Chapter 9

Instrumental Variable Estimators of Causal Effects

If a perfect stratification cannot be enacted with the available data, and thus neither matching nor regression nor any other type of basic conditioning technique can be used to effectively estimate a causal effect of D on Y, one possible solution is to find an exogenous source of variation that determines Y only by way of the causal variable D. The causal effect is then estimated by measuring how Y varies with the portion of the total variation in D that is attributable to the exogenous variation. The variable that indexes this variation in D is an instrumental variable (IV).

In this chapter, we orient the reader by first presenting IV estimation with a binary instrument and a simple demonstration. We then return to the origins of IV techniques, and we contrast this estimation strategy with the perspective on regression that was presented in Chapter 6. We then develop the same ideas using the potential outcome model, showing how the counterfactual perspective has led to a new literature on how to interpret IV estimates. This new literature suggests that IV techniques are more effective for estimating narrowly defined local average causal effects than for estimating more general average causal effects, such as the average treatment effect (ATE) and the average treatment effect for the treated (ATT). We also consider causal graphs that can represent this recent literature and conclude with a discussion of marginal treatment effects identified by local IVs.

9.1 Causal Effect Estimation with a Binary IV

We begin our presentation with the simplest scenario in which an instrumental variable can be used to effectively estimate a causal effect. Recall the causal regression setup in Equation (6.3):

$$Y = \alpha + \delta D + \varepsilon, \tag{9.1}$$

where Y is the outcome variable, D is a binary causal exposure variable, α is an intercept, δ is the causal effect of D on Y, and ε is a summary random variable

(a) Z is a valid instrumental variable for D

(b) Z is not a valid instrumental variable for D

Figure 9.1 Two graphs in which Z is a potential instrumental variable.

that represents all other causes of Y. As noted above, when Equation (9.1) is used to represent the causal effect of D on Y, the parameter δ is usually considered an invariant, structural causal effect that applies to all members of the population of interest. We will maintain this traditional assumption in this section.[1]

Suppose that the probability that D is equal to 1 rather than 0 is a function of a binary variable Z that takes on values of 0 and 1. Figure 9.1 presents two possible ways in which the variable Z could be related to both D and Y. Note first that, for both graphs, the back-door paths represented by $D \dashleftarrow\dashrightarrow \varepsilon \rightarrow Y$ prevent a least squares regression of Y on D from generating a consistent or unbiased estimate of the effect of D on Y. More generally, no conditioning estimator would effectively estimate the causal effect of D on Y for these graphs because no observed variables satisfy the back-door criterion.

Now consider how Z, which we have labeled a "potential instrumental variable," is related to the outcome variable Y according to the alternative structures of these two graphs. For both Figures 9.1(a) and 9.1(b), Z has an association with Y because of the directed path $Z \rightarrow D \rightarrow Y$. In addition, the paths collectively represented by $Z \rightarrow D \dashleftarrow\dashrightarrow \varepsilon \rightarrow Y$ do not contribute to the association between Z and Y because D is a collider variable along all of them. However, for Figure 9.1(b), the additional paths represented by $Z \dashleftarrow\dashrightarrow \varepsilon \rightarrow Y$ contribute to the association between Z and Y because of the common causes that determine both Z and ε.[2] This last difference is the reason that Z can be used to estimate the causal effect of D on Y for Figure 9.1(a) but not for Figure 9.1(b), as we will now explain.

If we continue to maintain the assumption that the effect of D on Y is a constant structural effect δ, then it is not necessary to relate all of the variation in D to all of the variation in Y in order to obtain a consistent estimate of the causal effect. Under this assumption, if we can find a way of isolating the covariation in D and Y that is causal, then we can ignore the other covariation in D and Y that is noncausal because it is generated by the common causes of D and ε. For Figure 9.1(a), the variable Z represents an isolated source of variation across which the analyst can examine the covariation in D and Y that is causal. In this sense, Z is an "instrument" for

[1] In Pearl's framework, we are assuming that $f_Y(D, e_Y)$ is defined as the linear function $\alpha + \delta D + \varepsilon$ and where α and β are scalar constants.

[2] We have drawn the two bidirected edges separately for Figure 9.1(b) for simplicity. The same reasoning holds when some of the common causes of Z and ε are also causes of D (and so on).

examining an isolated slice of the covariation in D and Y. For Figure 9.1(b), Z does not provide an isolated source of variation. The analyst can still examine the portion of the covariation in D and Y that can be calculated across levels of Z, but some of this covariation must be noncausal because D and Y are dependent on common causes embedded in the bidirected edge in the back-door paths collectively represented by $D \leftarrow Z \dashleftarrow\dashrightarrow \varepsilon \rightarrow Y$.

