Chapter 3

Causal Graphs, Identification, and Models of Causal Exposure

In this chapter, we present the basic conditioning strategy for the identification and estimation of causal effects. After introducing a methodology for building causal graphs, we present what has become known as the back-door criterion for sufficient conditioning to identify a causal effect. We then present models of causal exposure, introducing the treatment assignment and treatment selection literature from statistics and econometrics. We then return to the back-door criterion and discuss the two basic motivations of conditioning – balancing determinants of the cause of interest and adjusting for other causes of the outcome. We conclude with a discussion of the identification and estimation of conditional average causal effects by conditioning.

3.1 Causal Graphs and Conditioning as Back-Door Identification

In his 2000 book titled *Causality: Models, Reasoning, and Inference*, Judea Pearl lays out a powerful and extensive graphical theory of causality. Here, we present and use only the most basic elements of his theory. To the reader familiar with traditional linear path models, much of this material will look familiar. There are, however, important and subtle differences between traditional path models and Pearl's usage of directed acyclic graphs (DAGs).

Pearl's work provides a language and a framework for thinking about causality that differs from the potential outcome perspective presented in the last chapter. Beyond the alternative terminology and notation, Pearl (2000, Section 7.3) proves that the fundamental concepts underlying the potential outcome model and his more recent perspective are equivalent. In some cases, causal

statements in the potential outcome framework can be represented concisely by a causal graph. But it can be awkward to represent many of the complications created by causal effect heterogeneity. Accordingly, in this section we suppress potential outcome random variables and use only observed outcome variables.[1] Furthermore, we implicitly focus on only the unconditional average treatment effect first, although we will return to a discussion of the estimation of conditional average treatment effects in the final section of the chapter.

Even though we must suppress some of the very useful generality of the potential outcome framework, Pearl has shown that graphs nonetheless provide a direct and powerful way of thinking about causal systems of variables and the identification strategies that can be used to estimate the effects within them. Thus, some of the advantage of the framework is precisely that it permits suppression of what could be a dizzying amount of notation to reference each potential causal state in a system of equations. In this sense, Pearl's perspective is a reaffirmation of the utility of graphical models in general, and its appeal to us is similar to the general appeal of path models, which have retained their adherents in spite of some of their known limitations.

For our purposes in this chapter, Pearl's work is important for three different reasons. First, his framework is completely nonparametric, and as a result it is usually unnecessary to specify the nature of the functional dependence of an outcome Y on the variables that cause it. Thus, $X \rightarrow Y$ simply implies that X causes Y, without specifying whether the effect is linear, quadratic, or some other highly nonlinear function in the values of X. This generality allows for a theory of causality without side assumptions about functional form, such as the linearity assumptions that became the Achilles' heel of traditional path models. Second, Pearl's approach shows clearly the critical importance of what he labels "collider" variables, which are specific types of endogenous variables that must be treated with caution. Finally, Pearl shows that there are three basic methods for identifying a causal effect: conditioning on variables that block all back-door paths, conditioning on variables that allow for estimation by a mechanism, and estimating a causal effect by an instrumental variable that is an exogenous shock to to the cause. Each of these identification strategies was already introduced briefly in Section 1.6. Here, we provide the foundations of his approach and then offer a more detailed presentation of the conditioning strategy for estimating causal effects by invoking Pearl's back-door criterion. In Chapter 6, we then return to the framework to discuss the estimation of causal effects more generally when important variables are unobserved. There, we will then more fully present the front-door and instrumental variable approaches.

3.1.1 Basic Elements

As with standard path models, the basic goal of drawing a causal system as a DAG is to represent explicitly all causes of the outcome of interest. As we

[1] Nonetheless, we still maintain that the observed outcome variable is generated by a process of causal exposure to alternative causal states with their attendant potential outcome variables. See Pearl (2000, Chapter 7) for the formal connections between potential outcomes and causal graphs.

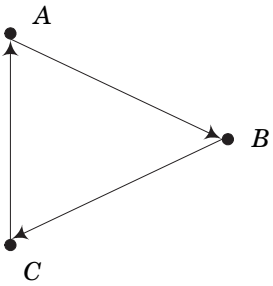


Figure 3.1: A directed graph that includes a cycle.

discussed earlier in Section 1.6, each node of a causal graph represents a random variable and is labeled by a letter such as A, B, or C. Nodes that are represented by a solid circle • are observed random variables, whereas nodes that are represented by a hollow circle ∘ are unobserved random variables. Causes are represented by directed edges → (i.e., single-headed arrows), such that an edge from one node to another signifies that the variable at the origin of the directed edge causes the variable at the terminus.[2]

Unlike standard path models, a DAG does not permit a representation of simultaneous causation. Only directed edges are permissible, and thus direct causation can run in only one direction, as in $X \rightarrow Y$. Furthermore, a DAG is defined to be an acyclic graph. Accordingly, no directed paths emanating from a causal variable also terminate at the same causal variable. Figure 3.1 presents a graph that includes a cycle, and as a result it is not a DAG, even though it includes only directed edges.

Under some circumstances it is useful to use a curved and dashed bidirected edge (as in Figures 1.1–1.3 earlier) as a shorthand device to indicate that two variables are mutually dependent on one or more (typically unobserved) common causes. In this shorthand, the two graphs presented in panels (a) and (b) of Figure 3.2 are equivalent. When this shorthand representation is used, the resulting graph is no longer a DAG by its formal definition. But, because the bidirected edge is a mere shorthand semantic substitution, the graph can be treated usually as if it were a DAG.[3] Such shorthand can be helpful in suppressing a complex set of background causal relationships that are irrelevant to the empirical analysis at hand. Nonetheless, these bidirected edges should not be interpreted in any way other than as we have just stated. They are not

[2]In Pearl's framework, each random variable is assumed to have an implicit probability distribution net of the causal effects represented by the directed edges. This position is equivalent to assuming that background causes of each variable exist that are independent of the causes explicitly represented in the graph by directed edges.

[3]Pearl would refer to such a graph as a semi-Markovian causal diagram rather than a fully Markovian causal model, but he would nonetheless treat it as if it were a full DAG when considering the identification of causal effects that are represented by directed edges in the graph (see Pearl 2000, Section 5.2).

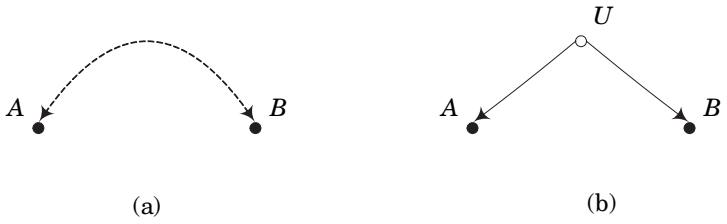


Figure 3.2: Two representations of the joint dependence of A and B on common causes.

indicators of mere correlations between the variables that they connect, and they do not signify that either of the two variables has a direct cause on the other one. Rather, they represent an unspecified set of common causes of the two variables that they connect.

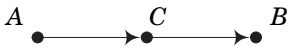
Figure 3.3 presents the three basic patterns of causal relationships that would be observed for any three variables that are related to each other: a chain of mediation, a fork of mutual dependence, and an inverted fork of mutual causation. Pearl's analysis of the first two types of relationship is conventional. For the graph in panel (a), A affects B through A's causal effect on C and C's causal effect on B. This type of a causal chain renders the variables A and B unconditionally associated. For the graph in panel (b), A and B are both caused by C. Here, A and B are also unconditionally associated, but now it is because they mutually depend on C.[4]

For the third graph in panel (c), A and B are again connected by a pathway through C. But now A and B are both causes of C. Pearl labels C a "collider" variable. Formally, a variable is a collider along a particular path if it has two arrows running into it. Figuratively, the causal effects of A and B "collide" with each other at C. Collider variables are common in social science applications: Any endogenous variable that has two or more causes is a collider along some path.

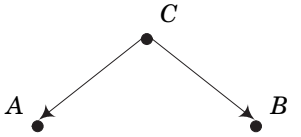
A path that is connected by a collider variable does not generate an unconditional association between the variables that cause the collider variable. For the mutual causation graph in panel (c) of Figure 3.3, the pathway between A and B through C does not generate an unconditional association between A and B. As a result, if nothing is known about the value that C takes on, then knowing the value that A takes on yields no information about the value that B takes on. Pearl's language is quite helpful here. The path $A \rightarrow C \leftarrow B$ does not generate an association between A and B because the collider variable C "blocks" the possible causal effects of A and B on each other.

Even though collider variables do not generate unconditional associations between the variables that determine them, we will show in the next subsection

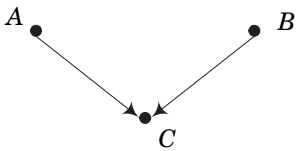
[4]The unconditional associations between A and B for both graphs mean that knowing the value that A takes on gives one some information on the likely value that B takes on. This unconditional association between A and B, however, is completely indirect, as neither A nor B has a direct causal effect on each other.



(a) Mediation



(b) Mutual dependence



(c) Mutual causation

Figure 3.3: Basic patterns of causal relationships for three variables.

that incautious handling of colliders can create conditional dependence that can sabotage a causal analysis. The importance of considering collider variables is a key insight of Pearl's framework, and it is closely related to the familiar concerns of selecting on the dependent variable and conditioning on an endogenous variable. But, to understand these complications, we first must introduce basic conditioning techniques in the context of graphical models.

3.1.2 Conditioning on Observable Variables

One of the most common modeling strategies to prosecute causal questions in quantitative research is to analyze a putative causal relationship within groups defined by one or more variables. Whether referred to as subgroup analysis, subclassification, stratification, or tabular decomposition, the usual motivation is to analyze the data after conditioning on membership in groups identified by values of a variable that is thought to be related to both the causal variable and the outcome variable.

From a graphical perspective, the result of such a modeling strategy is to generate simplified subgraphs for each subgroup or stratum of the data that correspond to each value of the conditioning variable. This procedure is analogous to disconnecting the conditioning variable from all other variables that it points

to in the original graph and rewriting the graph as many times as there are values for the conditioning variable. In the mutual dependence graph in panel (b) of Figure 3.3, conditioning on C results in separate graphs for each value c of C; in each of these subgraphs, A and B are disconnected. The reasoning here should be obvious: If analysis is carried out for a group in which all individuals have a particular value for the variable C, then the variable C is constant within the group and cannot therefore be associated with A or B. Thus, from a graphical perspective, conditioning is a means of transforming one graph into a simpler set of component graphs where fewer causes are represented.

As a technique for estimating a causal effect, conditioning is a very powerful and very general strategy. But one very important qualification must be noted: Conditioning on a collider variable does not simplify the original graph but rather adds complications by creating new associations. To see why, reconsider the mutual causation graph in panel (c) of Figure 3.3, where A and B are unrelated to each other but where both cause the collider variable C. In this case, conditioning on C will induce a relationship between A and B for at least one subgroup defined by the values of C.

