounders to Back-Door Paths

Consider Figure 4.1, which is an elaboration of the canonical confounding graph presented earlier in Figure 3.4(a). For this figure, the intermediate observed variable, O, expands the single-edge causal effect $C \rightarrow Y$ in Figure 3.4(a) to the directed path $C \rightarrow O \rightarrow Y$ in Figure 4.1. We noted in our prior discussion of Figure 3.4(a) that conditioning on C would allow us to generate a consistent and unbiased estimate of the causal effect of D on Y. This result holds for Figure 4.1 as well. However, for Figure 4.1, one could instead condition on O and achieve the same result. In fact, one could condition on both C and O as a third alternative.

The key goal of a conditioning strategy is not to adjust for any particular confounder but rather to remove the portion of the total association between D and Y that is noncausal. For Figure 4.1, the strategy to adjust for C is often referred to as "balancing the determinants of treatment assignment," and it is the standard motivation for matching estimators of causal effects. The alternative strategy to adjust for O is often referred to as "adjusting for all other causes of the outcome," and it is the standard motivation for regression estimators of causal effects.

Pearl characterizes both strategies in a novel way, using the language of back-door paths. As defined in Section 3.2.1, a *path* is any sequence of edges pointing in any direction that connects one variable to another. We now formally define a particular type of path that we have invoked informally already: A *back-door path* is a path between any causally ordered sequence of two variables that begins with a directed edge that points to the first variable.[1] For the directed graph in Figure 4.1, two paths connect D and Y: $D \leftarrow C \rightarrow O \rightarrow Y$ and $D \rightarrow Y$. The path $D \leftarrow C \rightarrow O \rightarrow Y$ is a

[1] "Causally ordered" means that the first variable causes the second variable by a directed path of some length. For the example discussed in this paragraph, D and Y are causally ordered because D causes Y by $D \rightarrow Y$. C and Y are also causally ordered because C causes Y by $C \rightarrow O \rightarrow Y$. D and O are *not* causally ordered because the only two paths that connect them ($D \leftarrow C \rightarrow O$ and $D \rightarrow Y \leftarrow O$) are not directed paths. Both of these paths have edges pointing in two directions. Recall also that we

back-door path because it begins with a directed edge pointing to D. Likewise, the path $D \rightarrow Y$ is not a back-door path because it does not begin with a directed edge pointing to D.[2]

In Pearl's language, the observed association between D and Y does not identify the causal effect of D on Y in Figure 4.1 because the total association between D and Y is an unknown composite of the true causal effect $D \rightarrow Y$ and a noncausal association between D and Y that is generated by the back-door path $D \leftarrow C \rightarrow O \rightarrow Y$. Fortunately, conditioning on C, O, or both C and O will block the back-door path, leaving within-stratum associations between D and Y that can be given causal interpretations (and can also be suitably averaged to obtain consistent and unbiased estimates of average causal effects of various types). The remainder of this chapter explains this result, as well as many others that are more complex.

4.1.2 Conditioning and Collider Variables

As a technique for estimating causal effects, conditioning is a very powerful and very general strategy. But, it is a much more complicated procedure in general than is suggested by our discussion of the simple directed graphs in Figures 3.4(a) and 4.1. Many of the complications arise when collider variables are present, and Pearl has explained systematically how to resolve these complications.

Recall that a variable is a collider along a particular path if it has two edges pointing directly to it, such as C in the path $A \rightarrow C \leftarrow B$ in Figure 3.3(c). For conditioning estimators of causal effects, collider variables must be handled with caution. Conditioning on a collider variable that lies along a back-door path does not help to block the back-door path but instead creates new associations.

The reasoning here is not intuitive, but it can be conveyed by a simple example with the mutual causation graph in Figure 3.3(c). Suppose that the population of interest is a set of applicants to a particular selective college and that C indicates whether applicants are admitted or rejected (i.e., $C = 1$ for admitted applicants and $C = 0$ for rejected applicants). Admissions decisions at this hypothetical college are determined entirely by two characteristics of students that are known to be independent within the population of applicants: SAT scores and a general rating of motivation based on an interview. These two factors are represented by A and B in Figure 3.3(c). Even though SAT scores and motivation are unrelated among applicants in general, they are not unrelated when the population is divided into admitted and rejected applicants. Among admitted applicants, the most motivated students will have lower than average SAT scores, and the least motivated students will have higher than average SAT scores. Thus, the college's sorting of applicants generates a pool of admitted students within which SAT scores and motivation are negatively related.[3]

stipulated in Chapter 3 that we will only consider acyclic graphs in this book. Without cycles, causal order is easily discerned by inspecting the directed paths in the graph.

[2] Recall that a directed path is a path in which all edges point in the same direction. Because all back-door paths have edges pointing in two directions, back-door paths are not directed paths. However, they can contain directed paths, such as the directed path $C \rightarrow O \rightarrow Y$ that is embedded in the back-door path $D \leftarrow C \rightarrow O \rightarrow Y$.

[3] A negative correlation will emerge for rejected students as well if (1) SAT scores and motivation have similarly shaped distributions and (2) both contribute equally to admissions decisions. As these

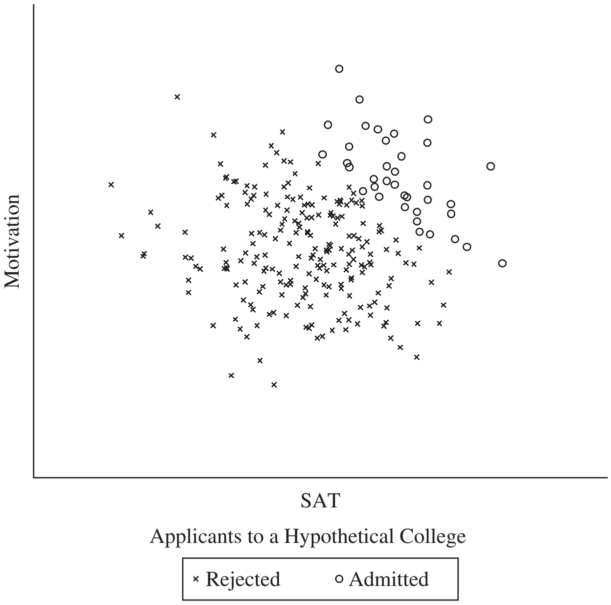


Figure 4.2 Simulation of conditional dependence within values of a collider variable.

This example is depicted in Figure 4.2 for 250 simulated applicants to this hypothetical college. For this set of applicants, SAT and motivation have a very small positive correlation of .035.[4] Offers of admission are then determined by the sum of SAT and motivation and granted to the top 15 percent of applicants (as shown in the upper right-hand portion of Figure 4.2).[5] Among admitted applicants, the correlation between SAT and motivation is −.641, whereas among rejected applicants the correlation between SAT and motivation is −.232. Thus, within values of the collider (the admissions decision), SAT and motivation are negatively related.

As Pearl documents comprehensively with a wide range of hypothetical examples, this is a very general feature of causal relationships and is present in many real-world applications. Elwert and Winship (2014) present many examples from the social science literature. In the next section, we show that care must be taken when attempting to estimate a causal effect by conditioning because conditioning on a collider variable can spoil an analysis.

conditions are altered, other patterns can emerge for rejected students, such as if admissions decisions are a nonlinear function of SAT and motivation.

[4] The values for SAT and motivation are 250 independent draws from standard normal variables. The draws result in an SAT variable with mean of .007 and a standard deviation of 1.01 as well as a motivation variable with mean of −.053 and a standard deviation of 1.02. Although the correlation between SAT and motivation is a small positive value for this simulation, we could drive the correlation arbitrarily close to 0 by increasing the number of applicants for the simulation.

[5] Admission is offered to the 37 of 250 students (14.8 percent) whose sum of SAT and motivation is greater than or equal to 1.5.

4.2 The Back-Door Criterion

With his language of back-door paths, colliders, and descendants, Pearl has developed what he labels the *back-door criterion* for determining whether or not conditioning on a given set of observed variables will identify the causal effect of interest.[6] The overall goal of a conditioning strategy guided by the back-door criterion is to block all paths that generate noncausal associations between the causal variable and the outcome variable without inadvertently blocking any of the paths that generate the causal effect itself. In practice, a conditioning strategy that utilizes the back-door criterion is implemented in two steps:

> Step 1: Write down the back-door paths from the causal variable to the outcome variable, determine which ones are unblocked, and then search for a candidate conditioning set of observed variables that will block all unblocked back-door paths.
>
> Step 2: If a candidate conditioning set is found that blocks all back-door paths, inspect the patterns of descent in the graph in order to verify that the variables in the candidate conditioning set do not block or otherwise adjust away any portion of the causal effect of interest.

This two-step procedure is justified by the following reasoning, which constitutes Pearl's back-door criterion:

> **Back-Door Criterion**
>
> If one or more back-door paths connect the causal variable to the outcome variable, the causal effect is identified by conditioning on a set of variables Z if
>
> > Condition 1. All back-door paths between the causal variable and the outcome variable are blocked after conditioning on Z, which will always be the case if each back-door path
> >
> > (a) contains a chain of mediation $A \to C \to B$, where the middle variable C is in Z, or
> > (b) contains a fork of mutual dependence $A \leftarrow C \to B$, where the middle variable C is in Z, or
> > (c) contains an inverted fork of mutual causation $A \to C \leftarrow B$, where the middle variable C and all of C's descendants are *not* in Z;

[6] The back-door criterion is meant to be used only when the causal effect of interest is specified as a component of a graph that is a Markovian causal model, or would be if all variables represented in the graph were observed; see Pearl's causal Markov condition for the existence of a causal model (Pearl 2009, section 1.4.2, theorem 1.4.1), as well as our discussion in the appendix to Chapter 3. This requirement can be weakened when the graph is only a locally Markovian causal model, as long as the underspecified causal relations are irrelevant to an evaluation of the back-door criterion for the particular causal effect under consideration. All of the examples we utilize in this book meet these conditions.

> and
>
> Condition 2. No variables in Z are descendants of the causal variable that lie on (or descend from other variables that lie on) any of the directed paths that begin at the causal variable and reach the outcome variable.[7]

To explain the back-door criterion, we will first consider Conditions 1(a), (b), and (c), using examples where Condition 2 is met by default because the only descendant of the causal variable D in the graph is the outcome variable Y (and where we assume that the analyst does not consider conditioning on Y itself).

Conditions 1(a) and 1(b) of the back-door criterion should be clear as stated. Return one last time to the simple example in Figure 4.1. Here, there is a single back-door path, $D \leftarrow C \rightarrow O \rightarrow Y$, which includes within it both a fork of mutual dependence ($D \leftarrow C \rightarrow O$) and a chain of mediation ($C \rightarrow O \rightarrow Y$). By the back-door criterion, conditioning on C blocks the path $D \leftarrow C \rightarrow O \rightarrow Y$ because C is the middle variable in a fork of mutual dependence. Likewise, conditioning on O blocks the path $D \leftarrow C \rightarrow O \rightarrow Y$ because O is the middle variable in a chain of mediation. As a result, the candidate conditioning set meets Pearl's back-door criterion if Z is C, O, or both C and O.

Condition 1(c), however, is quite different than Conditions 1(a) and 1(b) and is not intuitive. It states instead that the set of candidate conditioning variables Z *cannot* include collider variables that lie on back-door paths.[8] Consider the following example. A common but poorly justified practice in the social sciences is to salvage a regression model from suspected omitted-variable bias by adjusting for an endogenous variable that can be represented as a proxy for the omitted variable that is unobserved. In many cases, this strategy will fail because the endogenous variable is usually a collider.