To see this result without reference to graphs, take the population-level expectation of Equation (9.1), $E[Y] = E[\alpha + \delta D + \varepsilon] = \alpha + \delta E[D] + E[\varepsilon]$, and rewrite it as a difference equation in Z:

$$\begin{aligned} &E[Y|Z=1] - E[Y|Z=0] \\ &\quad = \delta(E[D|Z=1] - E[D|Z=0]) + (E[\varepsilon|Z=1] - E[\varepsilon|Z=0]). \end{aligned} \tag{9.2}$$

Equation (9.2) is now focused narrowly on the variation in Y, D, and ε that exists across levels of Z.[3] Now, take Equation (9.2) and divide both sides by $E[D|Z=1] - E[D|Z=0]$, yielding

$$\begin{aligned} &\frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} \\ &\quad = \frac{\delta(E[D|Z=1] - E[D|Z=0]) + (E[\varepsilon|Z=1] - E[\varepsilon|Z=0])}{E[D|Z=1] - E[D|Z=0]}. \end{aligned} \tag{9.3}$$

If the data are generated by the set of causal relationships depicted in Figure 9.1(a), then Z has no linear association with ε, and $E[\varepsilon|Z=1] - E[\varepsilon|Z=0]$ in Equation (9.3) is equal to 0. Consequently, the right-hand side of Equation (9.3) simplifies to δ:

$$\frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]} = \delta. \tag{9.4}$$

Under these conditions, the ratio of the population-level linear association between Y and Z and between D and Z is equal to the causal effect of D on Y. This result suggests that, if Z is in fact associated with D but not associated with ε (or with Y, except through D), then the following sample-based estimator will equal δ in an infinite sample:

$$\hat{\delta}_{\text{IV,WALD}} \equiv \frac{E_N[y_i|z_i=1] - E_N[y_i|z_i=0]}{E_N[d_i|z_i=1] - E_N[d_i|z_i=0]}. \tag{9.5}$$

As suggested by its subscript, this is the IV estimator, which is known as the Wald estimator when the instrument is binary. Although the Wald estimator is consistent for δ in this scenario, the assumption that δ is an invariant structural effect is crucial for this result.[4] From a potential outcomes perspective, in which we generally assume that

[3] Equation (9.2) is generated in the following way. First, write $E[Y] = \alpha + \delta E[D] + E[\varepsilon]$ conditional on the two values of Z, yielding $E[Y|Z=1] = \alpha + \delta E[D|Z=1] + E[\varepsilon|Z=1]$ and $E[Y|Z=0] = \alpha + \delta E[D|Z=0] + E[\varepsilon|Z=0]$. Note that, because α and δ are considered constant structural effects for this traditional motivation of IV estimation, they do not vary with Z. Now, subtract $E[Y|Z=0] = \alpha + \delta E[D|Z=0] + E[\varepsilon|Z=0]$ from $E[Y|Z=1] = \alpha + \delta E[D|Z=1] + E[\varepsilon|Z=1]$. The parameter α is eliminated by the subtraction, and δ can be factored out of its two terms, resulting in Equation (9.2).

[4] As for the origin of the Wald estimator, it is customarily traced to Wald (1940) by authors such as Angrist and Krueger (1999). As we discuss later, the Wald estimator is not generally unbiased in a finite sample and instead is only consistent.

causal effects vary meaningfully across individuals, this assumption is very limiting and quite likely unreasonable. Explaining when and how this assumption can be relaxed is one of the main goals of this chapter.

For completeness, return to consideration of Figure 9.1(b), in which Z has a nonzero association with ε. In this case, $E[\varepsilon|Z=1]-E[\varepsilon|Z=0]$ in Equations (9.2) and (9.3) cannot be equal to 0, and thus Equation (9.3) does not reduce further to Equation (9.4). Rather, it reduces only to

$$\frac{E[Y|Z=1]-E[Y|Z=0]}{E[D|Z=1]-E[D|Z=0]}=\delta+\frac{E[\varepsilon|Z=1]-E[\varepsilon|Z=0]}{E[D|Z=1]-E[D|Z=0]}\ . \tag{9.6}$$

In this case, the ratio of the population-level linear association between Y and Z and between D and Z does not equal the causal effect of D on Y but rather the causal effect of D on Y plus the last term on the right-hand side of Equation (9.6). The Wald estimator in Equation (9.5) is not consistent for δ in this case. Instead, it converges to the right-hand side of Equation (9.6), which is equal to δ plus a bias term that is a function of the net association between Z and ε.