The reasoning here is not intuitive, but it can be conveyed by a simple example with the mutual causation graph in panel (c) of Figure 3.3. 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 panel (c) of Figure 3.3. Even though SAT score 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, those with the highest SAT scores are on average the least motivated, and those with the lowest SAT scores are on average the most motivated. Thus, the college's sorting of applicants generates a pool of admitted students within which SAT and motivation are negatively related.[5]

This example is depicted in Figure 3.4 for 250 simulated applicants to this hypothetical college. For this set of applicants, SAT and motivation have a very small positive correlation of .035.[6] Offers of admission are then determined by the sum of SAT and motivation and granted to the top 15 percent of applicants

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

[6] 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.

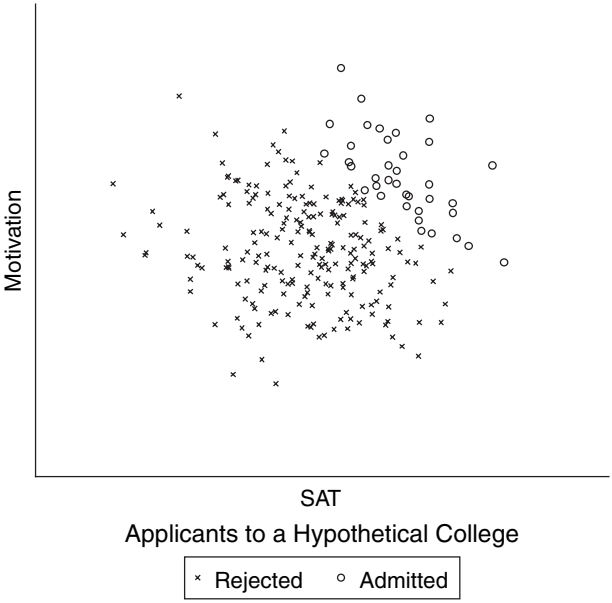


Figure 3.4: Simulation of conditional dependence within values of a collider variable.

(as shown in the upper right-hand portion of Figure 3.4).[7] 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 examples, this is a very general feature of causal relationships and is present in many real-world applications. 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.

3.1.3 Point Identification by Conditioning on Variables that Satisfy the Back-Door Criterion

Pearl elaborates three different approaches to identifying causal effects, which we already introduced in Section 1.6 as (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 an

[7] 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.

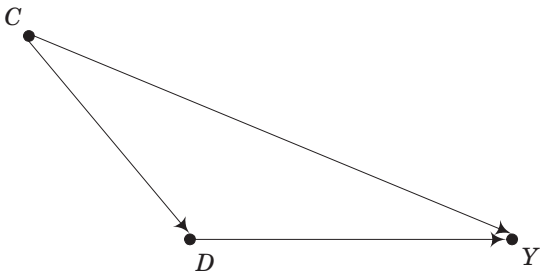


Figure 3.5: A causal diagram in which the effect of D on Y is confounded by C.

isolated and exhaustive mechanism (or set of mechanisms) 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 subsection, we consider the first of these strategies, which motivates the basic matching and regression techniques that we will present in the next two chapters.

Perhaps the most general concern of a researcher seeking to estimate a causal effect is that the causal variable D and the outcome variable Y are mutually dependent on a common third variable C. This common but simple scenario is represented by the DAG in Figure 3.5. In this case, the total association between D and Y represents the genuine causal effect of D on Y and the common dependence of D and Y on C. In this case, it is often said that the causal effect of D on Y is confounded by C. In particular, the presence of the causal effects $C \rightarrow D$ and $C \rightarrow Y$ confound the causal effect $D \rightarrow Y$.

For this example, the causal effect of D on Y can be consistently estimated by conditioning on C. We will explain this claim more formally and more generally in the remainder of this section. For now, consider conditioning only as a data analysis procedure in order to understand the end point of the discussion that follows. As an operational data analysis routine, the effect of D on Y can be estimated by conditioning on C in two steps: (1) calculate the association between D and Y for each subgroup with C equal to c and then (2) average these c-specific associations over the marginal distribution of the values c that the variable C takes on. The resulting weighted average is the causal effect of D on Y in Pearl's framework, which would be labeled the average treatment effect in the counterfactual causality literature. In this sense, conditioning on C identifies the causal effect of D on Y in this example.

Conditioning is a powerful strategy for estimating causal effects, and it is both successful and completely transparent for simple examples such as this one. But it is a much more complicated procedure in general than is suggested by our short accounting of this example. The complications arise when colliders are present, and Pearl has explained systematically how to resolve these complications.

Before considering a complex example, consider a more formal analysis of the example in Figure 3.5. Pearl characterizes the confounding created by both $C \rightarrow D$ and $C \rightarrow Y$ in a novel way, using the language of back-door paths. For Pearl, a path is any any sequence of edges pointing in any direction that connects one variable to another. A back-door path is then defined as a path between any causally ordered sequence of two variables that includes a directed edge $\rightarrow$ that points to the first variable. For the DAG in Figure 3.5, there are two paths that connect D and Y: $D \rightarrow Y$ and $D \leftarrow C \rightarrow Y$. For the causally ordered pair of variables D and Y, the path $D \leftarrow C \rightarrow Y$ is a back-door path because it includes a directed edge pointing to D. Likewise, the path $D \rightarrow Y$ is not a back-door path because it does not include a directed edge pointing to D.

The problem with back-door paths is that they may contribute to the association between D and Y. As a result, the observed association between D and Y may not consistently estimate the causal effect of D on Y. In Pearl's language, the observed association between D and Y does not identify the causal effect because the total association between D and Y is an unknown composite of the true causal effect $D \rightarrow Y$ and the back-door path $D \leftarrow C \rightarrow Y$.

With this language, Pearl then develops 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. If one or more back-door paths connects the causal variable to the outcome variable, Pearl shows that the causal effect is identified by conditioning on a set of variables Z if and only if all back-door paths between the causal variable and the outcome variable are blocked after conditioning on Z. He then proves that all back-door paths are blocked by Z if and only if each back-door path

1. contains a chain of mediation $A \rightarrow C \rightarrow B$, where the middle variable C is in Z, or

2. contains a fork of mutual dependence $A \leftarrow C \rightarrow B$, where the middle variable C is in Z, or

3. contains an inverted fork of mutual causation $A \rightarrow C \leftarrow B$, where the middle variable C and all of C's descendants are *not* in Z.[8]

Conditions 1 and 2 of the back-door criterion should be clear as stated; they imply that back-door associations between the causal variable and the outcome variable can be eliminated by conditioning on observed variables that block each back-door path. Condition 3, however, is quite different and is not intuitive. It states instead that the set of conditioning variables Z cannot include collider

[8]This claim is a combination of Pearl's definition of d-separation (Pearl 2000:16-17) and his definition of the back-door criterion (Pearl 2000:79). The back-door criterion is interpretable only when the causal effect of interest is specified as a component of a graph that is representable as a causal model (or at least a component of locally Markovian causal model, in which the underspecified causal relations are irrelevant to an evaluation of the back-door criterion for the particular causal effect under consideration). See Pearl's causal Markov condition for the existence of a causal model (Pearl 2000, Section 1.4.2, Theorem 1.4.1).

variables that lie along back-door paths.[9] We will explain the importance of condition 3, as well as the underlying rationale for it, in the examples that follow.[10]

First, return one last time to the simple example in Figure 3.5. Here, there is a single back-door path, which is a fork of mutual dependence where C causes both D and Y. By Pearl's back-door criterion, conditioning on C blocks $D \leftarrow C \rightarrow Y$ because C is the middle variable in a fork of mutual dependence. As a result, C satisfies Pearl's back-door criterion, and the causal effect of D on Y is identified by conditioning on C.

Consider now a more complex example, which involves the do-not-condition-on-colliders condition 3 of the back-door criterion. 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.

To set up this example, suppose that an analyst is confronted with a basic DAG, similar to the one in Figure 3.5, in which the causal effect of D on Y is confounded by a variable such as C. But suppose now that the confounder variable is unobserved, and thus cannot be conditioned on. 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 3.6.

For the DAG in Figure 3.6, there are two back-door paths from D to Y: $D \leftarrow V \rightarrow Y_{t-1} \rightarrow Y$ and $D \leftarrow V \rightarrow Y_{t-1} \leftarrow U \rightarrow Y$. The lagged outcome variable Y_{t-1} lies along 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 $D \leftarrow V \rightarrow Y_{t-1} \rightarrow Y$ because, for this path, Y_{t-1} is the middle variable of 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 $V \rightarrow Y_{t-1} \leftarrow U$. And, as a result, the back-door criterion states that, after conditioning on Y_{t-1}, at least one back-door path from D to Y will remain unblocked. For this example, it is the second path that includes the collider.

Having seen how condition 3 of the back-door criterion is applied, consider why it is so important. Pearl would state that, for this last example,

[9] Because the "or" in 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 or condition 2 with respect to another variable on the same back-door path.

[10] Note the stipulation in condition 3 that neither C nor the descendants of C can be in Z. We do not make much of these "descendants" in our presentation. But, see Hernan, Hernandez-Diaz, and Robins (2004) for a discussion of examples in epidemiology for which the distinction is important. For their examples, the collider is "getting sick enough to be admitted to a hospital for treatment" but 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.

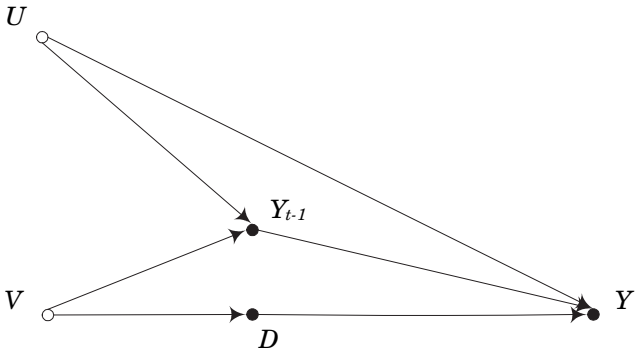


Figure 3.6: A causal diagram in which Y_{t-1} is a collider.

conditioning on Y_{t-1} eliminates part of the back-door association between D and Y because Y_{t-1} blocks the back-door path $D \leftarrow V \rightarrow Y_{t-1} \rightarrow Y$. But, at the same time, conditioning on Y_{t-1} creates 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 simple characterization of colliders and the inverted fork of mutual causation in panel (c) of Figure 3.3. For that graph, the path $A \rightarrow C \leftarrow B$ contains a collider C. As we noted earlier, the indirect causal effects of A and B on each other are absorbed by C, and the path $A \rightarrow C \leftarrow B$ does not on its own generate an unconditional association between A and B. This basic result can be applied to back-door paths between D and Y that include colliders, as in the example presented in Figure 3.6. If a back-door path between a causal variable D and an outcome variable Y includes an intermediate variable that is a collider, that back-door path does not contribute to the unconditional association between D and Y. Because no back-door association between D and Y is generated, a back-door path that contains a collider does not confound the causal effect of D on Y.