Suppose that an analyst is confronted with a directed graph similar to the one in Figure 3.4(b), in which the causal effect of D on Y is confounded by an unobserved variable, such as C. When in this situation, researchers often argue that the effects of the unobserved confounder can be decomposed in principle into a lagged process, using a prior variable for the outcome, Y_{t-1}, and two separate unobserved variables, U and V, as in Figure 4.3. For this graph, there are two back-door paths from D to Y:

1. $D \leftarrow V \rightarrow Y_{t-1} \rightarrow Y$ and
2. $D \leftarrow V \rightarrow Y_{t-1} \leftarrow U \rightarrow Y$.

[7] This representation of the back-door criterion is a combination of Pearl's definition of *d-separation* (Pearl 2009:16–17), his original specification of the back-door criterion (Pearl 2009:79), and the generalization of the back-door criterion that was developed and labeled the "adjustment criterion" by Shpitser, VanderWeele, and Robins (2010). In an appendix to this chapter, we clarify our specification of Condition 2, as incorporated from the adjustment criterion. In brief, for Pearl's original back-door criterion, Condition 2 requires more simply (but overly strongly) that no variables in Z can be descendants of the causal variable.

[8] Because the "or" in the Conditions 1(a), (b), and (c) of the back-door criterion is inclusive, one can condition on colliders and still satisfy the back-door criterion if the back-door paths along which the colliders lie are otherwise blocked because Z satisfies Condition 1(a) or Condition 1(b) with respect to another variable on the same back-door path.

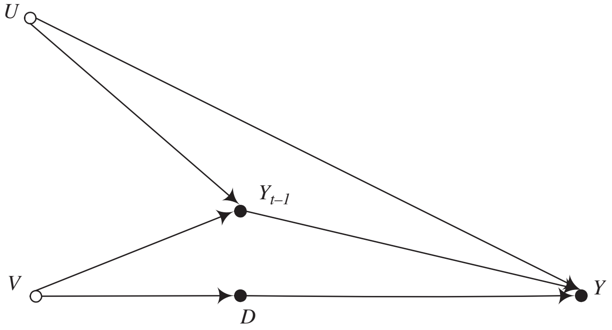


Figure 4.3 A causal diagram in which Y_{t-1} is a collider along a back-door path.

The lagged outcome variable Y_{t-1} lies on both of these back-door paths, but Y_{t-1} does not satisfy the back-door criterion. Notice first that Y_{t-1} blocks the first back-door path because, for this path, Y_{t-1} is the middle variable in a chain of mediation, $V \rightarrow Y_{t-1} \rightarrow Y$. But, for the second path, $D \leftarrow V \rightarrow Y_{t-1} \leftarrow U \rightarrow Y$, Y_{t-1} is a collider because it is the middle variable in an inverted fork of mutual causation, $V \rightarrow Y_{t-1} \leftarrow U$. Accordingly, conditioning on Y_{t-1} would eliminate part of the back-door association between D and Y because Y_{t-1} blocks the first back-door path $D \leftarrow V \rightarrow Y_{t-1} \rightarrow Y$. But, at the same time, conditioning on Y_{t-1} would create a new back-door association between D and Y because conditioning on Y_{t-1} unblocks the second back-door path $D \leftarrow V \rightarrow Y_{t-1} \leftarrow U \rightarrow Y$.

How can conditioning on a collider unblock a back-door path? To see the answer to this question, recall the discussion of conditioning in reference to Figure 3.3(c) and then as demonstrated in Figure 4.2. There, with the example of SAT and motivation effects on a hypothetical admissions decision to a college, we explained why conditioning on a collider variable induces an association between those variables that the collider is dependent on. That point applies here as well, when the causal effect of D on Y in Figure 4.3 is considered. Conditioning on a collider that lies along a back-door path unblocks the back-door path in the sense that it creates an association between D and Y within at least one of the subgroups enumerated by the collider.

Consider the slightly more complex example that is presented in Figure 4.4 (which is similar to Figure 1.1, except that the bidirected edges that signified unspecified and unobserved common causes have been replaced with two specific unobserved variables, U and V). Suppose, again, that we wish to estimate the causal effect of D on Y. For this directed graph, there are two back-door paths between D and Y:

1. $D \leftarrow A \leftarrow V \rightarrow F \rightarrow Y$ and
2. $D \leftarrow B \leftarrow U \rightarrow A \leftarrow V \rightarrow F \rightarrow Y$.

Notice that A is a collider variable in the second back-door path but not in the first back-door path. As a result, the first back-door path generates a noncausal association between D and Y, but the second back-door path does not. We therefore want to block the first path without unblocking the second path.

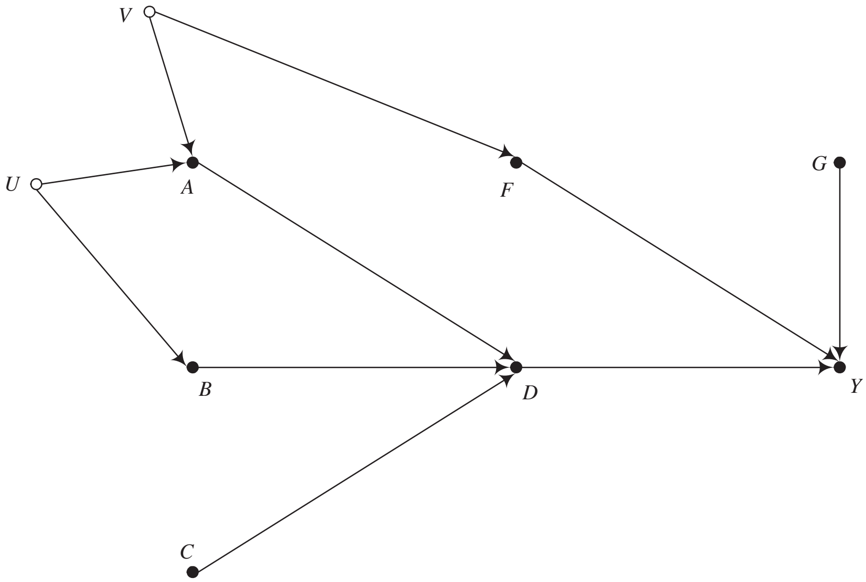


Figure 4.4 A causal diagram in which A is a collider on a back-door path.

For this example, there are two entirely different and effective conditioning strategies available that will identify the causal effect (numbers 1 and 3 in the following list) and a third one that may appear to work but that will fail (number 2 in the following list):

1. F is the middle variable in a chain of mediation, $V \rightarrow F \rightarrow Y$, for both back-door paths. As a result, F satisfies the back-door criterion, and conditioning on F identifies the causal effect of D on Y.

2. A is a middle variable in a chain of mediation, $D \leftarrow A \leftarrow V$, for the first back-door path. However, A is a collider variable for the second back-door path because it is the middle variable in a fork of mutual causation, $U \rightarrow A \leftarrow V$. As a result, A alone does not satisfy the back-door criterion. Conditioning on A does not identify the causal effect of D on Y, even though A lies along both back-door paths. Conditioning on A would unblock the second back-door path and thereby create a new, noncausal, back-door association between D and Y.

3. A is a middle variable in a chain of mediation, $D \leftarrow A \leftarrow V$, for the first back-door path. Likewise, B is a middle variable in a chain of mediation, $D \leftarrow B \leftarrow U$, for the second back-door path. Thus, even though A blocks only the first back-door path (and, in fact, conditioning on it unblocks the second back-door path), conditioning on B blocks the second back-door path. As a result, A and B together (but not alone) satisfy the back-door criterion, and conditioning on them together identifies the causal effect of D on Y.

In sum, for this example the causal effect can be identified by conditioning in one of two minimally sufficient ways: either condition on F or condition on both A and B.[9]

Now, we need to consider the complications introduced by descendants, as stipulated in both Condition 1(c) and Condition 2. Notice that Condition 1(c) states that neither C nor the descendants of C can be in Z. In words, conditioning on a collider *or* the descendant of a collider that lies on a back-door path will unblock the back-door path. Consider an extension of our prior analysis of Figure 4.3. For the graph in Figure 4.5, there are again two back-door paths from D to Y:

1. $D \leftarrow V \rightarrow Y_{t-2} \rightarrow Y_{t-1} \rightarrow Y$ and
2. $D \leftarrow V \rightarrow Y_{t-2} \leftarrow U \rightarrow Y$.

The first path does not contain any colliders and therefore confounds the causal effect of D on Y. The second back-door path, however, is blocked without any conditioning because Y_{t-2} is a collider on it. One might think, therefore, that conditioning on Y_{t-1} will block the first path without unblocking the second path. Condition 1(c) rules out this possibility because Y_{t-1} is a descendant of Y_{t-2}, and the latter is a collider on an already blocked path. The reasoning here is straightforward. The descendant Y_{t-1} is simply a noisy version of Y_{t-2} because its structural equation is

$$Y_{t-1} = f_{Y_{t-1}}(Y_{t-2}, e_{Y_{t-1}}),$$

where the error term, $e_{Y_{t-1}}$, is (as usual) assumed to be independent of all else in the graph. As a result, conditioning on Y_{t-1} has the same consequences for the second back-door path as conditioning on Y_{t-2}.[10]

Having explained Conditions 1(a), (b), and (c) of the back-door criterion, we can now consider Condition 2, which states that none of the variables in the possible conditioning set Z can be descendants of the causal variable that block the causal effect of interest by lying on or descending from any of the directed paths that begin at the causal variable and reach the outcome variable. Recall that a directed path is a path in which all edges point in the same direction. In all prior examples discussed in this section, the only descendant of D has been Y, and thus Condition 2 has been met by default (under the assumption that the analyst has not considered conditioning on the outcome variable Y itself).

We now consider two examples where additional descendants of D are present. Figure 4.6 presents a graph that elaborates Figure 4.1, where now the total effect of D on Y is separated into an indirect pathway through a mediating variable, $D \rightarrow N \rightarrow Y$, and a remaining direct effect, $D \rightarrow Y$. We will discuss models with such mediating

[9] One can of course condition in three additional ways that also satisfy the back-door criterion: F and A, F and B, and F, A, and B. These conditioning sets include unnecessary and redundant conditioning.

[10] Hernn, Hernandez-Diaz, and Robins (2004) offer an excellent discussion of examples in epidemiology for which such descendants of colliders are a primary focus. For their types of examples, the outcome is "death," the collider on the back-door path (or elsewhere in the causal graph) is "getting sick enough to be admitted to a hospital for treatment," and the variable that is conditioned on is "in a hospital." Conditioning on "in a hospital" (by undertaking a study of hospital patients) induces associations between the determinants of sickness that can spoil standard analyses. Elwert (2013) and Elwert and Winship (2014) cover many of the same issues from a social science perspective. In an appendix to this chapter, we also provide additional discussion and examples.

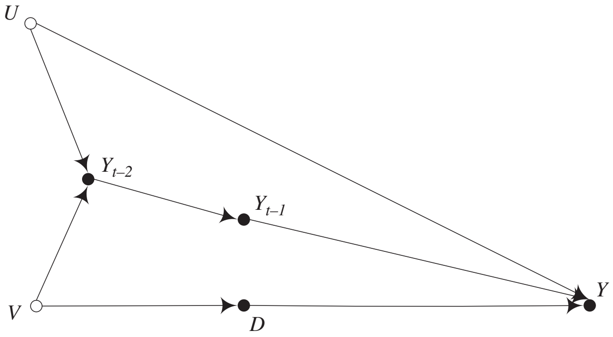


Figure 4.5 A causal diagram in which Y_{t-2} is a collider on a back-door path and Y_{t-1} is its descendant.