More generally, an IV estimator is a ratio that is a joint projection of Y and D onto a third dimension Z. In this sense, an IV estimator isolates a specific portion of the covariation in D and Y. For that selected covariation to be the basis of a valid causal inference for the effect of D on Y, it cannot be attributable to any extraneous common causes that determine both Z and Y. And, to justify the causal effect estimate generated by a subset of the covariation in D and Y as a consistent estimate of the population-level causal effect of D on Y, it is typically assumed in this tradition that δ is a constant for all members of the population.[5] Consider the following hypothetical demonstration, which is based on the school voucher example introduced in Section 1.3.2 (see page 23).

IV Demonstration 1

Suppose that a state education department wishes to determine whether private high schools outperform public high schools in a given metropolitan area, as measured by the achievement of ninth graders on a standardized test. For context, suppose that a school voucher program is operating in the city and that the state is considering whether to introduce the program in other areas in order to shift students out of public schools and into private schools.

To answer this policy question, the state department of education uses a census of the population in the metropolitan area to select a random sample of 10,000 ninth graders. They then give a standardized test to each sampled ninth grader at the end of the year, and they collect data as $\{y_i, d_i\}_{i=1}^{10{,}000}$, where Y is the score on the standardized

[5] There is one trivial way around this assumption. If the naive estimator of D on Y is consistent for the average causal effect of D on Y, then D is its own IV. The subset of the covariation in D and Y that is projected onto Z is then the full set of covariation in D and Y because Z is equal to D. In this case, no extrapolation is needed and the constant treatment effect assumption can be avoided. As we will discuss later, there are slightly less trivial ways to avoid the assumption as well. One alternative is to assert that δ is a constant among the treated and then stipulate instead that the IV identifies only the ATT. Although plausible, there are better ways to handle heterogeneity, as we will discuss as this chapter unfolds.

Table 9.1 The Distribution of Voucher Winners by School Sector for IV Demonstration 1

		Public school $d_i = 0$	Private school $d_i = 1$
Voucher loser	$z_i = 0$	8000	1000
Voucher winner	$z_i = 1$	800	200

test and D is equal to 1 for students who attend private high schools and equal to 0 for students who attend public high schools.

After the data are collected, suppose that the values of y_i are regressed on the values of d_i and that a predicted regression surface is obtained:

$$\hat{Y} = 50.0 + 9.67\,(D)\,. \tag{9.7}$$

The state officials recognize that private school students typically have more highly educated parents and therefore are more likely to have higher test scores no matter what curriculum and school culture they have been exposed to. Accordingly, they surmise that 9.67 is likely a poor causal effect estimate, or at least not one that they would want to defend in public as equal to the ATE.

The state officials therefore decide to merge the data with administrative records on the school voucher program in the area. For this program, all eighth graders in the city (both in public and private schools) are entered into a random lottery for $3,000 school vouchers that are redeemable at a private high school. By mandate, 10 percent of all eligible students win the voucher lottery.

After merging the data, the state officials cross-tabulate eighth grade lottery winners (where $z_i = 1$ for those who won the lottery and $z_i = 0$ for those who did not) by school sector attendance in the ninth grade, d_i. As shown in Table 9.1, 1,000 of the 10,000 sampled students were voucher lottery winners.[6] Of these 1,000 students, 200 were observed in private schools in the ninth grade. In comparison, of the 9,000 sampled students who were not voucher lottery winners, only 1,000 were observed in private schools.

The researchers assume that the dummy variable Z for winning the voucher lottery is a valid IV for D because they believe that (1) the randomization of the lottery renders Z independent of ε in the population-level causal regression equation $Y = \alpha + \delta D + \varepsilon$ and that (2) Z has a causal effect on Y only through Z.. They therefore estimate $\hat{\delta}_{\text{IV,WALD}}$ in Equation (9.5) as

$$\frac{E_N[y_i|z_i = 1] - E_N[y_i|z_i = 0]}{E_N[d_i|z_i = 1] - E_N[d_i|z_i = 0]} = \frac{51.600 - 51.111}{.200 - .111} = 5.5, \tag{9.8}$$

and conclude that the true causal effect of private schooling on ninth grade achievement is 5.5 rather than 9.67. Operationally, the Wald estimator takes the average

[6] For simplicity, we have assumed that the sample percentage of lottery winners is the same as the population percentage. Of course, some random variation would be expected in any sample.

difference in test scores among those students who have won a voucher and those who have not won a voucher and divides that difference by a corresponding difference in the proportion of high school students who attend private schools among those who have won a voucher and the proportion of high school students who attend private schools among those who have not won a voucher. The numerator of Equation (9.8) is equal to $51.600 - 51.111$ by an assumed construction of the outcome Y, which we will present later when we repeat this demonstration in more detail in Section 9.3.1 (as IV Demonstration 2, beginning on page 309). The denominator, however, can be calculated directly from Table 9.1. In particular, $E_N[d_i = 1|z_i = 1] = 200/1000 = .200$, whereas $E_N[d_i = 1|z_i = 0] = 1000/9000 = .111$.