At the same time, if a collider that lies along a back-door path is conditioned on, that conditioning will unblock the back-door path and thereby confound the causal effect. Recall the earlier discussion of conditioning in reference to panel (c) of Figure 3.3 and then as demonstrated in Figure 3.4. 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 3.6 is considered. Conditioning on a collider that lies along a back-door path unblocks the back-door path in the sense that it creates a net 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 the DAG in Figure 3.7 (which is similar to Figure 1.1, except that the bidirected edges

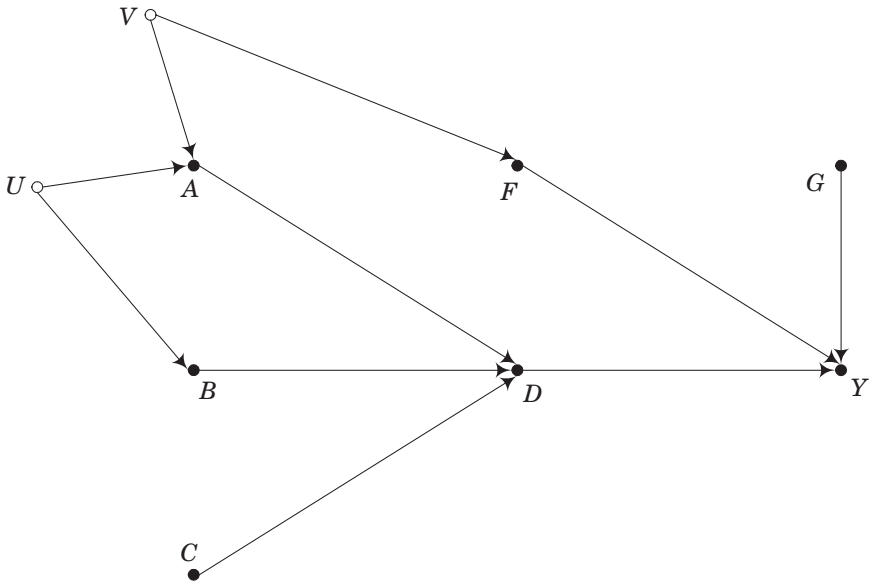


Figure 3.7: A causal diagram in which A is a collider.

that signified unspecified 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 DAG, 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 contributes to the association between D and Y, but the second back-door path does not contribute to the association between D and Y. We have to be careful that, whatever conditioning we enact to eliminate the confounding effect of the back-door path $D \leftarrow A \leftarrow V \rightarrow F \rightarrow Y$ does not unblock the back-door path $D \leftarrow B \leftarrow U \rightarrow A \leftarrow V \rightarrow F \rightarrow Y$ and thereby confound the causal effect in another way.

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 of a chain of mediation for both back-door paths, as in $V \rightarrow F \rightarrow Y$. 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 of a chain of mediation for the first back-door path, as in $D \leftarrow A \leftarrow V$. But A is a collider variable for the second back-door path, as in $U \rightarrow A \leftarrow V$. 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 back-door association between D and Y.

3. A is a middle variable of a chain of mediation for the first back-door path, as in $D \leftarrow A \leftarrow V$. Likewise, B is a middle variable of a chain of mediation for the second back-door path, as in $D \leftarrow B \leftarrow U$. 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.[11]

The key point of this section is that conditioning on variables that lie along 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. And, as a result, the remaining association between the causal variable and outcome variable identifies the causal effect. Even so, it must be kept in mind that conditioning on a collider variable has the opposite effect. It unblocks an already blocked back-door path. And thus, as the last two examples show, when a conditioning strategy is evaluated, each back-door path must be assessed carefully because a variable can be a collider along one back-door path but not a collider along another.

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. 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 renders all back-door paths blocked. Moreover, 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 these back-door paths is a much simpler way to begin to consider the full complications of a conditioning strategy than prior approaches.

[11] 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.

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 – can be thought of as more specific assertions of the general point that the average causal effect is identified when all back-door paths are blocked.

3.2 Models of Causal Exposure in the Counterfactual Tradition

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, in the counterfactual tradition 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 known, in the counterfactual framework, as modeling the treatment assignment mechanism or as modeling the treatment selection mechanism, based on which tradition of analysis is followed.

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.

3.2.1 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 earlier, if treatment assignment is completely random, 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 (2.4)]. In this case, the treatment assignment mechanism can be specified completely if $\Pr[D = 1]$ is set to a constant between 0 and 1. If a researcher desires treatment and control groups of approximately the same size, then $\Pr[D = 1]$ can be set to .5. Individual realized values of D for those in the study, denoted d_i generically, are then equal to 1 or 0. The values for d_i can be thought of as realized outcomes of a Bernoulli trial for the random variable D.

For more complex randomization schemes, more elaborate statements are required. If, for example, study subjects are stratified first by gender and then assigned with disproportionate probability to the treatment group if female, then the treatment assignment mechanism might instead be

$$\begin{aligned} \Pr[D = 1|\text{Gender} = \text{Female}] &= .7, \\ \Pr[D = 1|\text{Gender} = \text{Male}] &= .5. \end{aligned} \tag{3.1}$$

These conditional probabilities are often referred to as propensity scores, as they indicate the propensity that an individual with specific characteristics will be observed in the treatment group. Accordingly, for this example, Equations (3.1) are equivalent to stating that the propensity score for the treatment is .7 for females and .5 for males. In randomized experiments, the propensity scores are known to the researcher.

In contrast, a researcher with observational data only does not possess *a priori* knowledge of the propensity scores that apply to different types of individuals. However, she may know the characteristics of individuals that systematically determine their propensity scores. In this case, treatment selection patterns are represented by the general conditional probability distribution:

$$\Pr[D = 1|S], \tag{3.2}$$

where S now denotes all variables that systematically determine treatment assignment/selection. An observational researcher may know and have measures of all of the variables in S, even though he or she may not know the specific values of the propensity scores, which are defined as the values that $\Pr[D = 1|S]$ is equal to under different realized values s of the variables in S. Complete observation of S allows a researcher to assert that treatment selection is "ignorable" and then consistently estimate the average treatment effect, 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 ignorable. The core of the concept of ignorability is the independence assumption that was introduced earlier in Equation (2.4):

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

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. But ignorability also holds in the weaker case,

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid S, \tag{3.3}$$

and when S is fully observed. The treatment assignment mechanism is ignorable when the potential outcomes (and any function of them, such as δ) are independent of the treatment variable within strata defined by all combinations

of values on the observed variables in S that determine treatment selection.[12]

If ignorability of treatment assignment is asserted for an observational study, then a researcher must (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.[13] 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 selection 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 average treatment effect.

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 average treatment effect 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. This strategy is exactly the conditioning strategy introduced in the last section: Parents' religious identification blocks all back-door paths from Catholic school attendance to test scores.

Ignorability is thus directly related to conditioning on variables that satisfy the back-door criterion of Pearl. Suppose that we are confronted with the causal diagram in panel (a) of Figure 3.8, which includes the causal effect $D \rightarrow Y$ but also the bidirected edge $D \dashleftarrow\!\dashrightarrow Y$. 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 panel (a) of Figure 3.8. The simplest such model is presented in panel (b) of Figure 3.8, where $D \dashleftarrow\!\dashrightarrow Y$ has been replaced with

[12]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 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).

[13]This third step can be weakened if the data are merely sparse, as we discuss later when presenting propensity score techniques.

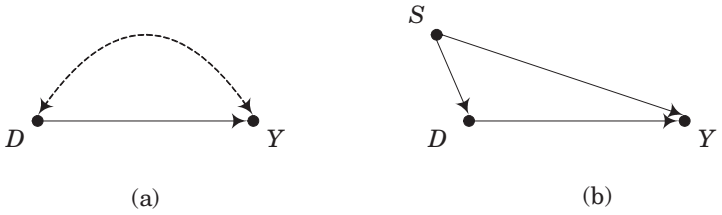


Figure 3.8: Causal diagrams in which treatment assignment is (a) nonignorable and (b) ignorable.

the back-door path $D \leftarrow S \rightarrow Y$. Thus, if S is observed, then conditioning on S will solve the causal inference problem. When identification by back-door conditioning is feasible, then treatment selection is ignorable.[14]

We will discuss these techniques in detail in Chapter 4, where we present matching estimators of causal effects. But the immediate complications of undertaking this 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, the treatment selection mechanism remains nonignorable, as treatment selection is then a function of unobserved characteristics of students.

3.2.2 Treatment Selection Modeling in Econometrics

The econometrics literature also has a long tradition of analyzing causal effects of these forms, and this literature may be more familiar to social scientists. Whereas concepts such as ignorability are used 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.

[14]The potential outcome random variables are not represented in Figure 3.8. Pearl introduces the causal states in a different way, using the semantics of an intervention and introducing the $do(.)$ operator. For the causal diagram in panel (a) of Figure 3.8, the average values of Y that accompany the two values of D do not correspond to the two values of Y that one would obtain by calculating the average values of Y that would result from intervening to set D to its two possible values. For the causal diagram in panel (a) of Figure 3.8, $E[Y|D = 1] - E[Y|D = 0]$ does not equal $E[Y|do(D = 1] - E[Y|do(D = 0)]$. As should be clear, holding all semantic issues aside, this last statement is equivalent to saying that, for observations of variables D and Y for the model in panel (a) of Figure 3.8, the naive estimator does not equal the average difference in potential outcomes. Thus, the $do(.)$ operator is equivalent to the assertion of well-defined causal states, and that assertion is attached to the causal variable rather than to potential outcomes for each causal state. Thus, for Pearl, $E[Y^1] - E[Y^0]$ is equal to $E[Y|do(D = 1)] - E[Y|do(D = 0)]$.

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:

$$\begin{aligned} Y &= Y^0 + (Y^1 - Y^0)D \\ &= Y^0 + \delta D \\ &= \mu^0 + \delta D + \upsilon^0, \end{aligned} \tag{3.4}$$

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 (3.4) so that potential variability of δ across individuals in the treatment and control groups is relegated to the error term, as in

$$Y = \mu^0 + (\mu^1 - \mu^0)D + \{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}, \tag{3.5}$$

where $\mu^1 \equiv E[Y^1]$, $\upsilon^1 \equiv Y^1 - E[Y^1]$, and all else is as defined for Equation (3.4).[15] Note that, in evolving from Equation (2.2) to Equation (3.5), the

[15]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 manpower training example. For each individual i, the earnings of individual i if trained are

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

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 average treatment effect, which we defined as $E[\delta]$ in Equation (2.3), and ε_i is the individual-level departure of δ_i from the average treatment effect $E[\delta]$. Although the notation in this last equation differs from the notation in Equation (3.5), the two equations are equivalent. Heckman and Vytlacil (2005) give a fully nonparametric version of this treatment

definition of the observed outcome variable Y has taken on the look and feel of a regression model.[16] 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 average causal effect $E[\delta]$. 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.[17]

Heckman and Robb use the specification of the treatment evaluation problem in Equation (3.5), 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 average treatment effect [in this case $(\mu^1 - \mu^0)$] when D is correlated with the population-level variant of the error term in braces in Equation (3.5), 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 (3.5), hoping to break the correlation between D and the error term.[18] Heckman and Robb show that this strategy is generally ineffective with the data available on manpower 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 5, 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 (3.5) can be broken. They proceed by proposing that treatment selection be modeled by specifying a latent continuous variable $\tilde{D}$:

$$\tilde{D} = Z\phi + U, \tag{3.6}$$

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 vari-

selection framework, which we draw on later.