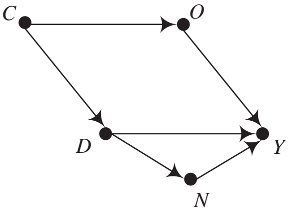


Figure 4.6 A confounded causal effect expressed as an indirect effect and a net direct effect.

mechanisms in substantial detail in Chapter 10, and for now we will use this graph only to introduce a discussion of Condition 2 of the back-door criterion.

As we have noted in this section and elsewhere (and, assuming now, without loss of generality that D is a two-valued treatment), conditioning on either C or O identifies the total average causal effect of D on Y, which we defined in Section 3.4 as $E[Y|do(D=1)]-E[Y|do(D=0)]$. The back-door criterion states that if, in addition to C and/or O, we also conditioned on N, the resulting estimate would no longer identify the casual effect of D on Y. The conditioning set of C, O, and N violates Condition 2 of the back-door criterion because N is a descendant of D that lies on a directed path, $D \to N \to Y$, that reaches Y. As a result, the back-door criterion indicates that the analyst should not condition on N.

We suspect that this conclusion would be clear to readers without consulting the back-door criterion (by reasoning that adjustment for N would rob the total causal effect of D on Y of some of its magnitude, leaving only a partial direct effect that is not equal to $E[Y|do(D=1)]-E[Y|do(D=0)]$).[11] Such reasoning is correct, and Condition 2 of the back-door criterion is, in part, a formalization of this intuition.

[11] In this section, we are considering only how to apply the back-door criterion when the goal is to estimate the total effect of the cause D on the outcome Y. If, instead, the analyst is interested in estimating the direct effect of D on Y, net of the indirect effect of D on Y through N, then adjustment for N is necessary. VanderWeele (in press) provides a comprehensive treatment of the identification and estimation of direct effects. We will return to these issues in Chapter 10, where we will consider how to identify causal effects using generative mechanisms.

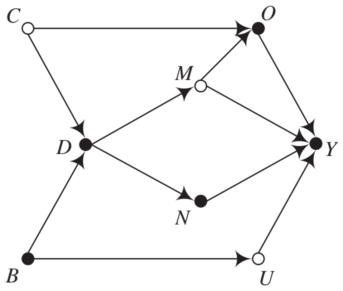


Figure 4.7 A graph where the effect of D on Y is not identified by conditioning on O and B because O is a descendant of D.

However, Condition 2 is more finely articulated than this intuition alone. It invalidates conditioning sets that adjust away any portion of the causal effect of interest, not just those portions that are carried by variables that mediate the causal effect. Just as conditioning on the descendant of a collider has the same consequences as conditioning on the collider itself, conditioning on a descendant of a variable that lies on a directed path from the causal variable to the outcome variable has the same consequences for identification results as conditioning directly on the variable from which it descends.[12]

Consider Figure 4.7, where we are again interested in the effect of D on Y and where we have now specified two mediating variables, M and N, that completely characterize the total effect of D on Y. For this graph, three back-door paths connect D to Y:

1. $D \leftarrow C \rightarrow O \rightarrow Y$,

2. $D \leftarrow C \rightarrow O \leftarrow M \rightarrow Y$, and

3. $D \leftarrow B \rightarrow U \rightarrow Y$.

In addition, the variables C, M, and U are unobserved, and thus they are not available to use in a conditioning estimator.

Suppose that one decides to condition on both O and B. First, note that these two variables do not satisfy Conditions 1(a), (b), and (c) of the back-door criterion.

[12] Notice further that Condition 2 applies to descendants of the cause that lie on or descend from variables on directed paths that begin at the causal variable and "reach the outcome variable" rather than "end at the outcome variable." The word "reach" has been chosen to allow for Condition 2 to invalidate conditioning sets that include descendants of the cause that are also descendants of the outcome. Suppose that an additional observed variable W is added to Figure 4.6 along with a directed edge from Y to W, as in $Y \rightarrow W$. The new variable W is a descendant of both D and Y via the two directed paths $D \rightarrow Y \rightarrow W$ and $D \rightarrow N \rightarrow Y \rightarrow W$ that begin at D and reach Y (and, in this case, carry on to W). There is, of course, no reason to condition on W in an attempt to estimate the causal effect of D and Y because W does not lie on a back-door path from D to Y that confounds the causal effect of interest. If an analyst does do so, by adding W to a conditioning set that includes C and/or O, then Conditions 1(a), (b), and (c) of the back-door criterion are met but Condition 2 is not. The variable W is simply a noisy version of Y, and the causal effect of D on Y is therefore embedded within it. Conditioning on W will, in fact, mistakenly suggest that D has no casual effect on Y because within strata defined by W, D and Y are independent (assuming no measurement error and an infinite sample that enables fully nonparametric estimation).

Conditioning on O will block path 1, and conditioning on B will block path 3. However, conditioning on O will unblock path 2 because it is a collider on an already blocked back-door path.

What we want to stress now is that the candidate conditioning set of O and B also does not satisfy Condition 2 of the back-door criterion. O is a descendant of D on a directed path, $D \rightarrow M \rightarrow O \rightarrow Y$, that reaches Y. Furthermore, O is a descendant of M, via $M \rightarrow O$, and M is a variable that lies on another directed path, $D \rightarrow M \rightarrow Y$, that begins at D and reaches Y.

Notice also that the graph in Figure 4.7 is realistic and not contrived. If a researcher follows the (sometimes wrong) conditioning strategy of "adjusting for all other causes of the outcome," and the researcher has not drawn the full causal graph in Figure 4.7, then the researcher would almost certainly condition on O (even if the researcher wisely decides not to condition on the clearly endogenous observed variable N but does decide to condition on the second-order cause B in place of the unobserved direct cause U). In practice, other causes of an outcome that are observed are hard to resist conditioning on, even though many of them are descendants of the cause along directed paths that reach the outcome and thereby violate Condition 2 of the back-door criterion. If a researcher does not observe a mechanistic variable like M, the analyst may not recognize the endogeneity of O with respect to D (especially if the analyst comes to believe that the only mechanistic variable has been observed as N and that the remaining effect of D on Y is an unmediated direct effect).

Condition 2 of the back-door criterion is a general requirement for sufficient conditioning sets, and it is easy to apply. One need only examine each candidate conditioning variable to determine whether it lies on (or descends from a variable that lies on) the directed paths that begin at the causal variable and reach the outcome variable.

Although Condition 2 is easy to apply, we discuss additional examples in an appendix to this chapter that are more challenging to explain. In particular, we reconsider the graphs just presented, building toward a full reexamination of the graph in Figure 4.7, which is even more complex than our presentation here reveals. In the appendix, we show that a full assessment of the consequences of conditioning on descendants of the cause is enabled by drawing the graphs under magnification so that it is easier to recognize all such descendants as colliders. Although these more complex cases are interesting to examine, so as to understand the full range of identification challenges that graphical models help to explain, one need not fully understand them or absorb them to effectively adhere to conditioning strategies that are warranted by the back-door criterion.

The two key points of this section are the following. First, conditioning on variables that lie on back-door paths can be an effective strategy to identify a causal effect. If all back-door paths between the causal variable and the outcome variable are blocked after the conditioning is enacted, then back-door paths do not contribute to the association between the causal variable and the outcome variable. However, it must be kept in mind that conditioning on a collider (or a descendant of a collider) has the opposite effect. Any such conditioning unblocks already blocked back-door paths. And thus, when a conditioning strategy is evaluated, each back-door path must be assessed carefully because a variable can be a collider on one back-door path but not a collider on another.

Second, if a set of conditioning variables blocks all back-door paths, the analyst must then verify that no variables within the conditioning set block the causal effect of interest or otherwise mistakenly adjust it away. If none of the candidate conditioning variables are also descendants of the cause that lie on or descend from directed paths that begin at the causal variable and reach the outcome variable, then the causal effect is identified and a conditioning estimator is consistent and unbiased for the average causal effect.

Pearl's back-door criterion for evaluating conditioning strategies is a generalization (and therefore a unification) of various traditions for how to solve problems that are frequently attributed to omitted-variable bias. From our perspective, Pearl's framework is particularly helpful in two respects. First, it shows clearly that researchers do not need to condition on all omitted direct causes of an outcome variable in order to solve an omitted-variable bias problem. This claim is not new, of course, but Pearl's back-door criterion shows clearly why researchers need to condition on only a minimally sufficient set of variables that (a) renders all back-door paths blocked and (b) does not block the causal effect itself. Second, Pearl's framework shows how to think clearly about the appropriateness of conditioning on endogenous variables. Writing down each back-door path and then determining whether or not each endogenous variable is a collider along any of the back-door paths is a much simpler way to begin to consider the full complications of a conditioning strategy than other approaches.[13]

In the next section, we consider models of causal exposure that have been used in the counterfactual tradition, starting first with the statistics literature and carrying on to the econometrics literature. We will show that the assumptions often introduced in these two traditions to justify conditioning estimation strategies – namely, ignorability and selection on the observables – have close connections to the back-door criterion presented in this section.

[13] The back-door criterion is not the only available graphical guide for the selection of conditioning variables. Elwert (2013:256–61) offers a crisp summary of the alternatives. Among these, the *adjustment criterion* is the most general and is guaranteed to select all possible sufficient conditioning sets. Nonetheless, we make the case in the appendix to this chapter that the version of the back-door criterion that we present in the main text of this chapter has advantages relative to the completeness of the adjustment criterion. Among the other criteria, the *disjunctive cause criterion* of VanderWeele and Shpitser (2011) is also particularly useful because it can serve as a guide to conditioning when the analyst is unwilling to commit to a set of assumptions that enable a full directed graph to be drawn for the generation of the outcome. The disjunctive cause criterion instructs the analyst to adjust for all variables that are causes of the treatment or the outcome, but not those variables that are causes neither of the treatment nor of the outcome. The analyst therefore does not need to construct a directed graph, only take a position on whether the candidate conditioning variables should be assumed to be causes of the treatment, causes of the outcome, or neither. Of course, the price to be paid for the simplicity of the disjunctive cause criterion is that it will not necessarily identify the causal effect (because one cannot know about all sources of confounding if one cannot draw the full directed graph for the generation of the outcome) and can suggest redundant conditioning (because one does not need to condition on variables that lie on back-door paths that are already blocked by colliders, variables that lie on back-door paths that are already blocked by conditioning on other variables, or variables that do not lie on any back-door paths). Nonetheless, the disjunctive cause criterion will prevent the analyst from inducing noncausal associations between the treatment and outcome that would result from conditioning on the relevant colliders that already block back-door paths.

4.3 Models of Causal Exposure and Point Identification Based on the Potential Outcome Model

With this general presentation of the conditioning strategy in mind, return to the familiar case of a binary cause D and an observed outcome variable Y. As discussed in Chapter 2, for the potential outcome model we consider Y to have been generated by a switching process between two potential outcome variables, as in $Y = DY^1 + (1-D)Y^0$, where the causal variable D is the switch. To model variation in Y and relate it to the individual-level causal effects defined by the potential outcome variables Y^1 and Y^0, a model for the variation in D must be adopted. This is typically known in the statistics literature as "modeling the treatment assignment mechanism" and in the econometrics literature as "modeling the treatment selection mechanism."