For this demonstration, Z is a valid instrument by the traditional assumptions maintained in this section of the chapter; it is randomly assigned to students and has no association with Y except for the one produced by the directed path $Z \rightarrow D \rightarrow Y$. The resulting estimator yields a point estimate that can be given a traditional causal interpretation based on the position that the casual effect of interest is a constant structural effect. However, as we will show later in this chapter when we reintroduce this demonstration as IV Demonstration 2, the particular causal effect that this IV identifies is quite a bit different when individual-level causal effect heterogeneity is not assumed away. The IV does not identify the ATE or the ATT. Instead, it identifies only the average causal effect for the subset of all students who would attend a private school if given a voucher but who would not attend a private school in the absence of a voucher. This means, for example, that the IV estimate is uninformative about the average causal effect among those who would enroll in private high schools in the absence of a voucher. This group of students represents the vast majority of private school students (in this example $1,000/1,200$ or 83 percent). The potential outcome literature has provided the insight that allows such a precise causal interpretation and clarifies what a valid IV estimate does not inform. Before presenting that newer material, we return to a more complete accounting of the traditional IV literature.

9.2 Traditional IV Estimators

As detailed by Goldberger (1972), Bowden and Turkington (1984), Heckman (2000), and Bollen (2012), IV estimators were developed first in the 1920s by biologists and economists analyzing equilibrium price determination in market exchange (see E. Working 1927; H. Working 1925; and Wright 1921, 1925). After subsequent development in the 1930s and 1940s (e.g., Wright 1934; Schultz 1938; Reiersl 1941; and Haavelmo 1943), IV estimators were brought into widespread use in economics by researchers associated with the Cowles commission (see Hood and Koopmans 1953; Koopmans and Reiersl 1950).[7] The structural equation tradition in sociology shares

[7]The canonical example for this early development was the estimation of price determination in markets. For a market in equilibrium, only one price and one quantity of goods sold is observable at any point in time. To make prospective predictions about the potential effects of exogenous supply-and-demand shocks on prices and quantities for a new market equilibrium, the shape of latent supply-and-demand curves must be determined. To estimate points on such curves, separate variables are

similar origins to that of the IV literature (see Duncan 1975). The most familiar deployment of IV estimation in the extant sociological research is as the order condition for identification of a system of structural equations (see Bollen 1989, 1995, 1996a, 1996b, 2001, 2012; Fox 1984).

9.2.1 The Linear Structural Equation IV Estimator

Consider the same basic ideas presented earlier for the Wald estimator in Equation (9.5). Again, recall the causal regression setup in Equation (9.1):

$$Y = \alpha + \delta D + \varepsilon, \tag{9.9}$$

and again assume that we are in the traditional setup where δ is an invariant, structural causal effect that applies to all members of the population of interest. The ordinary least squares (OLS) estimator of the regression coefficient on D is

$$\hat{\delta}_{\text{OLS, bivariate}} \equiv \frac{\text{Cov}_N(y_i, d_i)}{\text{Var}_N(d_i)}, \tag{9.10}$$

where $\text{Cov}_N(.)$ and $\text{Var}_N(.)$ denote unbiased, sample-based estimates from a sample of size N of the population-level covariance and variance.

Now, again suppose that the back-door association between D and ε renders the least squares estimator biased and inconsistent for δ in Equation (9.9). If least squares cannot be used to effectively estimate δ, an alternative IV estimator can be attempted, with an IV Z, as in

$$\hat{\delta}_{\text{IV}} \equiv \frac{\text{Cov}_N(y_i, z_i)}{\text{Cov}_N(d_i, z_i)}, \tag{9.11}$$

where Z can now take on more than two values. If the instrument Z is linearly associated with D but unassociated with ε, then the IV estimator in Equation (9.11) is consistent for δ in Equation (9.9).[8]

One way to see why IV estimators yield consistent estimates is to again consider the population-level relationships between Y, D, and Z, as in Equations (9.1)–(9.4). Manipulating Equation (9.1) as before, the covariance between the outcome Y and the instrument Z can be written as