[16] Sometimes, Equation (3.5) 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 fully articulated version of Equation (3.4).

[17] 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.

[18] 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.

able), 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 (3.6) is then related to the treatment selection dummy, D, by

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

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.[19] This is shown in the DAG in panel (b) of Figure 3.8. The corresponding causal diagram for the econometric selection equation is presented in two different graphs in Figure 3.9, as there are two scenarios corresponding to whether or not all elements of S have been observed as Z.

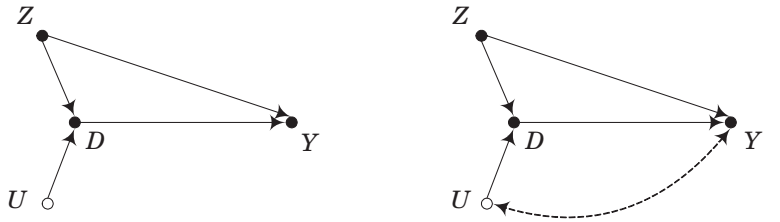
For the case in which Z in Equation (3.6) is equivalent to the set of variables in S in Equation (3.2), treatment selection is ignorable, as defined in Equation (3.3), because conditioning on Z is exactly equivalent to conditioning on S. In the econometric tradition, this situation would not, however, 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 of the ignorability assumption: All systematic determinants of treatment selection have been observed.

The scenario of selection on the observables is depicted in panel (a) of Figure 3.9. The variable S in panel (b) of Figure 3.8 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 panel (a) of Figure 3.9 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 panel (a) of Figure 3.9, which would render this DAG the same as the one in panel (b) of Figure 3.8.[20]

Now consider the case in which the observed treatment selection variables in Z are only a subset of the variables in S. In this case, some components of

[19] 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.

[20] 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.



(a) Selection on the observables (b) Selection on the unobservables

Figure 3.9: Causal diagrams for the terminology from econometric modeling of treatment selection.

S enter into the treatment selection latent variable $\tilde{D}$ through the error term, U, of Equation (3.6). In this case, treatment selection is nonignorable. Or, in the words of econometricians, "selection is on the unobservables." The scenario of selection on the unobservables is depicted in panel (b) of Figure 3.9, where there is now a back-door path from D to Y: $D \leftarrow U \dashleftarrow\dashrightarrow Y$. Conditioning on Z for this causal diagram 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 (3.5) (see Heckman and Vytlacil 2005, which we will discuss later). 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 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.

3.3 Conditioning to Balance versus 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 two chapters that present both matching and regression implementations of conditioning. Although we will show that both sets of techniques 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

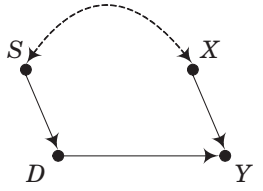


Figure 3.10: A causal diagram in which sufficient conditioning can be performed with respect to S or X.

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

$$\Pr[W|D = 1] = \Pr[W|D = 0], \tag{3.7}$$

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

Now consider the graph presented in Figure 3.10. A back-door path $D \leftarrow S \dashleftarrow\dashrightarrow X \rightarrow Y$ is present from D to Y, where S represents the complete set of variables that are direct causes of treatment assignment/selection, X represents 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.[21]

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 paths signified by 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

[21] 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 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.

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. For the graph in Figure 3.10, 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 3.10 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.

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. Perhaps the best answer is that one should do both.[22] Nonetheless, there is a specific advantage of balancing that may tip the scales in its favor if both strategies are feasible: It diminishes the inferential problems that can be induced by data-driven specification searches. We will discuss these issues in the course of presenting matching and regression conditioning strategies in the next two chapters.

3.4 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 graphical causal models and identification issues on the estimation of the unconditional average treatment effect. This narrow focus is entirely consistent with the graphical tradition, in which parameters such as the average treatment effect for the treated in Equation (2.5) and the average treatment effect for the untreated in Equation (2.6) are given considerably less attention than in the counterfactual 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 two chapters.

Identification When the Unconditional Average Treatment Effect is Identified

If one can identify and consistently estimate the unconditional average treatment effect with conditioning techniques, then one can usually estimate some of the

[22] As we discuss in Subsection 5.3.4, 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 break the back-door path between Y and D.

conditional average treatment effects that may be of interest as well. As we will show in the next two 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 average treatment effect 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 3.10. The back-door path from D to Y is blocked by both S and X. As a result, a consistent estimate of the average treatment effect in Equation (2.3) can be obtained by conditioning on either S or X. But, in addition, consistent estimates of the average treatment effect for the treated in Equation (2.5) and the average treatment effect for the untreated in Equation (2.6) can be obtained by properly weighting conditional differences in the observed values of Y. In particular, if conditioning is performed with respect to S, 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 average treatment effect for the treated and the average treatment effect for the untreated, respectively. If, on the other hand, conditioning is performed with respect to X, then alternative quantities are calculated as sample analogs to $E[Y|D = 1, X = x] - E[Y|D = 0, X = x]$, $\Pr[X|D = 1]$, and $\Pr[X|D = 0]$. These estimated quantities will differ from those that are generated by conditioning on S, but they can still be used in an analogous way to form consistent estimates of the average treatment effect for the treated and the average treatment effect for the untreated. We will present examples of these sorts of stratification and weighting estimators in the next chapter.

But note that the ingredients utilized to estimate the average treatment effect for the treated and the average treatment effect for the untreated in these two conditioning routines are quite different. 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 counterfactual 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.

Identification When the Unconditional Average Treatment Effect is Not Identified

If selection is on the unobservables, conditioning strategies will generally fail to identify unconditional average treatment effects. Nonetheless, weaker as-

sumptions 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 two 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 prior discussion in Subsection 2.6.4 of similar issues with regard to the bias of the naive estimator.)

Suppose, for example, that the graph in panel (b) of Figure 3.9 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 average treatment effect. 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 average treatment effect for the treated 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 $(Y^0, Y^1) \perp\!\!\!\perp D \mid S$ does not, then the average treatment effect for the untreated can be estimated consistently.[23]

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 average treatment effect for the treated 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 (2.14), 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 treatment effect for the treated and the treatment effect for the untreated in the next chapter.

3.5 Conclusions

In the next two chapters, we present details and connections between the two main types of conditioning estimation strategies: matching and regression. We show how they generally succeed when selection is on the observables and fail when selection is on the unobservables. We lay out the specific assumptions that

[23] And, as we will show later, the required assumptions are even simpler, as the entire distributions of Y^0 and Y^1 need not be conditionally independent of D. As long as SUTVA holds, only mean independence must be maintained.

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. In later chapters, we then present more complex methods for identifying and estimating causal effects when simple conditioning methods are insufficient.

Chapter 5

Regression Estimators of Causal Effects

Regression models are perhaps the most common form of data analysis used to evaluate alternative explanations for outcomes of interest to quantitatively oriented social scientists. In the past 40 years, a remarkable variety of regression models have been developed by statisticians. Accordingly, most major data analysis software packages allow for regression estimation of the relationships between interval and categorical variables, in cross sections and longitudinal panels, and in nested and multilevel patterns. In this chapter, however, we restrict our attention to OLS regression, focusing mostly on the regression of an interval-scaled variable on a binary causal variable. As we will show, the issues are complicated enough for these models. And it is our knowledge of how least squares models work that allows us to explain the complexity.

In this chapter, we present least squares regression from three different perspectives: (1) regression as a descriptive modeling tool, (2) regression as a parametric adjustment technique for estimating causal effects, and (3) regression as a matching estimator of causal effects. We give more attention to the third of these three perspectives on regression than is customary in methodological texts because this perspective allows one to understand the others from a counterfactual perspective. At the end of the chapter, we will draw some of the connections between least squares regression and more general models, and we will discuss the estimation of causal effects for many-valued causes.

5.1 Regression as a Descriptive Tool

Least squares regression can be justified without reference to causality, as it can be considered nothing more than a method for obtaining a best-fitting descriptive model under entailed linearity constraints. Goldberger (1991), for example, motivates least squares regression as a technique to estimate a best-fitting

linear approximation to a conditional expectation function that may be nonlinear in the population.

Consider this descriptive motivation of regression a bit more formally. If X is a collection of variables that are thought to be associated with Y in some way, then the conditional expectation function of Y, viewed as a function in X, is denoted $E[Y|X]$. Each particular value of the conditional expectation for a specific realization x of X is then denoted $E[Y|X = x]$.

Least squares regression yields a predicted surface $\hat{Y} = X\hat{\beta}$, where $\hat{\beta}$ is a vector of estimated coefficients from the regression of the realized values y_i on x_i. The predicted surface, $X\hat{\beta}$, does not necessarily run through the specific points of the conditional expectation function, even for an infinite sample, because (1) the conditional expectation function may be a nonlinear function of one or more of the variables in X and (2) a regression model can be fit without parameterizing all nonlinearities in X. An estimated regression surface simply represents a best-fitting linear approximation of $E[Y|X]$ under whatever linearity constraints are entailed by the chosen parameterization.[1]

The following demonstration of this usage of regression is simple. Most readers know this material well and can skip ahead to the next section. But, even so, it may be worthwhile to read the demonstration quickly because we will build directly on it when shifting to the consideration of regression as a causal effect estimator.

Regression Demonstration 1

Recall the stratification example presented as Matching Demonstration 1 (see page 92 in Chapter 4). Suppose that the same data are being analyzed, as generated by the distributions presented in Tables 4.1 and 4.2; features of these distributions are reproduced in Table 5.1 in more compact form. As before, assume that well-defined causal states continue to exist and that S serves as a perfect stratification of the data.[2] Accordingly, the conditional expectations in the last three panels of Table 5.1 are equal as shown.

But, for this demonstration of regression as a descriptive tool, assume that a cautious researcher does not wish to rush ahead and attempt to estimate the specific underlying causal effect of D on Y, either averaged across all individuals or averaged across particular subsets of the population. Instead, the researcher is cautious and is willing to assert only that the variables S, D, and Y constitute some portion of a larger system of causal relationships. In particular, the researcher is unwilling to assert anything about the existence or nonexistence of other variables that may also lie on the causal chain from S, through D, to Y. This is tantamount to doubting the claim that S offers a perfect stratification of the data, even though that claim is true by construction for this example.

[1] One can fit a large variety of nonlinear surfaces with regression by artful parameterizations of the variables in X, but these surfaces are always generated by a linear combination of a coefficient vector and values on some well-defined coding of the variables in X.

[2] For this section, we will also stipulate that the conditional variances of the potential outcomes are constant across both of the potential outcomes and across levels of S.