In this section, we first consider the notation and language developed by statisticians, and we then turn to the alternative notation and language developed by econometricians. Although both sets of ideas are equivalent, they each have some distinct conceptual advantages. In showing both, we hope to deepen the understanding of each.

4.3.1 Treatment Assignment Modeling in Statistics

The statistics literature on modeling the treatment assignment mechanism is an outgrowth of experimental methodology and the implementation of randomization research designs. Accordingly, we begin by considering a randomized experiment for which the phrase "treatment assignment" remains entirely appropriate.

As discussed in Chapter 2, if treatment assignment is completely randomized by design, then the treatment indicator variable D is completely independent of the potential outcomes Y^0 and Y^1 as well as any function of them, such as the distribution of δ; see the earlier discussion of Equation (2.6). In this case, the treatment assignment mechanism is known because it is set by the researcher who undertakes the randomization. If the researcher wants treatment and control groups of approximately the same size, then $\Pr[D=1]$ is set to .5. Individual realized values of D for those in the study, denoted d_i generically, are then equal to 1 or 0 and are determined by the flip of a fair coin (or by a computer that runs Bernoulli trials for the random variable D with .5 as the probability).

To facilitate the transition to designs for observational research, consider a slightly more elaborate design where study subjects are stratified first by gender and then assigned with disproportionate probability to the treatment group if female. In this case, the treatment assignment protocol would instead be represented by two conditional probabilities, such as

$$\Pr[D=1|\text{Gender}=\text{Female}] = .7, \tag{4.1}$$

$$\Pr[D=1|\text{Gender}=\text{Male}] = .5. \tag{4.2}$$

These conditional probabilities are typically referred to as "propensity scores" in the literature because they indicate the propensity that an individual with specific characteristics will be observed in the treatment group. Although labeled propensity scores in the literature, they are nonetheless nothing more than conditional probabilities that lie within an interval bounded by 0 and 1. For this example, the propensity score is .7 for female subjects and .5 for male subjects. The general point is that for randomized experiments, the propensity scores are known to the researcher.

In contrast, a researcher with observational data does not possess a priori knowledge of the exact propensity scores that apply to all individuals. However, the researcher may know all of the characteristics of individuals that systematically determine their propensity scores, even if the researcher does not know the specific values of the propensity scores.[14] In this case, treatment assignment patterns are represented by a general conditional probability distribution,

$$\Pr[D=1|S], \tag{4.3}$$

where S now denotes all variables that systematically determine all treatment assignment patterns. Complete observation of S then allows a researcher to assert that treatment assignment is "ignorable" and then consistently estimate the average treatment effect (ATE), as we now explain.

The general idea here is that, within strata defined by S, the remaining variation in the treatment D is completely random and hence the process that generates this remaining variation is labeled "ignorable." The core of the concept of ignorability is the independence assumption that was introduced in Equation (2.6),

$$(Y^0, Y^1) \perp\!\!\!\perp D,$$

where the symbol $\perp\!\!\!\perp$ denotes independence. As defined by Rubin (1978), ignorability of treatment assignment holds when the potential outcomes are independent of the treatment dummy indicator variable, as in this case all variation in D is completely random. Ignorability also holds in the weaker case where

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid S \tag{4.4}$$

and

all variables in S are observed.

In words, the treatment assignment mechanism is ignorable when the potential outcomes (and any function of them, such as δ) are independent of the treatment variable, D, within strata defined by all combinations of values on all variables, S, that systematically determine all treatment assignment patterns. If some components of S are unobserved, the conditional independence condition in Equation (4.4) may still hold, but treatment assignment cannot be considered ignorable with respect to the observed data. In this case, treatment assignment must be regarded as nonignorable, even if it is known that it would be ignorable if all variables in S had instead been observed.[15]

[14] This would be the situation for the randomized experiment represented by Equations (4.1) and (4.2) if the experimentalist knew that the probability of being assigned to the treatment differed by gender (and only by gender) but had forgotten the values of .7 and .5.

[15] Rosenbaum and Rubin (1983a) defined strong ignorability to develop the matching literature, which we will discuss later. To Rubin's ignorability assumption, Rosenbaum and Rubin (1983a) required for strong ignorability that each subject have a nonzero probability of being assigned to both the treatment and the control groups. Despite these clear definitions, the term ignorability is

In practice, in order to assert that treatment assignment is ignorable for an observational study, a researcher would

1. determine from related studies and supportable assumptions grounded in theory what the components of S are,
2. measure each of the component variables in S, and
3. collect enough data to be able to consistently estimate outcome differences on the observed variable Y within strata defined by S.

This third step can be weakened if the data are merely sparse, as we will discuss when presenting models based on estimate propensity scores in Chapters 5 and 7. The key point is that a researcher does not need to know the exact propensity scores (i.e., what $\Pr[D=1|S=s]$ is equal to for all s), only that the systematic features of treatment assignment can be exhaustively accounted for by the data in hand on the characteristics of individuals. The naive estimator can then be calculated within strata defined by values of the variables in S, and a weighted average of these stratified estimates can be formed as a consistent estimate of the ATE.[16]

Consider the Catholic school example. It is well known that students whose parents self-identify as Catholic are more likely to be enrolled in Catholic schools than students whose parents self-identify as non-Catholic. Suppose that parents' religious identity is the only characteristic of students that systematically determines whether they attend Catholic schools instead of public schools. In this case, a researcher can consistently estimate the ATE by collecting data on test scores, students' school sector attendance, and parent's religious identification. A researcher would then estimate the effect of Catholic schooling separately by using the naive estimator within groups of students defined by parents' religious identification and then take a weighted average of these estimates based on the proportion of the population of interest whose parents self-identify as Catholic and as non-Catholic. In the words of the prior section, the researcher can generate consistent and unbiased estimates of the ATE by conditioning on parents' religious identification.

Assumptions of ignorability have a close connection to the back-door criterion of Pearl, given the shared centrality of the conditioning operation. Even so, it is important to recognize some differences. Suppose that we are confronted with the graph in Figure 4.8(a), which includes the causal effect $D \rightarrow Y$ but also the bidirected edge $D \leftarrow\!\!\text{- - - -}\!\!\rightarrow Y$. The most common solution is to build an explicit causal model that represents the variables that generate the bidirected edge between D and Y in Figure 4.8(a). The simplest such model is presented in Figure 4.8(b), where $D \leftarrow\!\!\text{- - - -}\!\!\rightarrow Y$ has been replaced with the back-door path $D \leftarrow S \rightarrow Y$. If S is observed, then conditioning on S will solve the causal inference problem, according to the back-door criterion

often defined in different ways in the literature. We suspect that this varied history of usage explains why Rosenbaum (2002) rarely uses the term in his monograph on observational data analysis, even though he is generally credited, along with Rubin, with developing the ignorability semantics in this literature. And it also explains why some of the most recent econometrics literature uses the words unconfoundedness and exogeneity for the same set of independence and conditional-independence assumptions (see Imbens 2004).

[16] Again, we assume that measurement error does not exist, which requires that step 2 be undertaken without error.

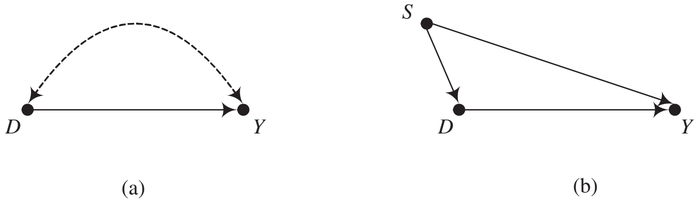


Figure 4.8 Causal diagrams in which treatment assignment is (a) nonignorable and (b) ignorable.

presented in the previous section. In the statistics literature, the same result is explained by noting that treatment assignment is ignorable when S is observed, which is then written as Equation (4.4).

When identification by back-door conditioning is feasible with the observed data, then treatment selection is ignorable with respect to the observed data. However, we do not mean to imply with this statement that the assertion of a valid ignorability assumption will lead the researcher to condition on the exact same variables suggested by the back-door criterion. Recall that the back-door criterion guides the researcher in selecting minimally sufficient conditioning sets. Ignorability can be asserted with respect to these sets or with respect to broader sets of conditioning variables. For an example, let S in Figure 4.8(b) be a set of variables. Suppose that, upon reflection, a researcher decides that one of the variables in S, S', has a direct effect on D but no direct effect on Y. Accordingly, it is appropriate to separate S' from the other variables in S and assert that S' is a cause of D that causes Y only indirectly through D.[17] In this case, S' does not generate any confounding because S' does not lie on a back-door path between D and Y. The minimally sufficient conditioning sets that are judged admissible by the back-door criterion would not include S', even though treatment assignment is ignorable with respect to S' and the remaining variables in S. Thus, when asserting ignorability with respect to the observed data, the researcher may decide to condition on S', even though such conditioning is unnecessary. After all, S' does determine treatment assignment and, hence, does structure the true propensity score. And it is possible that the strong assumption that S' does not have a direct effect on Y is incorrect.

We will begin to discuss specific techniques for conditioning estimators in Chapter 5, where we first present matching estimators of causal effects. But the immediate complications of undertaking a conditioning analysis strategy for the Catholic school example should be clear. How do we determine all of the factors that systematically determine whether a student enrolls in a Catholic school instead of a public school? And can we obtain measures of all of these factors? Attendance at a Catholic school is determined by more than just parents' religious self-identification, and some of these determinants are likely unmeasured. If this is the case, then the treatment assignment mechanism is likely to be nonignorable, as treatment selection is then a function of unobserved characteristics of students that generate confounding. These

[17] S' is an instrumental variable, as we will explain in full detail in Chapter 9.

issues are best approached by first drawing a directed graph. It may be that some of the unobserved determinants of treatment assignment do not generate confounding because (a) they do not lie on back-door paths, (b) they lie on back-door paths that are already blocked by colliders, or (c) they lie on back-door paths that can be blocked by conditioning on other variables that lie on the same back-door paths. If so, the researcher may be able to assert an ignorability assumption with respect to only a subset of the variables we have designated as S for this section. However, in most cases, the opposite challenge will dominate: The directed graph will show clearly that only a subset of the variables in S that generate confounding are observed, and the confounding that they generate cannot be eliminated by conditioning with the observed data. Treatment assignment is then nonignorable with respect to the observed data.

4.3.2 Treatment Selection Modeling in Econometrics

The econometrics literature also has a long tradition of analyzing causal effects, and this literature may be more familiar to social scientists. Whereas concepts such as ignorability are used somewhat infrequently in the social sciences, the language of selection bias is commonly used throughout the social sciences. This usage is due, in large part, to the energy that economists have devoted to exploring the complications of self-selection bias.