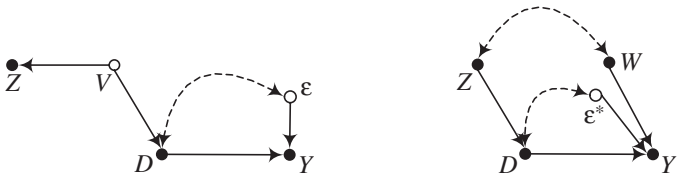
$$\text{Cov}(Y, Z) = \delta \,\text{Cov}(D, Z) + \text{Cov}(\varepsilon, Z), \tag{9.12}$$

again assuming that δ is a constant structural effect. Dividing by $\text{Cov}(D, Z)$ then yields

$$\frac{\text{Cov}(Y, Z)}{\text{Cov}(D, Z)} = \frac{\delta \,\text{Cov}(D, Z) + \text{Cov}(\varepsilon, Z)}{\text{Cov}(D, Z)}, \tag{9.13}$$

needed that uniquely index separate supply-and-demand shocks. Usually based on data from a set of exchangeable markets or alternative well-chosen equilibria for the same market from past time periods, these IVs are then used to identify different points of intersection between a shifted linear supply/demand curve and a fixed linear demand/supply curve.

[8] Notice that substituting d_i for z_i in Equation (9.11) results in the least squares regression estimator in Equation (9.10). Thus, the least squares regression estimator implicitly treats D as an instrument for itself.



(a) Z is a valid surrogate instrumental variable

(b) Z is a valid conditional instrumental variable

Figure 9.2 Two graphs in which Z is a valid IV.

which is directly analogous to Equation (9.3). When $\text{Cov}(\varepsilon, Z)$ is equal to 0 in the population, then the right-hand side of Equation (9.13) simplifies to δ. This suggests that

$$\frac{\text{Cov}_N(y_i, z_i)}{\text{Cov}_N(d_i, z_i)} \xrightarrow{p} \delta \tag{9.14}$$

if $\text{Cov}(\varepsilon, Z) = 0$ in the population and if $\text{Cov}(D, Z) \neq 0$. This would be the case for the causal diagram in Figure 9.1(a). But here the claim is more general and holds for cases in which Z is many-valued (and, in fact, for cases in which D is many-valued as well, assuming that the linear, constant-coefficient specification in Equation (9.9) is appropriate).

Our prior graphical presentation was also more restricted than it needed to be. As we discussed above, if Z is related to D and Y as in Figure 9.1(a), then IV estimation with Z is feasible.[9] Figure 9.2(a) presents another graph in which Z can be used as a valid instrumental variable for the effect of D on Y. In this case, the instrument V is unobserved. What we have instead is a surrogate instrumental variable, Z, which is not a direct cause of D but which has an association with D because both Z and D mutually depend on the unobserved instrument V. A complication, which we will discuss below, is that the association between Z and D may be very weak in this case, creating a particular set of statistical problems.[10]

Figure 9.2(b) represents perhaps the most common setup in the traditional applied literature. In this case, Z is not a "clean" instrument that has no association with Y other than the association that is generated by a directed path that begins at the instrument and ends at the outcome. Nonetheless, all of the additional associations between Z and Y are generated by back-door paths from the instrument Z to the outcome Y that can be blocked by simultaneous conditioning on observed variables, such as W in Figure 9.2(b). In this case, Z is sometimes referred to as a conditional

[9] The only difference in this section is that we are now allowing Z to take on more than two values. Because of this relaxation, in order for the graph to be compatible with the traditional setup, we need to make the further assumption that Z has a linear causal effect on D, which neither varies over individuals nor in any piecewise fashion across the levels of Z.

[10] As elsewhere up until this point in the chapter, in order to stay within the traditional setup we must now assume that linearity holds throughout, so that the effect of V on Z generates a linear relationship between Z and D.

instrumental variable, and estimation must proceed with a conditional version of Equation (9.11), usually the two-stage least squares estimator that we will discuss below. Consider the following examples of IV estimation from the applied literature.

9.2.2 Examples of IV Estimation

Among social scientists, demographic, and health researchers, economists are the most prone to the adoption of IV estimation.[11] Consider the causal effect of education on labor market earnings, as introduced earlier in Section 1.3.1 and as used as a focal example in prior chapters. As reviewed in pieces such as Card (1999) and Angrist and Krueger (2001), the causal effect of years of schooling on subsequent earnings has been estimated with a variety of IVs, including proximity to college, regional and temporal variation in school construction, tuition at local colleges, temporal variation in the minimum school-leaving age, and quarter of birth. The argument in these applications is that the IVs predict educational attainment but have no direct effects on earnings. For the quarter-of-birth instrument, which is one of the most widely discussed IVs in the literature, Angrist and Krueger (1991:981–82) reason:

> If the fraction of students who desire to leave school before they reach the legal dropout age is constant across birthdays, a student's birthday should be expected to influence his or her ultimate educational attainment. This relationship would be expected because, in the absence of rolling admissions to school, students born in different months of the year start school at different ages. This fact, in conjunction with compulsory schooling laws, which require students to attend school until they reach a specified birthday, produces a correlation between date of birth and years of schooling.... Students who are born early in the calendar year are typically older when they enter school than children born late in the year.... Hence, if a fixed fraction of students is constrained by the compulsory attendance law, those born in the beginning of the year will have less schooling, on average, than those born near the end of the year.