Table 5.1: The Joint Probability Distribution and Conditional Population Expectations for Regression Demonstration 1

	Joint probability distribution of S and D	
	Control group: $D = 0$	Treatment group: $D = 1$
$S = 1$	$\Pr[S = 1, D = 0] = .36$	$\Pr[S = 1, D = 1] = .08$
$S = 2$	$\Pr[S = 2, D = 0] = .12$	$\Pr[S = 2, D = 1] = .12$
$S = 3$	$\Pr[S = 3, D = 0] = .12$	$\Pr[S = 3, D = 1] = .2$
	Potential outcomes under the control state	
$S = 1$	$E[Y^0 \mid S = 1, D = 0] = 2$	$E[Y^0 \mid S = 1, D = 1] = 2$
$S = 2$	$E[Y^0 \mid S = 2, D = 0] = 6$	$E[Y^0 \mid S = 2, D = 1] = 6$
$S = 3$	$E[Y^0 \mid S = 3, D = 0] = 10$	$E[Y^0 \mid S = 3, D = 1] = 10$
	Potential outcomes under the treatment state	
$S = 1$	$E[Y^1 \mid S = 1, D = 0] = 4$	$E[Y^1 \mid S = 1, D = 1] = 4$
$S = 2$	$E[Y^1 \mid S = 2, D = 0] = 8$	$E[Y^1 \mid S = 2, D = 1] = 8$
$S = 3$	$E[Y^1 \mid S = 3, D = 0] = 14$	$E[Y^1 \mid S = 3, D = 1] = 14$
	Observed outcomes	
$S = 1$	$E[Y \mid S = 1, D = 0] = 2$	$E[Y \mid S = 1, D = 1] = 4$
$S = 2$	$E[Y \mid S = 2, D = 0] = 6$	$E[Y \mid S = 2, D = 1] = 8$
$S = 3$	$E[Y \mid S = 3, D = 0] = 10$	$E[Y \mid S = 3, D = 1] = 14$

In this situation, suppose that the researcher simply wishes to estimate the best linear approximation to the conditional expectation $E[Y|D, S]$ and does not wish to then give a causal interpretation to any of the coefficients that define the linear approximation. The six true values of $E[Y|D, S]$ are given in the last panel of Table 5.1. Notice that the linearity of $E[Y|D, S]$ in D and S is present only when $S \leq 2$. The value of 14 for $E[Y|D = 1, S = 3]$ makes $E[Y|D, S]$ nonlinear in D and S over their full distributions.

Now consider the predicted surfaces that would result from the estimation of two alternative least squares regression equations with data from a sample of infinite size (to render sampling error zero). A regression of Y on D and S that treats D as a dummy variable and S as an interval-scaled variable would yield a predictive surface of

$$\hat{Y} = -2.71 + 2.69(D) + 4.45(S). \tag{5.1}$$

This model constrains the partial association between Y and S to be linear. It represents a sensible predicted regression surface because it is a best-fitting,

linear-in-the-parameters model of the association between Y and the two variables D and S, where "best" is defined as minimizing the average squared differences between the fitted values and the true values of the conditional expectation function.

For this example, one can offer a better descriptive fit at little interpretive cost by using a more flexible parameterization of S. An alternative regression that treats S as a discrete variable represented in the estimation routine by dummy variables $S2$ and $S3$ (for S equal to 2 and S equal to 3, respectively) would yield a predictive surface of

$$\hat{Y} = 1.86 + 2.75(D) + 3.76(S2) + 8.92(S3). \tag{5.2}$$

Like the predicted surface for the model in Equation (5.1), this model is also a best linear approximation to the six values of the true conditional expectation $E[Y|D, S]$. The specific estimated values are

$$\begin{aligned}
D = 0, S = 1 &: \ \hat{Y} = 1.86, \\
D = 0, S = 2 &: \ \hat{Y} = 5.62, \\
D = 0, S = 3 &: \ \hat{Y} = 10.78, \\
D = 1, S = 1 &: \ \hat{Y} = 4.61, \\
D = 1, S = 2 &: \ \hat{Y} = 8.37, \\
D = 1, S = 3 &: \ \hat{Y} = 13.53.
\end{aligned}$$

In contrast to the model in Equation (5.1), for this model the variable S is given a fully flexible coding. As a result, parameters are fit that uniquely represent all values of S.[3] The predicted change in Y for a shift in S from 1 to 2 is 3.76

[3]The difference between a model in which a variable is given a fully flexible coding and one in which it is given a more constrained coding is clearer for a simpler conditional expectation function. For $E[Y|S]$, consider the values in the cells of Table 5.1. The three values of $E[Y|S]$ can be obtained from the first and fourth panels of Table 5.1 as follows:

$$\begin{aligned}
E[Y|S=1] &= \frac{.36}{(.36+.08)}(2) + \frac{.08}{(.36+.08)}(4) = 2.36, \\
E[Y|S=2] &= \frac{.12}{(.12+.12)}(6) + \frac{.12}{(.12+.12)}(8) = 7, \\
E[Y|S=3] &= \frac{.12}{(.12+.2)}(10) + \frac{.2}{(.12+.2)}(14) = 12.5.
\end{aligned}$$

Notice that these three values of $E[Y|S]$ do not fall on a straight line; the middle value of 7 is closer to 2.36 than it is to 12.5.

For $E[Y|S]$, a least squares regression of Y on S, treating S as an interval-scaled variable, would yield a predictive surface of

$$\hat{Y} = -2.78 + 5.05(S).$$

The three values of this estimated regression surface lie on a straight line −2.27, 7.32, and 12.37 – and they do not match the corresponding true values of 2.36, 7, and 12.5. A regression of Y on S, treating S as a discrete variable with dummy variables $S2$ and $S3$, would yield an alternative predictive surface of

$$\hat{Y} = 2.36 + 4.64(S2) + 10.14(S3).$$

(i.e., $5.62 - 1.86 = 3.76$ and $8.37 - 4.61 = 3.76$) whereas the predicted change in Y for a shift in S from 2 to 3 is 5.16 (i.e., $10.78 - 5.62 = 5.16$ and $13.53 - 8.37 = 5.16$).

Even so, the model in Equation (5.2) constrains the parameter for D to be the same without regard to the value of S. And, because the level of Y depends on the interaction of S and D, specifying more than one parameter for the three values of S does not bring the predicted regression surface into alignment with the six values of $E[Y|D, S]$ presented in the last panel of Table 5.1. Thus, even when S is given a fully flexible coding (and even for an infinitely large sample), the fitted values do not equal the true values of $E[Y|D, S]$.[4] As we discuss later, a model that is saturated fully in both S and D – that is, one that adds two additional parameters for the interactions between D and both $S2$ and $S3$ – would yield predicted values that would exactly match the six true values of $E[Y|D, S]$ in a dataset of sufficient size.

Recall the more general statement of the descriptive motivation of regression analysis presented earlier, in which the predicted surface $\hat{Y} = X\hat{\beta}$ is estimated for the sole purpose of obtaining a best-fitting linear approximation to the true conditional expectation function $E[Y|X]$. When the purposes of regression are so narrowly restricted, the outcome variable of interest, Y, is not generally thought to be a function of potential outcomes associated with well-defined causal states. Consequently, it would be inappropriate to give a causal interpretation to any of the estimated coefficients in $\hat{\beta}$.

This perspective implies that if one were to add more variables to the predictors, embedding X in a more encompassing set of variables W, then a new set of least squares estimates $\hat{\gamma}$ could be obtained by regressing Y on W. The estimated surface $W\hat{\gamma}$ then represents a best-fitting, linear-in-the-parameters, descriptive fit to a more encompassing conditional expectation function, $E[Y|W]$. Whether one then prefers $W\hat{\gamma}$ to $X\hat{\beta}$ as a description of the variation in Y depends on whether one finds it more useful to approximate $E[Y|W]$ than $E[Y|X]$. The former regression approximation is often referred to as the long regression, with the latter representing the short regression. These labels are aptly chosen, when regression is considered nothing more than a descriptive tool, as there is no inherent reason to prefer a short to a long regression if neither is meant to

This second model uses a fully flexible coding of S, and each value of the conditional expectation function is a unique function of the parameters in the model (that is, $2.36 = 2.36$, $4.64 + 2.36 = 7$, and $10.14 + 2.36 = 12.5$). Thus, in this case, the regression model would, in a suitably large sample, estimate the three values of $E[Y|S]$ exactly.

[4] Why would one ever prefer a constrained regression model of this sort? Consider a conditional expectation function, $E[Y|X]$, where Y is earnings and X is years of education (with 21 values from 0 to 20). A fully flexible coding of X would fit 20 dummy variables for the 21 values of X. This would allow the predicted surface to change only modestly between some years (such as between 7 and 8 and between 12 and 13) and more dramatically between other years (such as between 11 and 12 and between 15 and 16). However, one might wish to treat X as an interval-scaled variable, smoothing these increases from year to year by constraining them to a best-fitting line parameterized only by an intercept and a constant slope. This constrained model would not fit the conditional expectation function as closely as the model with 20 dummy variables, but it might be preferred in some situations because it is easier to present and easier to estimate for a relatively small sample.

be interpreted as anything other than a best-fitting linear approximation to its respective true conditional expectation function.

In many applied regression textbooks, the descriptive motivation of regression receives no direct explication. And, in fact, many textbooks state that the only correct specification of a regression model is one that includes all explanatory variables. Goldberger (1991) admonishes such textbook writers, countering their claims with:

> An alternative position is less stringent and is free of causal language. Nothing in the CR [classical regression] model itself requires an exhaustive list of explanatory variables, nor any assumption about the direction of causality. (Goldberger 1991:173)

Goldberger is surely correct, but his perspective nonetheless begs an important question on the ultimate utility of descriptively motivated regression. Clearly, if one wishes to know only predicted values of the outcome Y for those not originally studied but whose variables in X are known, then being able to form the surface $X\hat{\beta}$ is a good first step (and perhaps a good last step). And, if one wishes to build a more elaborate regression model, allowing for an additional variable in W or explicitly accounting for multilevel variability by modeling the nested structure of the data, then regression results will be useful if the aim is merely to generate descriptive reductions of the data. But, if one wishes to know the value of Y that would result for any individual in the population if a variable in X were shifted from a value k to a value k', then regression results may be uninformative.

Many researchers (perhaps a clear majority) who use regression models in their research are very much interested in causal effects. Knowing the interests of their readers, many textbook presentations of regression sidestep these issues artfully by, for example, discussing how biased regression coefficients result from the omission of important explanatory variables but without introducing explicit, formal notions of causality into their presentations. Draper and Smith (1998:236), for example, write of the bias that enters into estimated regression coefficients when only a subset of the variables in the "true response relationship" are included in the fitted model. Similarly, Greene (2000:334) writes of the same form of bias that results from estimating coefficients for a subset of the variables from the "correctly specified regression model."[5] And, in his presentation of regression models for social scientists, Stolzenberg (2004:188) equivocates:

> Philosophical arguments about the nature of causation notwithstanding (see Holland, 1986), in most social science uses of regression, the *effect* of an independent variable on a dependent variable is the *rate* at which differences in the independent variable are associated with (or cause) differences or changes in the dependent variable. [Italics in the original.]

[5]There are, of course, other textbooks that do present a more complete perspective, such as Berk (2004), Freedman (2005), and Gelman and Hill (2007).