The selection-bias literature in econometrics is vast, but the most relevant piece that we focus on here is James Heckman's specification of the random-coefficient model for the treatment effects of training programs, which he attributes, despite the difference in substance, to Roy (1951). The clearest specification of this model was presented in a series of papers that Heckman wrote with Richard Robb (see Heckman and Robb 1985, 1986, 1989), but Heckman worked out many of these ideas in the 1970s. Using the notation we have adopted in this book, take Equation (2.2),

$$Y = DY^1 + (1-D)Y^0,$$

and then rearrange and relabel terms as follows:

$$\begin{aligned} Y &= Y^0 + (Y^1 - Y^0)D \\ &= Y^0 + \delta D \\ &= \mu^0 + \delta D + \upsilon^0, \end{aligned} \tag{4.5}$$

where $\mu^0 \equiv E[Y^0]$ and $\upsilon^0 \equiv Y^0 - E[Y^0]$. The standard outcome model from the econometrics of treatment evaluation simply reexpresses Equation (4.5) so that potential variability of δ across individuals in the treatment and control groups is relegated to the error term, as in

$$Y = \mu^0 + (\mu^1 - \mu^0)D + \{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}, \tag{4.6}$$

where $\mu^1 \equiv E[Y^1]$, $\upsilon^1 \equiv Y^1 - E[Y^1]$, and all else is as defined for Equation (4.5).[18] Note that, in evolving from Equation (2.2) to Equation (4.6), the definition of the

[18] The original notation is a bit different, but the ideas are the same. Without much usage of the language of potential outcomes, Heckman and Robb (1985, section 1.4) offered the following setup for the random coefficient model of treatment effects to analyze posttreatment earnings differences for a fictitious worker training example. For each individual i, the earnings of individual i if trained are

observed outcome variable Y has taken on the look and feel of a regression model.[19] The first μ^0 term is akin to an intercept, even though it is defined as $E[Y^0]$. The term $(\mu^1 - \mu^0)$ that precedes the first appearance of D is akin to a coefficient on the primary causal variable of interest D, even though $(\mu^1 - \mu^0)$ is defined as the true ATE, $E[\delta]$. Finally, the term in braces, $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$, is akin to an error term, even though it represents both heterogeneity of the baseline no-treatment potential outcome and of the causal effect, δ, and even though it includes within it the observed variable D.[20]

Heckman and Robb use the specification of the treatment evaluation problem in Equation (4.6), and many others similar to it, to demonstrate all of the major problems created by selection bias in program evaluation contexts when simple regression estimators are used. Heckman and Robb show why a regression of Y on D does not in general identify the ATE, in this case $(\mu^1 - \mu^0)$, when D is correlated with the population-level variant of the error term in braces in Equation (4.6), as would be the case when the size of the individual-level treatment effect, in this case $(\mu^1 - \mu^0) + \{\upsilon_i^0 + d_i(\upsilon_i^1 - \upsilon_i^0)\}$, differs among those who select the treatment and those who do not.

The standard regression strategy that prevailed in the literature at the time was to include additional variables in a regression model of the form of Equation (4.6), hoping

$$y_i^1 = \beta^1 + U_i^1,$$

and the earnings of individual i in the absence of training are

$$y_i^0 = \beta^0 + U_i^0,$$

(where we have suppressed subscripting on t for time from the original presentation and also shifted the treatment state descriptors from subscript to superscript position). With observed training status represented by a binary variable, d_i, Heckman and Robb then substitute the right-hand sides of these equations into the definition of the observed outcome in Equation (2.2) and rearrange terms to obtain

$$y_i = \beta^0 + (\beta^1 - \beta^0)d_i + U_i^0 + (U_i^1 - U_i^0)d_i,$$

which they then collapse into

$$y_i = \beta^0 + \bar{\alpha} d_i + \{U_i^0 + \varepsilon_i d_i\},$$

where $\bar{\alpha} \equiv \beta^1 - \beta^0$ and $\varepsilon_i \equiv U_i^1 - U_i^0$ (see Heckman and Robb 1985, equation 1.13). As a result, $\bar{\alpha}$ is the ATE, which we defined as $E[\delta]$ in Equation (2.3), and ε_i is the individual-level departure of δ_i from the ATE, $E[\delta]$. Although the notation in this last equation differs from the notation in Equation (4.6), the two equations are equivalent. Heckman and Vytlacil (2005, 2007) give a fully nonparametric version of this treatment selection framework, which we draw on below.

[19] Sometimes, Equation (4.6) is written as

$$Y = \mu^0 + [(\mu^1 - \mu^0) + (\upsilon^1 - \upsilon^0)]D + \upsilon^0$$

in order to preserve its random-coefficient interpretation. This alternative representation is nothing other than a more fully articulated version of Equation (4.5).

[20] Statisticians sometimes object to the specification of "error terms" because, among other things, they are said to represent a hidden assumption of linearity. In this case, however, the specification of this error term is nothing other than an expression of the definition of the individual-level causal effect as the linear difference between y_i^1 and y_i^0.

to break the correlation between D and the error term.[21] Heckman and Robb show that this strategy is generally ineffective with the data available on worker training programs because (1) some individuals are thought to enter the programs based on anticipation of the treatment effect itself and (2) none of the available data sources have measures of such anticipation. We will return to this case in detail in Chapter 6, where we discuss regression models.

To explain these complications, Heckman and Robb explore how effectively the dependence between D and the error term in Equation (4.6) can be broken. They proceed by proposing that treatment selection be modeled by specifying a latent continuous variable $\tilde{D}$ as

$$\tilde{D} = Z\phi + U, \tag{4.7}$$

where Z represents all observed variables that determine treatment selection, ϕ is a coefficient (or a vector of coefficients if Z includes more than one variable), and U represents both systematic unobserved determinants of treatment selection and completely random idiosyncratic determinants of treatment selection. The latent continuous variable $\tilde{D}$ in Equation (4.7) is then related to the treatment selection dummy, D, by

$$\begin{aligned} D &= 1 \quad \text{if } \tilde{D} \geq 0, \\ D &= 0 \quad \text{if } \tilde{D} < 0, \end{aligned}$$

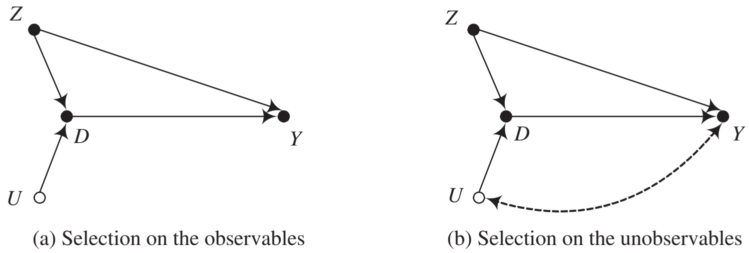
where the threshold 0 is arbitrary because the term U has no inherent metric (because it is composed of unobserved and possibly unknown variables).

To see the connection between this econometric specification and the one from the statistics literature introduced in the last section, first recall that statisticians typically specify the treatment selection mechanism as the general conditional probability distribution $\Pr[D = 1|S]$, where S is a vector of all systematic observed determinants of treatment selection.[22] This is shown in the graph in Figure 4.8(b). The corresponding causal diagram for the econometric selection equation is presented in two different graphs in Figure 4.9, as there are two scenarios corresponding to whether or not all elements of S have been observed as Z.

For the case in which Z in Equation (4.7) is equivalent to the set of variables in S in Equation (4.3), treatment selection is ignorable, as defined in Equation (4.4), because conditioning on Z is exactly equivalent to conditioning on S. In the econometric tradition, this situation would not, however, be referred to as a case for which treatment assignment/selection is ignorable. Rather, treatment selection would be characterized as "selection on the observables" because all systematic determinants of treatment selection are included in the observed treatment selection variables Z. This phrase is widely used by social scientists because it conveys the essential content

[21] Barnow, Cain, and Goldberger (1980:52) noted that "the most common approach" is to "simply assume away the selection bias after a diligent attempt to include a large number of variables" in the regression equation.

[22] When more specific, the basic model is usually a Bernoulli trial, in which $\Pr[D = 1|S = s]$ gives the specific probability of drawing a 1 and the complement of drawing a 0 for individuals with S equal to s.



(a) Selection on the observables

(b) Selection on the unobservables

Figure 4.9 Causal diagrams for the terminology from econometric modeling of treatment selection.

of the ignorability assumption: All systematic determinants of treatment selection have been observed.[23]

The scenario of selection on the observables is depicted in Figure 4.9(a). The variable S in Figure 4.8(b) is simply relabeled Z, and there are no back-door paths from D to Y other than the one that is blocked by Z. The remaining idiosyncratic random variation in D is attributed in the econometric tradition to a variable U, which is presented in Figure 4.9(a) as a cause of D that is conditionally independent of both Z and Y. This error term U represents nothing other than completely idiosyncratic determinants of treatment selection. It could therefore be suppressed in Figure 4.9(a), which would render this graph the same as the one in Figure 4.8(b).[24]

Now consider the case in which the observed treatment selection variables in Z are only a subset of the variables in S. In this particular case, some components of S enter into the treatment selection latent variable $\tilde{D}$ through the error term, U, of Equation (4.7). In this case, treatment selection is nonignorable. In the words of econometricians, "selection is on the unobservables" (or, more completely, "selection is on the observables Z and the unobservables U"). The scenario of selection on the unobservables is depicted in Figure 4.9(b), where there are now back-door paths from D to Y represented by $D \leftarrow U \dashleftarrow\dashrightarrow Y$. Conditioning on Z for this graph does not block all back-door paths.

In spite of differences in language and notation, there is little that differentiates the statistics and econometrics models of treatment selection, especially now that the outcome equations used by economists are often completely general nonparametric versions of Equation (4.6) (see Heckman and Vytlacil 2005, which we will discuss in a few different places later in the book, such as Chapters 9 and 12). For now, the key point is that both the statistics and econometric specifications consider the treatment indicator variable, D, to be determined by a set of systematic treatment selection variables

[23] And, as with ignorability, the variables in Z are not necessarily equivalent to those in the minimally sufficient conditioning sets suggested by the back-door criterion. The admissible sets according to the back-door criterion may be subsets of the variables in Z or even variables not in Z that are proximate determinants of Y and that, when conditioned on, block all back-door paths generated by the variables in Z.

[24] The latent variable specification in the econometric tradition can be made equivalent to almost all particular specifications of the statement $\Pr[D=1|S]$ in the statistics tradition by the choice of an explicit probability distribution for U. The full nonparametric equivalence in the causal graph tradition would be $D = f_D(S, e_D) = f_D(Z, U)$.

in S. When all of these variables are observed, the treatment selection mechanism is ignorable and selection is on the observables only. When some of the variables in S are unobserved, the treatment selection mechanism is nonignorable and selection is on the unobservables.

Finally, the qualifications we noted in the prior section about the differences between ignorability assumptions and the back-door criterion apply in analogous fashion to assumptions of selection on the observables. Minimally sufficient conditioning sets suggested by the back-door criterion may be subsets of the variables in Z or entirely different variables that, when conditioned on, eliminate the confounding generated by Z. Variables of the latter type would typically be proximate determinants of Y that intercept the effects of the variables in Z on Y, such as the variable F in Figure 4.4.

4.3.3 Point Identification of Conditional Average Treatment Effects by Conditioning

At the beginning of this chapter, we indicated that we would implicitly focus our presentation of directed graphs and identification issues on the estimation of the unconditional ATE. This narrow focus is entirely consistent with the graphical tradition, in which parameters such as the average treatment effect for the treated (ATT) in Equation (2.7) and the average treatment effect for the controls (ATC) in Equation (2.8) are given considerably less attention than in the potential outcome modeling tradition in both statistics and econometrics. Some comments on the connections may be helpful at this point to foreshadow some of the specific material on causal effect heterogeneity that we will present in the next three chapters.

Identification When the Unconditional ATE Is Identified

If one can identify and consistently estimate the unconditional ATE with conditioning techniques, then one can usually estimate some of the conditional average treatment effects that may be of interest as well. As we will show in the next three chapters, consistent estimates of conditional average treatment effects can usually be formed by specification of alternative weighted averages of the average treatment effects for subgroups defined by values of the conditioning variables. Thus, calculating average effects other than the unconditional ATE may be no more complicated than simply adding one step to the more general conditioning strategy we have presented in this chapter.