The results of Angrist and Krueger (1991) were sufficiently persuasive that many additional researchers have since been convinced to use related IVs. Braakmann (2011), for example, uses an IV based on month of birth to estimate the effect of education on health-related behaviors. Other researchers have used IVs that index over-time variation in compulsory schooling laws, such as Oreopoulos (2006) for the effect of education on earnings, Machin, Marie, and Vuji (2011) for the effect of education on crime, and Kemptner, Jürges, and Reinhold (2011) for the effect of education on health.[12]

[11] Some of the examples we introduce in this section adopt the traditional setup presented above, wherein the causal effect of interest is a structural constant. Many of the more recent examples adopt the alternative setup that we will introduce later in the chapter, where it is assumed as a first principle that treatment effects vary at the individual level. For now, we offer these examples only to demonstrate the typical identifying assumptions presented in Figures 9.1(a) and 9.2.

[12] And, in turn, each of these effects has been modeled with alternative IVs. For example, in the case of health and health behaviors, Lindahl (2005) used lottery winnings as an IV for the effect of income

For another example, consider the literature on the effects of military service on subsequent labor market outcomes, a topic that both economists and sociologists have studied for several decades. Some have argued that military service can serve as an effective quasi-job-training program for young men likely to otherwise experience low earnings because of weak attachment to more traditional forms of education (Browning, Lopreato, and Poston 1973) or as a more general productivity signal that generates a veteran premium (De Tray 1982). In one of the earliest studies, which focused primarily on veterans who served around the time of the Korean War, Cutright (1974) concluded that

> two years in the military environment, coupled with the extensive benefits awarded to veterans, do *not* result in a clear cut, large net positive effect of service. ... Therefore, one may question the likely utility of social programs that offer minority men and whites likely to have low earnings a career contingency in a bridging environment similar to that provided by military service. (Cutright 1974:326)

Following the war in Vietnam, attention then focused on whether any military service premium had declined (see Rosen and Taubman 1982; Schwartz 1986; see also Berger and Hirsch 1983). Angrist (1990) used the randomization created by the Vietnam-era draft lottery to estimate the veteran effect. Here, the draft lottery turns date of birth into an IV, in much the same way that compulsory school entry and school-leaving laws turn date of birth into an IV for years of education. Angrist determined that veteran status had a negative effect on earnings, which he attributed to a loss of labor market experience.[13]

The quarter-of-birth and draft-lottery IVs are widely discussed because they are determined by processes that are very unlikely to be structured by any of the unobserved determinants of the outcome of interest. Many IVs in the economics literature do not have this feature. As discussed in Section 1.3.2, much of the early debate on the effectiveness of private schooling relative to its alternatives was carried out with observational survey data on high school students from national samples of public and Catholic high schools. In attempts to resolve the concerns over selection bias, Evans and Schwab (1995), Hoxby (1996), and Neal (1997) introduced plausible IVs for

on health and mortality, while Cawley, Moran, and Simon (2010) used an over-time discontinuity in Social Security benefits as an IV for the effect of income on the obesity of the elderly.

[13] We do not mean to imply that Angrist's work on this issue ended in 1990. Not only was his work very important for understanding what IV estimators accomplish in the presence of causal effect heterogeneity (see next section), but he also continued with subsequent substantive work on the military service effect. Angrist and Krueger (1994) estimated the World War II veteran effect, and, in an attempt to reconcile past estimates, concluded: "Empirical results using the 1960, 1970, and 1980 censuses support a conclusion that World War II veterans earn no more than comparable nonveterans, and may well earn less. These findings suggest that the answer to the question 'Why do World War II veterans earn more than nonveterans?' appears to be that World War II veterans would have earned more than nonveterans even had they not served in the military. Military service, in fact, may have reduced World War II veterans' earnings from what they otherwise would have been" (Angrist and Krueger 1994:92). Then, Angrist (1998) assessed the effects of voluntary service in the military in the 1980s, and he found mixed evidence. In Angrist and Chen (2011), he revisited the Vietnam-era veteran effect, analyzing 2000 census data, arguing that the effect on education is positive because of the GI bill but is nonetheless close to zero for earnings.