We also assume that the readers of our book are interested in causal effect estimators. And thus, although we recognize the classical regression tradition, perhaps best defended by Goldberger (1991) as interpretable merely as a descriptive data reduction tool, we will consider regression primarily as a causal effect estimator in the following sections of this chapter. And we further note that, in spite of our reference to Goldberger (1991), in other writing Goldberger has made it absolutely clear that he too was very much interested in the proper usage of regression models to offer warranted causal claims. This is perhaps most clear in work in which he criticized what he regarded as unwarranted causal claims generated by others using regression techniques, such as in his robust critique of Coleman's Catholic schools research that we summarized in Subsection 1.3.2 (see Goldberger and Cain 1982). We will return to a discussion of the notion of a correct specification of a regression model at the end of the chapter, where we discuss the connections between theoretical models and regressions as all-cause perfect specifications. Until then, however, we return to the same basic scenario considered in our presentation of matching in Chapter 4: the estimation of a single causal effect that may be confounded by other variables.

5.2 Regression Adjustment as a Strategy to Estimate Causal Effects

In this section, we consider the estimation of causal effects in which least squares regression is used to adjust for variables thought to be correlated with both the causal and the outcome variables. We first consider the textbook treatment of the concept of omitted-variable bias, with which most readers are probably well acquainted. Thereafter, we consider the same set of ideas after specifying the potential outcome variables that the counterfactual tradition assumes lie beneath the observed data.

5.2.1 Regression Models and Omitted-Variable Bias

Suppose that one is interested in estimating the causal effect of a binary variable D on an observed outcome Y. This goal can be motivated as an attempt to obtain an unbiased and consistent estimate of a coefficient δ in a generic bivariate regression equation:

$$Y = \alpha + \delta D + \varepsilon, \tag{5.3}$$

where α is an intercept and ε is a summary random variable that represents all other causes of Y (some of which may be related to the causal variable of interest, D). When Equation (5.3) is used to represent the causal effect of D on Y without any reference to individual-varying potential outcome variables, the parameter δ is implicitly cast as an invariant, structural causal effect that applies to all members of the population of interest.[6]

[6] Although this is generally the case, there are of course introductions to regression that explicitly define δ as the mean effect of D on Y across units in the population of interest or,

The OLS estimator of this bivariate regression coefficient is then:

$$\hat{\delta}_{\text{OLS, bivariate}} \equiv \frac{\text{Cov}_N(y_i, d_i)}{\text{Var}_N(d_i)}, \tag{5.4}$$

where $\text{Cov}_N(.)$ and $\text{Var}_N(.)$ are unbiased, sample-based estimates from a sample of size N of the population-level covariance and variance of the variables that are their arguments.[7] Because D is a binary variable, $\hat{\delta}_{\text{OLS, bivariate}}$ is exactly equivalent to the naive estimator, $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$, presented earlier in Equation (2.7) (i.e., sample mean of y_i for those in the treatment group minus the sample mean of y_i for those in the control group). Our analysis thus follows quite closely the prior discussion of the naive estimator in Subsection 2.6.3. The difference is that here we will develop the same basic claims with reference to the relationship between D and ε rather than the general implications of heterogeneity of the causal effect.

Consider first a case in which D is randomly assigned, as when individuals are randomly assigned to the treatment and control groups. In this case, D would be uncorrelated with ε in Equation (5.3), even though there may be a chance correlation between D and ε in any finite set of study subjects.[8] The literature on regression, when presented as a causal effect estimator, maintains that, in this case, (1) the estimator $\hat{\delta}_{\text{OLS, bivariate}}$ is unbiased and consistent for δ in Equation (5.3) and (2) δ can be interpreted as the causal effect of D on Y.

To understand this claim, it is best to consider a counterexample in which D is correlated with ε in the population because D is correlated with other causes of Y that are implicitly embedded in ε. For a familiar example, consider again the effect of education on earnings. Individuals are not randomly assigned to the treatment "completed a bachelor's degree." It is generally thought that those who complete college would be more likely to have had high levels of earnings

as was noted in the last section, without regard to causality at all.

[7]Notice that we are again focusing on the essential features of the methods, and thus we maintain our perfect measurement assumption (which allows us to avoid talking about measurement error in D or in Y, the latter of which would be embedded in ε). We also ignore degree-of-freedom adjustments because we assume that the available sample is again large. To be more precise, of course, we would want to indicate that the sample variance of D does not equal the population-level variance of D in the absence of such a degree-of-freedom adjustment, and so on. We merely label $\text{Var}_N(.)$ as signifying such an unbiased estimate of the population-level-variance of that which is its argument. Thus, $\text{Var}_N(.)$ implicitly includes the proper degree-of-freedom adjustment, which would be $N/(N-1)$ and which would then be multiplied by the average of squared deviations from the sample mean.

[8]We will frequently refer to D and ε as being uncorrelated for this type of assumption, as this is the semantics that most social scientists seem to use and understand when discussing these issues. Most textbook presentations of regression discuss very specific exogeneity assumptions for D that imply a correlation of 0 between D and ε. Usually, in the social sciences, the assumption is defined either by mean independence of D and ε or as an assumed covariance of 0 between D and ε. Both of these imply a correlation between D and ε of 0. In statistics, one often finds a stronger assumption: D and ε must be completely independent of each other. The argument in favor of this stronger assumption, which is convincing to statisticians, is that an inference is strongest when it holds under any transformation of Y (and thus any transformation of ε). When full independence of D and ε holds, mean independence of D and ε, a covariance of 0 between D and ε, and a 0 correlation between D and ε are all implied.

(a) A graph in which the causal effect of D on Y is unidentified

(b) An equivalent regression representation of the same graph

Figure 5.1: Graphs for a regression equation of the causal effect of D on Y.

in the absence of a college education. If this is true, D and the population-level error term ε are correlated because those who have a 1 on D are more likely to have high values rather than low values for ε. For this example, the least squares regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4) would not yield a consistent and unbiased estimate of δ that can be regarded as an unbiased and consistent estimate of the causal effect of D on Y. Instead, $\hat{\delta}_{\text{OLS,bivariate}}$ must be interpreted as an upwardly biased and inconsistent estimate of the causal effect of D on Y. In the substance of the college-degree example, $\hat{\delta}_{\text{OLS, bivariate}}$ would be a poor estimate of the causal effect of a college degree on earnings, as it would suggest that the effect of obtaining a college degree is larger than it really is.[9]

Figure 5.1 presents two causal graphs. In panel (a), D and Y are connected by two types of paths, the direct causal effect $D \rightarrow Y$ and an unspecified number of back-door paths signified by $D \dashleftarrow\!\dashrightarrow Y$. (Recall that bidirected edges $\dashleftarrow\!\dashrightarrow$ represent an unspecified number of common causes of the two variables that they connect.) For the graph in panel (a), the causal effect of D on Y is unidentified because no observable variables block the back-door paths represented by $D \dashleftarrow\!\dashrightarrow Y$.

The graph in panel (b) is the regression analog to the causal graph panel (a). It contains three edges: $D \rightarrow Y$, $\varepsilon \rightarrow Y$, and $D \dashleftarrow\!\dashrightarrow \varepsilon$, where the node for ε is represented by a hollow circle $\circ$ rather than a solid circle $\bullet$ in order to indicate that ε is an unobserved variable. The back-door paths from D to Y now run through the error term ε, and the dependence represented by the bidirected edge

[9] Consider for one last time the alternative and permissible descriptive interpretation: The least squares regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4) could be interpreted as an unbiased and consistent estimate of δ, in which the regression surface generated by the estimation of δ in Equation (5.3) can be interpreted as only a descriptively motivated, best linear prediction of the conditional expectation function, $E[Y|D]$ (i.e., where $\hat{\alpha}$ is an unbiased and consistent estimate of $E[Y|D = 0]$ and $\hat{\alpha} + \hat{\delta}$ *is* an unbiased and consistent estimate of $E[Y|D = 1]$). And, in the substance of the college-degree example, it could be regarded as an efficient estimate of the mean difference between the earnings of those who have obtained a college degree and those who have not. For this second type of interpretation, see the last section of this chapter.

contaminates the bivariate least squares regression coefficient for the regression of Y on D. Bivariate regression results, when interpreted as warranted causal effect estimates, assume that there are no unblocked back-door paths from the causal variable to the outcome variable.

For many applications in the social sciences, a correlation between D and ε is conceptualized as a problem of omitted variables. For the example in this section, a bivariate OLS estimate of the effect of a college degree on labor market earnings would be said to be biased because intelligence is unobserved but is correlated with both education and earnings. Its omission from Equation (5.3) leads the estimate of the effect of a college degree on earnings from that equation to be larger than it would have been if a variable for intelligence were instead included in the equation.

This perspective, however, has led to much confusion, especially in cases in which a correlation between D and ε emerges because subjects choose different levels of D based on their expectations about the variability of Y, and hence their own expectations of the causal effect itself. For example, those who attend college may be more likely to benefit from college than those who do not, even independent of the unobserved ability factor. Although this latent form of anticipation can be labeled an omitted variable, it is generally not. Instead, the language of research shifts toward notions such as self-selection bias, and this is less comfortable territory for the typical applied researcher.

To clarify the connections between omitted-variable bias and self-selection bias within a more general presentation, we draw on the counterfactual model in the next section. We break the error term in Equation (5.3) into component pieces defined by underlying potential outcome variables and allow for the more general forms of causal effect heterogeneity that are implicitly ruled out by constant-coefficient models.

5.2.2 Potential Outcomes and Omitted-Variable Bias

Consider the same set of ideas but now use the potential outcome framework to define the observable variables. We build directly on the variant of the counterfactual model presented in Subsection 3.2.2. From that presentation, recall Equation (3.5), which we reintroduce here as

$$Y = \mu^0 + (\mu^1 - \mu^0)D + \{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}, \tag{5.5}$$

where $\mu^0 \equiv E[Y^0]$, $\mu^1 \equiv E[Y^1]$, $\upsilon^0 \equiv Y^0 - E[Y^0]$, and $\upsilon^1 \equiv Y^1 - E[Y^1]$. We could rewrite this equation to bring it into closer alignment with Equation (5.3) by stipulating that $\alpha = \mu^0$, $\delta = (\mu^1 - \mu^0)$, and $\varepsilon = \upsilon^0 + D(\upsilon^1 - \upsilon^0)$. But note that this is not what is typically meant by the terms α, δ, and ε in Equation (5.3). The parameters α and δ in Equation (5.3) are not considered to be equal to $E[Y^0]$ or $E[\delta]$ for two reasons: (1) models are usually asserted in the regression tradition (e.g., in Draper and Smith 1998) without any reference to underlying causal states tied to potential outcomes and (2) the parameters α and δ are usually implicitly held to be constant structural effects that do not

vary over individuals in the population. Similarly, the error term, ε, in Equation (5.3) is not separated into two pieces as a function of the definition of potential outcomes and their relationship to D. For these reasons, Equation (5.5) is quite different from the traditional bivariate regression in Equation (5.3), in the sense that it is more finely articulated but also irretrievably tied to a particular formalization of a causal effect.