Consider again the graph presented in Figure 4.8(b). The back-door path from D to Y is blocked by S. As a result, a consistent estimate of the ATE in Equation (2.3) can be obtained by conditioning on S. But, in addition, consistent estimates of the ATT in Equation (2.7) and the ATC in Equation (2.8) can be obtained by properly weighting conditional differences in the observed values of Y. In particular, first calculate the sample analogs to the differences $E[Y|D=1,S=s]-E[Y|D=0,S=s]$ for all values s of S. Then, weight these differences by the conditional distributions $\Pr[S|D=1]$ and $\Pr[S|D=0]$ to calculate the ATT and the ATC, respectively.

Identification When the Unconditional ATE Is Not Identified

If selection is on the unobservables, conditioning strategies will typically fail to identify unconditional ATEs. Nonetheless, weaker assumptions may still allow for the identification and subsequent estimation by conditioning of various conditional average treatment effects. We will present these specific weaker assumptions in the course of explaining matching and regression techniques in the next three chapters, but for now we give a brief overview of the identification issues in relation to the graphical models presented in this chapter. (See also the discussion in Section 2.7.4 of similar issues with regard to the inconsistency and bias of the naive estimator.)

Suppose, for example, that the graph in Figure 4.9(b) now obtains, and hence that a back-door path from D to Y exists via unobserved determinants of the cause, U. In this case, conditioning on Z will not identify the unconditional ATE. Nonetheless, conditioning on Z may still identify a conditional average treatment effect of interest, as narrower effects can be identified if weaker assumptions can be maintained even though unblocked back-door paths may still exist between D and Y.

Consider a case for which partial ignorability holds, such that $Y^0 \perp\!\!\!\perp D|S$ is true but $(Y^0, Y^1) \perp\!\!\!\perp D \mid S$ is not. Here, conditioning on S generates a consistent estimate of the ATT even though S does not block the back-door path from D to Y. The opposite is, of course, also true. If partial ignorability holds in the other direction, such that $Y^1 \perp\!\!\!\perp D|S$ holds but $(Y^0, Y^1) \perp\!\!\!\perp D|S$ does not, then the ATC can be estimated consistently.[25]

Consider the first case, in which only $Y^0 \perp\!\!\!\perp D \mid S$ holds. Even after conditioning on S, a back-door path remains between D and Y because Y^1 still differs systematically between those in the treatment and control groups and Y is determined in part by Y^1; see Equation (2.2). Nonetheless, if, after conditioning on S, the outcome under the no-treatment-state, Y^0, is independent of exposure to the treatment, then the ATT can be estimated consistently. The average values of Y, conditional on S, can be used to consistently estimate the average what-if values for the treated if they were instead in the control state. This type of partial ignorability is akin to Assumption 2 in Equation (2.16), except that it is conditional on S. We will give a full explanation of the utility of such assumptions when discussing matching estimates of the ATT and the ATC in the next chapter.

Graphs Do Not Clearly Reveal the Identification Possibilities for the ATT and ATC When the ATE Is Not Also Identified

In the prior section, we noted in our presentations of ignorability and selection on the observables that graphs help guide researchers toward minimally sufficient conditioning sets that may differ from the conditioning sets suggested by the statistics and econometrics literature. The back-door criterion is especially helpful in this regard because of its targeted focus on back-door paths that generate noncausal associations between the causal variable and the outcome variable. However, it must also be stated, as implied by this section, that directed graphs will not clearly reveal effective analysis

[25] And, as we will show in the next chapter, the required assumptions are even simpler because the entire distributions of Y^0 and Y^1 need not be conditionally independent of D. As long as the stable unit treatment value assumption (SUTVA) holds, only mean independence must be maintained.

strategies to identify either the ATT or the ATC in situations where the ATE cannot also be identified by conditioning.

The overall implication of this point is that researchers should learn all three frameworks for approaching causal identification challenges. Graphs help immensely in selecting conditioning sets when the target parameter is the ATE. When the ATE is not identified by any feasible conditioning sets, the potential outcome model – either as deployed in statistics or econometrics – can still guide the researcher to identification strategies for narrower conditional average effects, most commonly the ATT or the ATC. Directed graphs remain a useful tool in these situations, as they help to organize one's thinking about the full system of causal relationships that are relevant. But, the identification strategy that is then selected to estimate either the ATT or the ATC is likely to emerge from thinking through the possibilities from within the potential outcome model.

4.4 Conditioning to Balance and Conditioning to Adjust

When presenting Pearl's back-door criterion for determining a sufficient set of conditioning variables, we noted that for some applications more than one set of conditioning variables is sufficient. In this section, we return to this point as a bridge to the following three chapters that present both matching and regression implementations of conditioning. Although we will show that matching and regression can be considered variants of each other, here we point to the different ways in which they are usually invoked in applied research. Matching is most often considered a technique to balance the determinants of the causal variable, and regression is most often considered a technique to adjust for other causes of the outcome.

To frame this discussion, consider first the origins of the balancing approach in the randomized experiment tradition. Here, the most familiar approach is a randomized experiment that ensures that treatment status is unassociated with all observed and unobserved variables that determine the outcome (although only in expectation). When treatment status is unassociated with an observed set of variables W, the data are balanced with respect to W. More formally, the data are balanced if

$$\Pr[W|D=1]=\Pr[W|D=0], \tag{4.8}$$

which requires that the probability distribution of W be the same within the treatment and control groups.

Now consider the graph presented in Figure 4.10. Back-door paths are present from D to Y, represented by $D \leftarrow S \dashleftarrow\dashrightarrow X \rightarrow Y$, where S is the complete set of variables that are direct causes of treatment assignment/selection, X is the complete set of variables other than D that are direct causes of Y, and the bidirected edge between S and X signifies that they are mutually caused by some set of common unobserved causes.[26]

[26] For this example, we could have motivated the same set of conclusions with other types of causal graphs. The same basic conclusions would hold even if X and S include several variables within them in which some members of X cause D directly and some members of S cause Y directly. In other

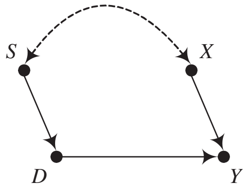


Figure 4.10 A causal diagram in which sufficient conditioning can be performed with respect to S or X.

Because neither S nor X is a collider, all back-door paths in the graph can be blocked by conditioning on either S or X (and we write "paths" because there may be many back-door paths through the bidirected edge between S and X). Conditioning on S is considered a balancing conditioning strategy, whereas conditioning on X is considered an adjustment-for-other-causes conditioning strategy. If one observes and then conditions on S, the variables in S and D are no longer associated within the subgroups defined by the conditioning. The treatment and control groups are thereby balanced with respect to the distribution of S. Alternatively, if one conditions on X, the resulting subgroup differences in Y across D within X can be attributed to D alone. In this case, the goal is not to balance X but rather to partial out its effects on Y in order to isolate the net effect of D on Y.

The distinction between balancing and adjustment for other causes is somewhat artificial (see Hansen 2008). For the graph in Figure 4.10, balancing X identifies the causal effect. Thus, it is technically valid to say that one can identify a causal effect by balancing a sufficient set of other causes of Y. Nonetheless, the graph in Figure 4.10 demonstrates why the distinction is important. The ultimate set of systematic causes that generates the relationship between S and X is unobserved, as it often is in many applied research situations. Because one cannot condition on these unobserved variables, one must condition on either S or X in order to identify the causal effect. These two alternatives may be quite different in their practical implementation.[27]

Should one balance the determinants of a cause, or should one adjust for other causes of the outcome? The answer to this question is situation specific, and it depends on the quality of our knowledge and measurement of the determinants of D and Y.

words, all that we need to make the distinction between balancing and adjustment for other direct causes is two sets of variables that are related to each other, with at least one variable in one set that causes D but not Y and at least one variable in the other set that causes Y but not D.

[27] Note also that the ingredients utilized to estimate the ATE (as well as the ATT and the ATC) will differ based on the particular conditioning routine, and this will allow alternative expressions of underlying heterogeneity. If S is observed, then conditional average treatment effects can be calculated for those who are subject to the cause for different reasons, based on the values of S that determine D. If X is observed, then conditional average treatment effects can be calculated holding other causes of Y at chosen values of X. Each of these sets of conditional average treatment effects has its own appeal, with the relative appeal of each depending on the application. In the potential outcome tradition, average treatment effects conditional on S would likely be of more interest than average treatment effects conditional on X. But for those who are accustomed to working within an all-cause regression tradition, then average treatments effects conditional on X might be more appealing.

One answer is that the researcher should do both.[28] Nonetheless, there is a specific advantage of balancing that may tip the scales in its favor if both strategies are feasible: Balancing diminishes the inferential problems that can be induced by data-driven specification searches, as we will explain in Chapter 6.[29]

4.5 Conclusions

In this chapter, we have used the causal graphs introduced in Chapter 3 to explain the rationale for conditioning estimators of causal effects, focusing on the back-door criterion for selecting sufficient sets of conditioning variables that identify the ATE. We then introduced models of treatment assignment and treatment selection from statistics and econometrics that are used to assert similar claims about conditioning estimators, focusing also on the ATT and ATC.

In the next three chapters, we present details and connections between the three main types of conditioning estimation strategies utilized in the social sciences: matching, regression, and weighted regression. We show how they typically succeed when selection is on the observables and fail when selection is on the unobservables. We lay out the specific assumptions that allow for the identification of unconditional average treatment effects, as well as the weaker assumptions that allow for the identification of narrower conditional average treatment effects, such as the ATT and ATC. In later chapters, we then present additional methods for identifying and estimating causal effects when conditioning methods do not suffice because crucial variables on back-door paths are unobserved.

4.6 Appendix to Chapter 4: The Back-Door and Adjustment Criteria, Descendants, and Colliders Under Magnification

In order to properly utilize the back-door criterion when evaluating alternative conditioning sets, an analyst does not need to understand all details of all scenarios in which its violation will prevent a conditioning estimator from generating consistent and unbiased estimates of causal effects of interest. However, a deeper examination of additional scenarios provides insight into common methodological challenges, while also demonstrating the powerful contribution that graphical methods are likely to make to social science research in the coming decades.

In this appendix, we offer additional examples where colliders and descendants must be carefully considered in order to understand the fine points of why the back-door criterion warrants causal inference. We show how Condition 2 of the back-door

[28] As we discuss in later chapters, many scholars have argued for conditioning on both S and X. Robins, for example, argues for this option as a double protection strategy that offers two chances to effectively block the back-door paths between D and Y (see Robins and Rotnitzky 2001).

[29] We phrase this guidance with "may" because it must be evaluated alongside another concern. Balancing all variables that determine D, including those that do not generate confounding, is unnecessary and inefficient. As a result, it may render conditioning infeasible in datasets of the size typically available to social scientists.