Catholic school attendance. Hoxby and Neal argued that the share of the local population that is Catholic is a valid IV for Catholic school attendance, maintaining that exogenous differences in population composition influence the likelihood of attending a Catholic school (by lowering the costs of opening such schools, which then lowers the tuition that schools need to charge and to which parents respond when making school sector selection decisions). Variation in the number of Catholics in each county was attributed to lagged effects from past immigration patterns and was therefore assumed to have no direct effect on learning.[14]

These examples of IV estimation in economics are but the tip of an iceberg of applications. Some economists have argued that the attention given to IV estimation in the applied literature in the past two decades has been excessive, and a thriving debate is now under way on the future direction of observational research in economics, as revealed by a comparison of Angrist and Pischke (2010) and Imbens (2010) to Deaton (2010), Heckman and Urzua (2010), and Leamer (2010). We will discuss this debate at several points in this chapter and the next.

Moving beyond research in economics, examples of IV estimation follow more varied patterns. For political science, Dunning (2012), Sekhon and Titiunik (2012), and Sovey and Green (2011) review the growing list of applications that draw on "natural experiments," some of which are instrumental variable designs. In sociology, Bollen (2012) shows that IV estimation is uncommon but still utilized in models with a structural equations foundation.[15] Finally, epidemiologists and health researchers have not utilized instrumental variables with substantial frequency, to some extent heeding the warning of Hernán and Robins (2006b:364) about the "risk [of] transforming the methodologic dream of avoiding unmeasured confounding into a nightmare of conflicting biased estimates."

9.2.3 The IV Identifying Assumption Cannot Be Tested

The basic identification assumption underlying all of the studies just summarized – that Z has no net association with Y except for the one generated by the directed path $Z \rightarrow D \rightarrow Y$ – is a strong and untestable assumption. Some researchers believe mistakenly that this assumption is empirically testable. In particular, they believe that the assumption that Z has no direct effect on Y implies that there is no association

[14] Evans and Schwab (1995) used a student's religious identification as an IV. This IV has proven to be rather unconvincing to many scholars (see Altonji, Elder, and Taber 2005a, 2005b) because the assumed exclusion restriction appears unreasonable. Neal (1997) also used this IV but only selectively in his analysis.

[15] As noted, IV estimation is an inherent part of structural equation modeling because point estimates of coefficients are not infrequently generated by IV-based order identification in this tradition of analysis. For example, Messner, Baumer, and Rosenfeld (2004) estimate a two-equation model, with communities as the unit of analysis, where (1) the effect of community-level social trust on the homicide rate is identified by an instrument for community-level subjective alienation and (2) the effect of community-level social activism on the homicide rate is identified by instruments for the community-level average amount of television watched and extreme political attitudes. This type of application relies quite heavily on the theoretical rationale for instrument selection, more so than for the "natural experiment" instruments prized in economics where the prima facie case for the independence of the instrument Z from causes of Y other than D is stronger.

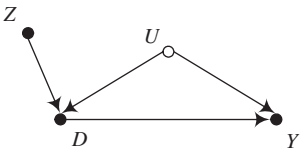


Figure 9.3 A graph with an unblocked back-door path and a valid IV.

between Z and Y conditional on D. Thus, if an association between Z and Y is detected after conditioning on D, then the assumption must be incorrect.

Causal graphs show clearly why individuals might believe that the identifying assumption can be tested in this way and also why it cannot. Consider the simplest possible causal graph, presented in Figure 9.3, in which (1) an unblocked back-door path between D and Y exists because of the unobserved common cause U, but (2) Z is a valid IV that identifies the causal effect of D on Y.[16] Here, the instrument Z is valid because it causes D, has no causal effect on Y other than through D, and is unconditionally unassociated with U.

Why is the suggested test faulty? Again, the rationale for the test is that conditioning on D will block the indirect relationship between Z and Y through D. Accordingly, if the only association between Z and Y is indirect through D, then it is thought that there should be no association between Z and Y after conditioning on D. If such a net association is detected, then it may seem reasonable to conclude that the IV identifying assumption must be false.

Although this rationale feels convincing, it is incorrect. It is certainly true that if the IV assumption is invalid, then Z and Y will be associated after conditioning on D. But the converse is not true. In fact, Z and Y will always be associated after conditioning on D when the IV assumption is valid. The explanation follows from the fact that D in Figure 9.3 is a collider that is mutually caused by both Z and U. As discussed extensively in this book, conditioning on a collider variable creates dependence between the variables that cause it. Accordingly, conditioning on D in this graph creates dependence between Z and U, even though the IV identifying assumption is valid. And, as a result, Z and Y will always be associated within at least one stratum of D even if the IV is valid. The faulty test yields an association between Z and Y when conditioning on D regardless of whether the IV identifying assumption is valid.[17]

[16] For our purposes here, this graph is isomorphic with Figure 9.1(a). We have replaced the bidirected edge from D to ε by representing the association between D and ε as a single back-door path between D and Y generated by an unobserved common cause U. We have then eliminated the remainder of the distribution of ε from the graph because it is unconditionally unassociated with Z and D (and is, in Pearl's framework, now the usual error term e_Y that comes into view only under magnification).