Suppose that we are interested in estimating the average treatment effect, denoted $(\mu^1 - \mu^0)$ here. D could be correlated with the population-level variant of the error term $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$ in Equation (5.5) in two ways. First, suppose that there is a net baseline difference in the hypothetical no-treatment state that is correlated with membership in the treatment group, but the size of the individual-level treatment effect does not differ on average between those in the treatment group and those in the control group. In this case, υ^0 would be correlated with D, generating a correlation between $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$ and D, even though the $D(\upsilon^1 - \upsilon^0)$ term in $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$ would be equal to zero on average because $\upsilon^1 - \upsilon^0$ does not vary with D. Second, suppose there is a net treatment effect difference that is correlated with membership in the treatment group, but there is no net baseline difference in the absence of treatment. Now, $D(\upsilon^1 - \upsilon^0)$ would be correlated with D, even though υ^0 is not, because the average difference in $\upsilon^1 - \upsilon^0$ varies across those in the treatment group and those in the control group. In either case, an OLS regression of the realized values of Y on D would yield a biased and inconsistent estimate of $(\mu^1 - \mu^0)$.

It may be helpful to see precisely how these sorts of bias come about with reference to the potential outcomes of individuals. Table 5.2 presents three simple two-person examples in which the least squares bivariate regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4) is biased. Each panel presents the potential outcome values for two individuals and then the implied observed data and error term in the braces from Equation (5.5). Assume for convenience that there are only two types of individuals in the population, both of which are homogeneous with respect to the outcomes under study and both of which comprise one half of the population. For the three examples in Table 5.2, we have sampled one of each of these two types of individuals for study.

For the example in the first panel, the true average treatment effect is 15, because for the individual in the treatment group δ_i is 10 whereas for the individual in the control group δ_i is 20. The values of υ_i^1 and υ_i^0 are deviations of the values of y_i^1 and y_i^0 from $E[Y^1]$ and $E[Y^0]$, respectively. Because these expectations are equal to 20 and 5, the values of υ_i^1 are both equal to 0 because each individual's value of y_i^1 is equal to 20. In contrast, the values of υ_i^0 are equal to 5 and -5 for the individuals in the treatment and control groups, respectively, because their two values of y_i^0 are 10 and 0.

As noted earlier, the bivariate regression estimate of the coefficient on D is equal to the naive estimator, $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$. Accordingly, a regression of the values for y_i on d_i would yield a value of 0 for the intercept and a value of 20 for the coefficient on D. This estimated value of 20 is an upwardly biased estimate for the true average causal effect because the values of d_i are positively correlated with the values of the error term $\upsilon_i^0 + d_i(\upsilon_i^1 - \upsilon_i^0)$. In this

Table 5.2: Examples of the Two Basic Forms of Bias for Least Squares Regression

			Differential baseline bias only				
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	0	5	20	1	0
In control group	20	0	0	−5	0	0	−5
			Differential treatment effect bias only				
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	2.5	0	20	1	2.5
In control group	15	10	−2.5	0	10	0	0
			Both types of bias				
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	25	5	5	−2.5	25	1	5
In control group	15	10	−5	2.5	10	0	2.5

case, the individual with a value of 1 for d_i has a value of 0 for the error term whereas the individual with a value of 0 for d_i has a value of −5 for the error term.

For the example in the second panel, the relevant difference between the individual in the treatment group and the individual in the control group is in the value of y_i^1 rather than y_i^0. In this variant, both individuals would have had the same outcome if they were both in the control state, but the individual in the treatment group would benefit relatively more from being in the treatment state. Consequently, the values of d_i are correlated with the values of the error term in the last column because the true treatment effect is larger for the individual in the treatment group than for the individual in the control group. A bivariate regression would yield an estimate of 10 for the average causal effect, even though the true average causal effect is only 7.5 in this case.

Finally, in the third panel of the table, both forms of baseline and net treatment effect bias are present, and in opposite directions. In combination, however, they still generate a positive correlation between the values of d_i and the error term in the last column. This pattern results in a bivariate regression estimate of 15, which is upwardly biased for the true average causal effect of 12.5.

For symmetry, and some additional insight, now consider two additional two-person examples in which regression gives an unbiased estimate of the average causal effect. For the first panel of Table 5.3, the potential outcomes are independent of D, and as a result a bivariate regression of the values y_i on d_i would

Table 5.3: Two-Person Examples in Which Least Squares Regression Estimates are Unbiased

	Independence of (Y^1, Y^0) from D						
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	0	0	20	1	0
In control group	20	10	0	0	10	0	0
	Offsetting dependence of $Y^1$and Y^0 on D						
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	5	−5	20	1	5
In control group	10	20	−5	5	20	0	5

yield an unbiased estimate of 10 for the true average causal effect. But the example in the second panel is quite different. Here, the values of v_i^1 and v_i^0 are each correlated with the values of d_i, but they cancel each other out when they jointly constitute the error term in the final column. Thus, a bivariate regression yields an unbiased estimate of 0 for the true average causal effect of 0. And, yet, with knowledge of the values for y_i^1 and y_i^0, it is clear that these results mask important heterogeneity of the causal effect. Even though the average causal effect is indeed 0, the individual-level causal effects are equal to 10 and −10 for the individuals in the treatment group and control group, respectively. Thus, regression gives the right answer, but it hides the underlying heterogeneity that one would almost certainly wish to know.

Having considered these examples, we are now in a position to answer, from within the counterfactual framework, the question that so often confounds students when first introduced to regression as a causal effect estimator: What is the error term of a regression equation? Compare the third and fourth columns with the final column in Tables 5.2 and 5.3. The regression error term, $v^0 + D(v^1 - v^0)$, is equal to v^0 for those in the control group and v^1 for those in the treatment group. This can be seen without reference to the examples in the tables. Simply rearrange $v^0 + D(v^1 - v^0)$ as $Dv^1 + (1 - D)v^0$ and then rewrite Equation (5.5) as

$$Y = \mu^0 + (\mu^1 - \mu^0)D + \{Dv^1 + (1 - D)v^0\}. \tag{5.6}$$

It should be clear that the error term now appears very much like the observability of Y definition presented earlier as $DY^1 + (1-D)Y^0$ in Equation (2.2). Just as Y switches between Y^1 and Y^0 as a function of D, the error term switches between v^1 and v^0 as a function of D. Given that v^1 and v^0 can be interpreted as Y^1 and Y^0 centered around their respective population-level expectations $E[Y^1]$ and $E[Y^0]$, this should not be surprising.

Even so, few presentations of regression characterize the error term of a bivariate regression in this way. Some notable exceptions do exist. The connection is made to the counterfactual tradition by specifying Equation (5.3) as

$$Y = \alpha + \delta D + \varepsilon_{(D)}, \tag{5.7}$$

where the error term $\varepsilon_{(D)}$ is considered to be an entirely different random variable for each value of D (see Pratt and Schlaifer 1988). Consequently, the error term ε in Equation (5.3) switches between $\varepsilon_{(1)}$ and $\varepsilon_{(0)}$ in Equation (5.7) depending on whether D is equal to 1 or 0.[10]

Before moving on to adjustment techniques, it seems proper to ask one final question. If both υ^1 and υ^0 are uncorrelated with D, will the bivariate least squares regression coefficient for D be an unbiased and consistent estimate of the average causal effect? Yes, but two qualifications should be noted, both of which were revealed in the second example in Table 5.3. First, bivariate regression can yield an unbiased and consistent estimate in other cases, as when the nonzero correlations that υ^1 and υ^0 have with D "cancel out" in the construction of the combined error term $D\upsilon^1 + (1 - D)\upsilon^0$. Second, an unbiased and consistent regression estimate of the average causal effect may still mask important heterogeneity of causal effects. The first of these qualifications would rarely apply to real-world applications, but the second qualification, we suspect, obtains widely and is less frequently recognized than it should be.

5.2.3 Regression as Adjustment for Otherwise Omitted Variables

How well can regression adjust for an omitted variable if that variable is observed and included in an expanded regression equation? The basic strategy behind regression analysis as an adjustment technique to estimate a causal effect is to add a sufficient set of "control variables" to the bivariate regression in Equation (5.3) in order to break a correlation between the treatment variable D and the error term ε, as in

$$Y = \alpha + \delta D + X\beta + \varepsilon^*, \tag{5.8}$$

where X represents one or more variables, β is a coefficient (or a conformable vector of coefficients if X represents more than one variable), ε^* is a residualized version of the original error term ε from Equation (5.3), and all else is as defined for Equation (5.3).

For the multiple regression analog to the least squares bivariate regression estimator $\hat{\delta}_{\text{OLS, bivariate}}$ in Equation (5.4), the observed data values d_i and x_i are embedded in an all-encompassing $\mathbf{Q}$ matrix, which is $N \times K$, where N is the number of respondents and K is the number of variables in X plus 2 (one

[10] This is the same approach taken by Freedman (see Berk 2004, Freedman 2005), and he refers to Equation (5.7) as a response schedule. See also the discussion of Sobel (1995). For a continuous variable, Garen (1984) notes that there would be an infinite number of error terms (see discussion of Garen's Equation 10).

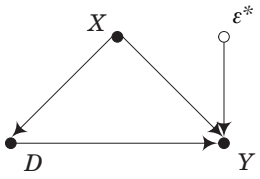


Figure 5.2: A causal graph for a regression equation in which the causal effect of D on Y is identified by conditioning on X.

for the constant and one for the treatment variable D). The OLS estimator for the parameters in Equation (5.8) is then written in matrix notation as

$$\hat{\delta}_{\text{OLS, multiple}} \equiv (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{y}, \tag{5.9}$$

where $\mathbf{y}$ is an $N \times 1$ vector for the observed outcomes y_i. As all regression textbooks show, there is nothing magical about these least squares computations, even though the matrix representation may appear unfamiliar to some readers. OLS regression is equivalent to the following three-step regression procedure with reference to Equation (5.8) [and without reference to the perhaps overly compact Equation (5.9)]: (1) Regress y_i on x_i and calculate $y_i^* = y_i - \hat{y}_i$; (2) regress d_i on x_i and calculate $d_i^* = d_i - \hat{d}_i$; (3) regress y_i^* on d_i^*. The regression coefficient on d_i^* yielded by step (3) is the OLS estimate of δ, which is typically declared unbiased and consistent for δ in Equation (5.8) if the true correlation between D and ε^* is assumed to be equal to zero. Thus, in this simple example, OLS regression is equivalent to estimating the relationship between residualized versions of Y and D from which their common dependence on other variables in X has been "subtracted out."

Even though the variables in X might be labeled control variables in a regression analysis of a causal effect, this label expresses the intent rather than the outcome of their utilization. The goal of such a regression adjustment strategy is to find variables in X that can be used to redraw the causal graph in panel (b) of Figure 5.1 as the DAG in Figure 5.2. If this can be done, then one can condition on X in order to consistently estimate the causal effect of D on Y because X blocks the only back-door path between D and Y.

If D is uncorrelated with ε^* (i.e., the error term net of adjustment for X), then least squares regression yields an estimate that is ostensibly freed of the bias generated by the correlation of the treatment D with the error term ε in Equation (5.3). However, even in this case some complications remain when one invokes the potential outcome model.

First, if one assumes that δ is truly constant across individuals (i.e., that $y_i^1 - y_i^0$ is equal to the same constant for all individuals i), then the OLS estimate is unbiased and consistent for δ and for $(\mu^1 - \mu^0)$. If, however, $y_i^1 - y_i^0$ is not constant, then the OLS estimate represents a conditional-variance-weighted estimate of the underlying causal effects of individuals, δ_i, in which the weights are a function of the conditional variance of D (see Angrist 1998, as well as our

explanation of this result in the next section). Under these conditions, the OLS estimate is unbiased and consistent for this particular weighted average, which is usually not a causal parameter of interest.