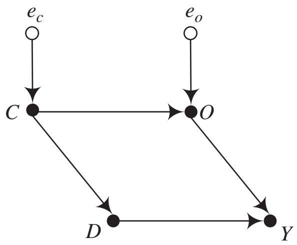


Figure 4.11 A causal graph with a confounded causal effect and where the variables along the back-door path are viewed under magnification.

criterion, which appears only to prevent the analyst from inadvertently conditioning on variables that adjust away the causal effect of interest, also plays a role in blocking hidden as-if back-door paths that also result from conditioning on endogenous variables. We show that this result can only be seen when endogenous variables are viewed under magnification, so that it becomes clear that these variables are also colliders. We conclude with an examination of the connections between Pearl's original back-door criterion, its generalization as the adjustment criterion of Shpitser, VanderWeele, and Robins (2010), and our blended alternative, which retains the targeted spirit and inherent practicality of Pearl's original back-door criterion while also incorporating some of the insight furnished by the admirably broad adjustment criterion.

Like the appendix to Chapter 3, nothing in this appendix is necessary for understanding the next three chapters that follow. We offer it for curious readers who crave a deeper understanding of the back-door criterion and who wish to dive into the original literature after reading this book. We also offer it to explain to the community of causal graph methodologists why we have chosen the version of the back-door criterion that we have.

The Back-Door Criterion Explained with Causal Graphs Viewed Under Selective Magnification. The first step in fully considering why violations of the back-door criterion prevent conditioning estimators from delivering consistent and unbiased estimates is to repeat some of the material from the main body of the chapter, after redrawing the causal graphs under magnification. As we introduced in Section 3.3.2 using Figure 3.7, under magnification the unobserved structural error terms of the associated structural equations are brought into view. For Figure 4.11, we have redrawn Figure 4.1 under what we will call "selective magnification." In this case, we magnify the nodes of all variables that we might consider conditioning on, leaving the error terms on the causal variable and the outcome variable hidden as in the standard representation for visual simplicity. Accordingly, Figure 4.11 shows two additional causal effects, $e_C \rightarrow C$ and $e_O \rightarrow O$, that were only implicit in Figure 4.1.

Now, reconsider our explanation of the back-door criterion for this graph. We wrote earlier that conditioning on C and/or O would identify the causal effect. As a transition to the additional examples in this appendix, we will now present the same explanation considering how the inclusion of $e_C \rightarrow C$ and $e_O \rightarrow O$ in the graph does not change this claim.

In the directed graph tradition, C is a "root" variable because it has no causes other than those embedded in its structural error term. Root variables do not need to be considered in any other way when viewed under magnification because their structural error terms are nothing more than the source of the variation observed for these variables. Accordingly, it is still the case that conditioning on C blocks the sole back-door path $D \leftarrow C \rightarrow O \rightarrow Y$ because C is the middle variable of a fork of mutual dependence that lies within the path. Because C is also not a descendant of D, it satisfies the back-door criterion.

For O, an additional complication should be clear. Under magnification, it is revealed that O is a collider variable because $C \rightarrow O$ is now accompanied by $e_O \rightarrow O$. In general, all variables in a causal graph that are not root variables will appear as colliders when viewed under magnification. We have urged caution when handling colliders, and yet only now do we reveal that all variables such as O are, in fact, colliders as well. Accordingly, conditioning on O will induce an association between C and e_O. Fortunately, this conditioning does not invalidate the back-door criterion by unblocking an already blocked back-door path, as we will now explain.

Recall our prior discussions of colliders in Sections 3.2.2 and 4.1.2. We have said that for a blocked path, $A \rightarrow C \leftarrow B$, where C is the collider, conditioning on C opens up this path. We now introduce a piece of notation to convey this result: a conditioning "button," $\odot$, that can be deployed to show the genesis of a new induced association, $\cdots \odot \cdots$.[30] For example, for the blocked path $A \rightarrow C \leftarrow B$, conditioning on C generates a new association, $A \cdots \odot \cdots B$, where the conditioning action itself, $\odot$, generates the association between A and B. For our hypothetical college admissions example in 4.1.2, the notation represents the following action: A button is pushed for "admissions decisions," and the population of applicants is then divided into those admitted and those rejected. Within both groups, A and B are now associated, as shown in Figure 4.2. The only way to eliminate the induced association is to undo the conditioning that generates it (i.e., release the conditioning button).

Return to Figure 4.11 so that we can demonstrate this new notation when considering the conclusions suggested by an evaluation of the back-door criterion. Suppose again that O is our candidate conditioning variable, and we know that conditioning on O will block the back-door path $D \leftarrow C \rightarrow O \rightarrow Y$. When seen under magnification, it should be clear that now conditioning on O also generates a new association, $C \cdots \odot \cdots e_O$. In fact, conditioning on O also generates a second association, $e_C \cdots \odot \cdots e_O$.[31] Pearl (2009:339) explains, for examples such as this one, that these induced associations create two as-if back-door paths:

1. $D \leftarrow C \cdots \odot \cdots e_O \rightarrow O \rightarrow Y$ and

2. $D \leftarrow C \leftarrow e_C \cdots \odot \cdots e_O \rightarrow O \rightarrow Y$.

[30] Pearl and other authors often use alternative notation to convey the same associations, most commonly A—B. We prefer our more active "button," which can also be directly extended to actual operator status, as in $\cdots \odot (.) \cdots$. If one wanted to show conditioning on more than one variable, we would have $\cdots \odot (C,O) \cdots$, etc.

[31] Conditioning on a collider, or a descendant of a collider, induces associations between all ancestors of the collider on separate directed paths that reach the collider.

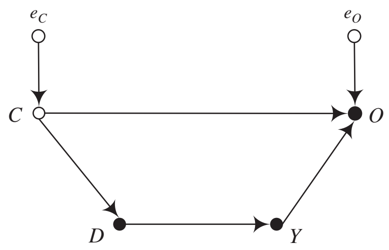


Figure 4.12 A diagram where the causal effect of D on Y is not confounded and where the observed variable O on the back-door path is a descendant of both D and Y.

Fortunately, both of these as-if back-door paths are also blocked when O is conditioned on because O is the middle variable in a chain of mediation, $e_O \rightarrow O \rightarrow Y$, for both of them. Thus, O satisfies the back-door criterion because O is not a descendant of D and because conditioning on O eliminates all back-door associations between D and Y, including two associations created by as-if back-door paths that are only revealed under magnification.

Of course, we already knew that O satisfied the back-door criterion from our consideration of Figure 4.1. Viewing the graph under selective magnification has forced us to reckon with the unobserved structural error terms, and little insight has been gained. We will now consider two additional examples where consideration of the structural error terms leads to additional insight that will convince the reader of how simple and effective the back-door criterion is.

Consider Figure 4.12, where again we have four variables, and where all variables but D and Y are displayed under magnification. The only back-door path between D and Y, which is $D \leftarrow C \rightarrow O \leftarrow Y$, is blocked by the collider O. Although it may seem awkward to refer to $D \leftarrow C \rightarrow O \leftarrow Y$ as a back-door path between D and Y (because it does not end with $\rightarrow Y$), it satisfies our definition because it is a path between two causally ordered variables D and Y that begins with $D \leftarrow$. As a result, it is still a back-door path, even though it does not terminate with $\rightarrow Y$ (unlike all other back-door paths displayed so far in our examples).[32] Because the sole back-door path between D and Y does not generate a noncausal association between D and Y, there is no need to adjust for any variables in order to generate a consistent and unbiased estimate of the causal effect of D on Y.

In fact, all that one can do is make the situation worse. Suppose that an analyst believes mistakenly that the path $D \leftarrow C \rightarrow O \leftarrow Y$ is an unblocked back-door path that must be blocked in order to generate a consistent and unbiased estimate of the effect of D on Y. Because C is unobserved, suppose that the analyst decides to condition on O instead, under the rationale that it is the only observed variable that lies on the back-door path.

[32] We could bring this graph in line with prior ones by replacing $Y \rightarrow O$ with $O \leftarrow U \rightarrow Y$. The explanation would then change a bit because O would no longer be a descendant of D via a directed path, but the analysis with respect to Condition 1 of the back-door criterion would be the same.

The variable O violates both conditions of the back-door criterion. Consider Condition 1 first. Because O is a collider variable, conditioning on O generates many new associations, including $C \cdots \odot \cdots Y$, $e_C \cdots \odot \cdots Y$, $C \cdots \odot \cdots e_O$, and $e_O \cdots \odot \cdots Y$. These induced associations generate three as-if back-door paths:

1. $D \leftarrow C \cdots \odot \cdots Y$,
2. $D \leftarrow C \leftarrow e_C \cdots \odot \cdots Y$, and
3. $D \leftarrow C \cdots \odot \cdots e_O \cdots \odot \cdots Y$.

Because C is unobserved, these as-if back-door paths remain unblocked after conditioning on O, and therefore O does not satisfy Condition 1 of the back-door criterion.

Now consider Condition 2, and notice that O lies on a directed path, $D \rightarrow Y \rightarrow O$, that begins at D and reaches Y. As we noted earlier, we chose to specify Condition 2 of the back-door criterion with the words "reaches the outcome variable" rather than "ends at at the outcome variable" in order to capture cases such as this one. In this case, O lies on the directed path that represents the causal effect of interest, even though O does not mediate the casual effect itself. The intuition of Condition 2 of the back-door criterion should nonetheless still be clear. Because the causal effect of D on Y is fully embedded within the variation in O, adjusting for O would explain away the causal effect itself.[33]

Overall, then, conditioning on O in this graph would make the situation considerably worse. No unblocked back-door paths needed to be blocked in the first place, and conditioning on O would generate as-if back-door paths that induce new noncausal associations between D and Y while at the same time robbing the causal effect of (possibly all) of its magnitude.

To now transition to a more complicated case, pause to consider how our language in this appendix differs from the language used in the main body of the chapter. Up until now, we have expressed similar results to those for Figure 4.12 on the perils of conditioning on colliders using more brief language. For this graph, the simpler language would be the following:

> Conditioning on the collider variable O unblocks the already blocked back-door path, $D \leftarrow C \rightarrow O \leftarrow Y$.

This explanatory syntax is concise and correct. Even so, it should now be clear that a more laborious way to understand this result is the following:

> Conditioning on the variable O that is a collider on the back-door path $D \leftarrow C \rightarrow O \leftarrow Y$ creates additional as-if back-door paths such as $D \leftarrow C \cdots \odot \cdots Y$, $D \leftarrow C \leftarrow e_C \cdots \odot \cdots Y$, and $D \leftarrow C \cdots \odot \cdots e_O \cdots \odot \cdots Y$. These as-if back-door paths are unblocked when conditioning on O.

[33] In an infinite sample, wherein one could stratify the data on all values of O, D and Y would be independent within the strata of O. In a finite sample, possibly necessitating the use of a different type of conditioning estimator, the conditional association between D and Y might not equal zero but would still not be a consistent estimate of the causal effect.

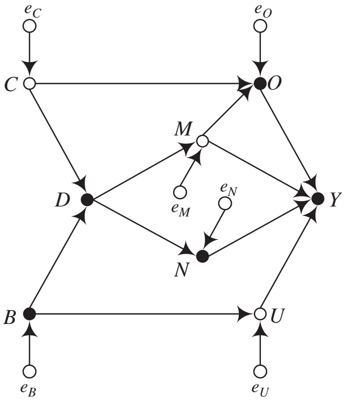


Figure 4.13 A graph where the effect of D on Y is not identified by conditioning on O and B because O is a descendant of D.

For the graphs considered in the main body of this chapter, this more tortured language would not have been helpful and therefore was not used. However, for this particular graph, the more tortured language may offer a clearer explanation. The brief explanation does not make clear why conditioning on O generates a back-door association by unblocking $D \leftarrow C \rightarrow O \leftarrow Y$, given that this back-door path does not end with $\rightarrow Y$. The more tortured language, which uses the conditioning button, does provide the proper imagery because the as-if back-door path, $D \leftarrow C \cdots \odot \cdots Y$, does not end with $\leftarrow Y$.

For an additional example, which draws many of these issues together, consider Figure 4.13. This graph is a redrawn version Figure 4.7, but now with all variables other than D and Y displayed under magnification. As noted earlier for Figure 4.7, there are three back-door paths from D to Y:

1. $D \leftarrow C \rightarrow O \rightarrow Y$,
2. $D \leftarrow C \rightarrow O \leftarrow M \rightarrow Y$, and
3. $D \leftarrow B \rightarrow U \rightarrow Y$.

Again, suppose that one decides to condition on both O and B. Using the conditioning button to reveal all of the associations generated by the conditioning action itself would generate many new as-if back-door paths. Consider just two of these. The induced association $C \cdots \odot \cdots M$ creates the as-if back-door path $D \leftarrow C \cdots \odot \cdots M \rightarrow Y$ that remains unblocked after conditioning on O and B. This as-if back-door path is similar to those considered for Figure 4.12. But now consider how the induced association $D \cdots \odot \cdots e_M$ creates the as-if back-door path $D \cdots \odot \cdots e_M \rightarrow M \rightarrow Y$. This path is also unblocked after conditioning on O and B. But, most importantly, this as-if back-door path only comes to the foreground when the causal graph is viewed under magnification. Except under magnification, it is all too easy to fail to recognize that M is a collider and that conditioning on a descendant of M will induce an association between D and e_M. For all but the most experienced causal graph practitioners, this as-if back-door path would be hidden by the standard representation of the graph.

As a result, the conditioning set of O and B does not satisfy Conditions 1(a), (b), and (c) of the back-door criterion. As we used this graph to show in the main body of the chapter, this graph also does not satisfy Condition 2 of the back-door criterion either. This result does not need to be explained in any different fashion when the graph is viewed under selective magnification. Overall, the adjusted effect of D on Y, which results from conditioning on O and B, would steal some of the total effect of D on Y. Yet, as we show here, the conditioning action also generates new back-door associations, one through the unobserved mediating variable M. This last as-if back-door path is completely hidden from view in the standard representation of a causal graph. Fortunately, all such hidden as-if back-door paths that emerge under conditioning will never creep into an empirical analysis as long as the back-door criterion is properly evaluated.

And this is the key point of this last example: Condition 2 of the back-door criterion must be heeded for two reasons. First, as was also presented in the main body of the chapter, conditioning on descendants of the cause that lie on (or descend from) directed paths that begin at D and that reach Y will always mistakenly adjust away some of the causal effect the analyst is interested in estimating. Second, as we have now shown in this appendix, conditioning on the descendants of the cause (such as O) that are also descendants of unobserved colliders that are are themselves descendants of the cause (such as M) will always generate unblocked back-door associations that spoil the analysis. Fortunately, if one uses to the back-door criterion to select conditioning sets, then the threat of such hidden unblocked back-door associations can be avoided.

The Back-Door and Adjustment Criteria Considered. As we noted earlier, our version of the back-door criterion incorporates insight gained from the more recent development of the *adjustment criterion* by Shpitser et al. (2010). Figure 4.14 presents a graph that reveals the differences between Pearl's original back-door criterion and the more recent generalization that is the adjustment criterion. We will use this graph to explain why we have modified Pearl's original back-door criterion to a small degree but also why we have not adopted the adjustment criterion as a whole.

As we have specified it in the main body of this chapter, the back-door criterion is evaluated in the following two steps. The first step is to write down the back-door paths from D to Y, determine which ones are unblocked, and then search for a conditioning set that will block all unblocked back-door paths. For Figure 4.14, the two back-door paths are

1. $D \leftarrow A \rightarrow B \leftarrow C \rightarrow Y$ and

2. $D \leftarrow E \rightarrow Y$.

Path 1 is already blocked by the collider B, and so only path 2 must be blocked by conditioning. Here, the solution is simple: Condition on E because it is the middle variable of a fork of mutual dependence.

The second step is to verify that the variables in the candidate conditioning set are not descendants of the causal variable that lie along, or descend from, all directed paths from the causal variable that reach the outcome variable. For Figure 4.14, it

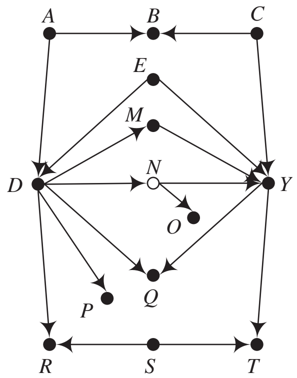


Figure 4.14 A directed graph that reveals the differences between the back-door criterion and the adjustment criterion.

is easy to see that E is a root variable and as such is not a descendant of any other variable in the graph.

However, to understand the differences between the back-door criterion in its original form and the adjustment criterion, we need to consider this step in more detail. For more elaborate causal graphs, in order to evaluate Condition 2 of the back-door criterion one needs to ascend the family tree by looking "upstream" through the arrows that point to each candidate conditioning variable (i.e., from the parent, to the grandparent, to the great grandparent, and so on) to determine whether the causal variable is a direct ancestor of any of the candidate conditioning variables. If the causal variable is not found to be an ancestor of any of the candidate conditioning variables in the set under consideration, then by definition none of the candidate conditioning variables lies on or descends from directed paths that begin at the causal variable and that reach the outcome variable. In this case, a conditioning estimator that uses the variables under consideration will generate consistent and unbiased estimates of the causal effect of interest.

If the causal variable is discovered to be an ancestor of any of the candidate conditioning variables in the set under consideration, then for our version of the back-door criterion (which incorporates insight from the adjustment criterion), one has to determine whether the directed paths from the causal variable that establish the descent of the conditioning variable(s) also reach the outcome variable. If not, then one can still condition on any such variables and generate consistent and unbiased estimates. If, instead, any of the candidate conditioning variables are descendants of the causal variable *and* lie on or descend from directed paths that begin at the causal variable and reach the outcome variable, then the conditioning set will block or adjust away some of the causal effect. In this case, the conditioning set will not generate consistent and unbiased estimates of the causal effect of interest.

To now transition to a parallel consideration of the adjustment criterion, and why we have not adopted it in whole, return to the case of Figure 4.14 and consider how the back-door criterion might be utilized in practice by an experienced analyst. As we noted above in this appendix, conditioning on E alone satisfies the back-door criterion. Notice, however, that, in addition to E, one could also condition on A, on C, on A and B, on B and C, or on A, B, and C. The back-door criterion does not rule out these possibilities, but it does not encourage the analyst to consider A and C as members of the candidate conditioning set because the back-door path on which these variables lie is already blocked in the absence of any conditioning by B. Instead, the back-door criterion guides the analyst to the essential information: (1) path 1 is already blocked by B and can be left alone; (2) path 2 must be blocked by conditioning on E.

Nonetheless, an experienced analyst might reason that, although one should not condition only on B, one could condition on B as long as either A or C is also conditioned on. The experienced analyst might then choose to offer two estimates, one that conditions only on E and one that conditions on B and C as well (perhaps because a fair critic has a theory says that the $B \leftarrow C$ in Figure 4.14 should instead be $B \rightarrow C$). The experienced analyst might reach this position even though she regards conditioning on B and C as unnecessary and inefficient, given that she truly does believe that her theoretical assumptions represented by the original graph are beyond reproach. Nevertheless, she might reason that it is prudent to show a fair critic that even if his alternative assumptions are correct, the main results of the ensuing empirical analysis remain the same.

Once one begins to allow supplementary but unnecessary conditioning variables into the conditioning set, a question arises as to whether the back-door criterion covers all cases. This is the inspiration for the adjustment criterion, developed by Shpitser et al. (2010). It is designed to allow the analyst to identify all permissible conditioning sets, including those that include unnecessary and redundant conditioning, as we now explain.

The adjustment criterion also has two steps. For the first step, the analyst considers all paths in the graph that begin at the causal variable and then categorizes all paths into (1) "causal paths," which are defined as all directed paths from the causal variable to the outcome variable, and (2) "noncausal paths," which are defined as all paths from the causal variable that are not causal paths. For Figure 4.14, one would enumerate the following paths, categorizing them as follows:

1. $D \leftarrow A \rightarrow B \leftarrow C \rightarrow D$ (noncausal),
2. $D \leftarrow E \rightarrow Y$ (noncausal),
3. $D \rightarrow M \rightarrow Y$ (causal),
4. $D \rightarrow N \rightarrow Y$ (causal),
5. $D \rightarrow N \rightarrow O$ (noncausal),
6. $D \rightarrow P$ (noncausal),
7. $D \rightarrow Q \leftarrow Y$ (noncausal),
8. $D \rightarrow R \leftarrow S \rightarrow T \leftarrow Y$ (noncausal).

Notice that paths 1 and 2 are back-door paths for the back-door criterion but, for the adjustment criterion, are classified as members of the more encompassing set of noncausal paths. The first step of the adjustment criterion then requires that the analyst search for variables that, when conditioned on, will ensure that all noncausal paths between D and Y are blocked. This is analogous to the back-door criterion, but now all noncausal paths, not just back-door paths, must be considered. For Figure 4.14, this part of the first step requires the analyst to search for all permissible conditioning sets that will block the noncausal paths 1, 2, 5, 6, 7, and 8. For the case of Figure 4.14, more sets than we wish to enumerate would be selected. In brief, all sets would include E and exclude Q. Yet, these sets would include many combinations of additional unnecessary conditioning variables, including appropriate combinations of A, B, C, N, O, P, R, S, and T, such as the maximal blocking set $\{A,B,C,E,N,O,P,R,S,T\}$.

For the second step of the adjustment criterion, one then implements the same second step as the version of the back-door criterion we have specified in the main body of the chapter. For Figure 4.14, this step requires that the candidate conditioning sets not include the variables M, N, or O. Presumably M would not have been considered as a conditioning variable in the first place because it is not a variable that lies on any of the noncausal paths that the analyst is attempting to block. Nonetheless, the analyst must also strike all conditioning sets that include N and O, and this conclusion has to be discerned from the graph, not the list of the paths that does not reveal full patterns of descent. (This is also the case for the back-door criterion.)

The advantage of the adjustment criteria is that it will reveal all permissible conditioning sets, such as the maximal permissible conditioning set $\{A,B,C,E,P,R,S,T\}$, assuming that the analyst devotes the necessary energy to delineating all patterns of permissible conditioning. The disadvantage is that, for graphs such as the one in Figure 4.14, the adjustment criterion is laborious and is likely to lead one to condition on more variables than is necessary to identify the causal effect.

We have therefore chosen to stick with the back-door criterion in the main body of the chapter, with only one substantial modification to Pearl's original specification in deference to the completeness of the adjustment criterion. As we noted earlier in our footnote on Condition 2 of our version of the back-door criterion, Pearl's original back-door criterion requires more simply (but overly strongly) that no variables in the conditioning set Z can be descendants of the causal variable. For our version of the back-door criterion, we weaken Pearl's original no-descendants condition to allow conditioning on variables, such as P in Figure 4.14, that are descendants of the cause D but that do not lie on or descend from directed paths that begin at D and reach Y. We doubt an analyst would think to condition on P when evaluating the back-door criterion for this graph (because the back-door criterion focuses attention more narrowly on back-door paths, rather than all noncausal paths). Still, this modification of the back-door criterion allows it to identify more of the permissible conditioning sets that are identified by the more laborious but admirably complete adjustment criterion.