Second, note that the residualized error term, ε^*, in Equation (5.8) is not equivalent to either ε from Equation (5.3) or to the multipart error term $\{\upsilon^0 + D(\upsilon^1 - \upsilon^0)\}$ from Equation (5.5). Rather, it is defined by whatever adjustment occurs within Equation (5.8), as represented by the term $X\beta$. Consequently, the residualized error term ε^* cannot be interpreted independently of decisions about how to specify the vector of adjustment variables in X, and this can make it difficult to define when a net covariance between D and ε^* can be assumed to be zero.

We explain these two complications and their important implications in the following sections of this chapter, where we consider a variety of examples that demonstrate the connections between matching and regression estimators of causal effects. Before developing these explanations, however, we conclude this section with two final small-N examples that demonstrate how the regression adjustment strategy does and does not work.

Table 5.4 presents two six-person examples. For both examples, a regression of Y on D yields a biased estimate of the true average treatment effect. And, in fact, both examples yield the same biased estimate because the observed values y_i and d_i are the same for both examples. Moreover, an adjustment variable X is also available for both examples, and its observed values x_i have the same associations with the observed values y_i and d_i for both examples. But the underlying potential outcomes differ substantially between the two examples. These differences render regression adjustment by X effective for only the first example.

For the example in the first panel, a regression of Y on D would yield an estimate of the coefficient for D of 11.67, which is an upwardly biased estimate of the true average causal effect of 10. The bias arises because the correlation between the error term in the last column and the realized values for d_i is not zero but is instead .33.

For the example in the second panel, a regression of Y on D would yield an estimate of the coefficient for D of 11.67 because the values for y_i and d_i are exactly the same as for the example in the first panel. Moreover, this estimate is also upwardly biased because the error term in the last column is positively correlated with the realized values of d_i. However, here the patterns are more complex. The underlying potential outcomes are different, and individual-level heterogeneity of the causal effect is now present. One member of the control group has an individual-level treatment effect of only 8, and as a result the true average treatment effect is only 9.67. Consequently, the same bivariate regression coefficient of 11.67 has a larger upward bias in this second example, and the correlation between the values of d_i and the error term in the last column is now .39 rather than .33.[11]

[11] Moreover, the correlation between the values of d_i and both υ_i^1 and υ_i^0 differs, with the former generating a correlation coefficient of .51 and the latter generating a correlation coefficient of .33.

Table 5.4: Two Six-Person Examples in Which Regression Adjustment is Differentially Effective

					Regression adjustment with X generates an unbiased estimate for D			
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	x_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	2.5	2.5	20	1	1	2.5
In treatment group	20	10	2.5	2.5	20	1	1	2.5
In treatment group	15	5	−2.5	−2.5	15	1	0	−2.5
In control group	20	10	2.5	2.5	10	0	1	2.5
In control group	15	5	−2.5	−2.5	5	0	0	−2.5
In control group	15	5	−2.5	−2.5	5	0	0	−2.5

					Regression adjustment with X does not generate an unbiased estimate for D			
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	x_i	$v_i^0+d_i(v_i^1-v_i^0)$
In treatment group	20	10	2.83	2.5	20	1	1	2.83
In treatment group	20	10	2.83	2.5	20	1	1	2.83
In treatment group	15	5	−2.17	−2.5	15	1	0	−2.17
In control group	18	10	.83	2.5	10	0	1	2.5
In control group	15	5	−2.17	−2.5	5	0	0	−2.5
In control group	15	5	−2.17	−2.5	5	0	0	−2.5

This underlying difference in potential outcomes also has consequences for the capacity of regression adjustment to effectively generate unbiased estimates of the average treatment effect. This is easiest to see by rearranging the rows in Table 5.4 for each of the two examples based on the values of X for each individual, as in Table 5.5. For the first example, the values of d_i are uncorrelated with the error term within subsets of individuals defined by the two values of X. In contrast, for the second example, the values of d_i remain positively correlated with the error term within subsets of individuals defined by the two values of X. Thus, conditioning on X breaks the correlation between D and the error term in the first example but not in the second example. Because the observed data are the same for both examples, this difference is entirely a function of the underlying potential outcomes that generate the data.

This example demonstrates an important conceptual point. Recall that the basic strategy behind regression analysis as an adjustment technique is to estimate

$$Y = \alpha + \delta D + X\beta + \varepsilon^*,$$

Table 5.5: A Rearrangement to Show How Regression Adjustment is Differentially Effective

					Regression adjustment with X generates an unbiased estimate for D			
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	x_i	$v_i^0+d_i(v_i^1-v_i^0)$
					For those with $X=1$			
In treatment group	20	10	2.5	2.5	20	1	1	2.5
In treatment group	20	10	2.5	2.5	20	1	1	2.5
In control group	20	10	2.5	2.5	10	0	1	2.5
					For those with $X=0$			
In treatment group	15	5	−2.5	−2.5	15	1	0	−2.5
In control group	15	5	−2.5	−2.5	5	0	0	−2.5
In control group	15	5	−2.5	−2.5	5	0	0	−2.5
					Regression adjustment with X does not generate an unbiased estimate for D			
	y_i^1	y_i^0	v_i^1	v_i^0	y_i	d_i	x_i	$v_i^0+d_i(v_i^1-v_i^0)$
					For those with $X=1$			
In treatment group	20	10	2.83	2.5	20	1	1	2.83
In treatment group	20	10	2.83	2.5	20	1	1	2.83
In control group	18	10	.83	2.5	10	0	1	2.5
					For those with $X=0$			
In treatment group	15	5	−2.17	−2.5	15	1	0	−2.17
In control group	15	5	−2.17	−2.5	5	0	0	−2.5
In control group	15	5	−2.17	−2.5	5	0	0	−2.5

where X represents one or more control variables, β is a coefficient (or a conformable vector of coefficients if X represents more than one variable), and ε^* is a residualized version of the original error term ε from Equation (5.3) [see our earlier presentation of Equation (5.8)]. The literature on regression often states that an estimated coefficient $\hat{\delta}$ from this regression equation is unbiased and consistent for the average causal effect if ε^* is uncorrelated with D. But, because the specific definition of ε^* is conditional on the specification of X, many researchers find this requirement of a zero correlation difficult to interpret and hence difficult to evaluate.

The crux of the idea, however, can be understood without reference to the error term ε^* but rather with reference to the simpler (and, as we have argued earlier) more clearly defined error term $v^0 + D(v^1 - v^0)$ from Equation (5.5) [or, equivalently, $Dv^1 + (1-D)v^0$ from Equation (5.6)]. Regression adjustment

by X in Equation (5.8) will yield an unbiased and consistent estimate of the average causal effect of D when

1. D is uncorrelated with $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$ for each subset of respondents identified by distinct values on the variables in X,
2. the causal effect of D does not vary with X, and
3. a fully flexible parameterization of X is used.[12]

Consider the relationship between this set of conditions and what was described earlier in Subsection 3.2.1 as an assumption that treatment assignment is ignorable. Switching notation from S to X in Equation (3.3) results in

$$(Y^0, Y^1) \perp\!\!\!\perp D \mid X, \tag{5.10}$$

where, again, the symbol $\perp\!\!\!\perp$ denotes independence. Now, rewrite the assumption, deviating Y^0 and Y^1 from their population-level expectations:

$$(\upsilon^0, \upsilon^1) \perp\!\!\!\perp D \mid X. \tag{5.11}$$

This switch from (Y^0, Y^1) to (υ^0, υ^1) does not change the assumption, at least insofar as is relevant here (because we have defined the individual-level causal effect as a linear difference, because the expectation operator is linear, and because $E[Y^0]$ and $E[Y^1]$ do not depend on who is in the treatment state and who is in the control state). Consequently, ignorability of treatment assignment can be defined only with respect to individual-level departures from the true average potential outcomes across all members of the population under the assumptions already introduced.

Given that an assumption of ignorable treatment assignment can be written as Equation (5.11), the connections between this assumption and the set of conditions that justify a regression estimator as unbiased and consistent for the effect of D on Y should now be clear. If treatment assignment is ignorable as defined in Equation (5.11), then a regression equation that conditions fully on all values of X by including a fully flexible coding of X as a set of dummy variables will yield unbiased and consistent regression estimates of the average causal effect of D on Y. Even so, ignorability is not equivalent to the set of conditions just laid out. Instead, υ^0 and υ^1 [as well as functions of them, such as $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$] must only be mean independent of D conditional on X, not fully independent of D conditional on X.

Stepping back from this correspondence, we should note that this is not the only set of conditions that would establish least squares estimation unbiased and consistent for the average causal effect, but it is the most common

[12] Here again, we use the word uncorrelated to characterize the necessary association between D and $\upsilon^0 + D\,(\upsilon^1 - \upsilon^0)$. More formally, it would be best to state that D and $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$ must be mean independent, so that a 0 covariance of D and $\upsilon^0 + D(\upsilon^1 - \upsilon^0)$ is implied.

set of conditions that would apply in most research situations.[13] Our point in laying it out is not to provide a rigid guideline applicable to all types of regression models but instead to show why the earlier statement that "ε^* must be uncorrelated with D" is insufficiently articulated from a counterfactual perspective.

A larger point of this section, however, is that much of the received wisdom on regression modeling breaks down in the presence of individual-level heterogeneity of a causal effect, as would be present in general when causal effects are defined with reference to underlying potential outcomes tied to well-defined causal states. In the next section, we begin to explain these complications more systematically, starting from the assumption, as in prior chapters, that causal effects are inherently heterogeneous and likely to vary systematically between those in the treatment and control groups. We then present the connections among regression, matching, and stratification, building directly on our presentation of matching as conditioning by stratification in Chapter 4.

5.3 The Connections Between Regression and Matching

In this section, we return to the demonstrations utilized to motivate matching estimators in Chapter 4. Our goal is to establish when matching and regression yield different results, even though a researcher is attempting to adjust for the same set of variables. We then show how regression estimators can be reformulated to yield the same results as matching estimators – as a full parameterization of a perfect stratification of the data and then as weighted least squares estimators in which the weights are a function of the propensity score. In these cases, we show that regression is an effective estimator of causal effects defined by potential outcomes.

5.3.1 Regression as Conditional-Variance-Weighted Matching

We first show why least squares regression can yield misleading causal effect estimates in the presence of individual-level heterogeneity of causal effects, even though the only variable that needs to be adjusted for is given a fully flexible coding (i.e., when the adjustment variable is parameterized with a dummy variable for each of its values, save one for the reference category).[14] When

[13] For example, the second condition can be dropped if the heterogeneity of the causal effect is modeled as a function of X (i.e., the parameterization is fully saturated in both D and X). In this case, however, regression then becomes a way of enacting a stratification of the data, as for the matching techniques presented in the last chapter.

[14] When we write of a fully flexible coding of a variable, we are referring to a dummy variable coding of that variable only. As we will discuss later, a saturated model entails a fully flexible coding of each variable *as well as all interactions between them.* For the models discussed