[17] For completeness, consider what the faulty test reveals when the IV assumption is invalid. Suppose that Figure 9.3 is augmented by an unobserved cause E and then two edges $E \rightarrow Z$ and $E \rightarrow U$. In this case, Z and Y would be associated within levels of D for two reasons: (1) conditioning on the collider D generates a net association between Z and U (and hence Z and Y) and (2) the common cause E of Z and U generates an unconditional association between Z and Y.

9.2.4 Recognized Pitfalls of Traditional IV Estimation

The traditional IV literature suggests that, as long as there is an instrument that predicts the causal variable of interest but is linearly unrelated to the outcome variable except by way of the causal variable, then an IV estimator will effectively estimate the causal effect. Even within this traditional setup, however, there are some recognized pitfalls of an IV estimation strategy. First, the assumption that an IV does not have a net direct effect on the outcome variable is often hard to defend. Second, even when an IV does not have a net direct effect on the outcome variable, IV estimators are biased in finite samples. Moreover, this bias can be substantial when an instrument only weakly predicts the causal variable. We discuss each of these weaknesses here.

Even "natural experiments" that generate compelling IVs are not immune from criticism. Consider date of birth as an instrument for veteran status in estimating the effect of military service in Vietnam on subsequent earnings (Angrist 1990). The case for excluding date of birth from the earnings equation that is the primary interest of the study is the randomization of the draft lottery, in which draft numbers were assigned at random to different dates of birth. Even though a type of randomization generates the IV, this does not necessarily imply that date of birth has no net direct effect on earnings. After the randomization occurs and the lottery outcomes are known, employers may behave differently with respect to individuals with different lottery numbers, investing more heavily in individuals who are less likely to be drafted. As a result, lottery number may be a direct, though probably weak, determinant of future earnings (see Heckman 1997; Moffitt 1996). IVs that are not generated by randomization are even more susceptible to causal narratives that challenge the assumption that the purported IV does not have a net direct effect on the outcome of interest.

Even in the absence of this complication, there are well-recognized statistical pitfalls. By using only a portion of the covariation in the causal variable and the outcome variable, IV estimators use only a portion of the information in the data. This represents a direct loss in statistical power, and as a result IV estimators tend to exhibit substantially more expected sampling variance than other estimators. By the criterion of mean-squared error, a consistent and asymptotically unbiased IV estimator can be outperformed by a biased and inconsistent regression estimator.

The problem can be especially acute in some cases. It has been shown that instruments that only weakly predict the causal variable of interest should be avoided entirely, even if they generate point estimates with acceptably small estimated standard errors (see Bound, Jaeger, and Baker 1995). In brief, the argument here is fourfold: (1) in finite samples, IV point estimates can always be computed because sample covariances are never exactly equal to zero; (2) as a result, an IV point estimate can be computed even for an instrument that is invalid because it does not predict the endogenous variable in the population (i.e., even if $\mathrm{Cov}(D, Z) = 0$ in the population, rendering Equation (9.13) undefined because its denominator is equal to 0); (3) at the same time, the formulas for calculating the standard errors of IV estimates fail in such situations, giving artificially small standard errors (when in fact the true standard error for the undefined parameter is infinity); and (4) the bias imparted by a small violation of the assumption that the IV affects the outcome variable only by way of

the causal variable can explode if the instrument is weak.[18] To see this last result, consider Equation (9.6), which depicts the expected bias in the Wald estimator for a binary IV as the term

$$\frac{E[\varepsilon|Z=1]-E[\varepsilon|Z=0]}{E[D|Z=1]-E[D|Z=0]}. \tag{9.15}$$

When the identifying assumption is violated, the numerator of Equation (9.15) is nonzero because Z is associated with Y through ε. The bias is then an inverse function of the strength of the instrument; the weaker the instrument, the smaller the denominator and the larger the bias. If the denominator is close to zero, even a tiny violation of the identifying assumption can generate a large amount of bias. And, unfortunately, this relationship is independent of the sample size. Thus, even though a weak instrument may suggest a reasonable (or perhaps even intriguing) point estimate, and one with an acceptably small estimated standard error, the IV estimate may contain no genuine information whatsoever about the true causal effect of interest (see Hahn and Hausman 2003; Small and Rosenbaum 2008; Staiger and Stock 1997; Wooldridge 2010).[19]

Beyond these widely recognized pitfalls of standard IV estimation in economics, a third criticism is emerging of current practice, especially in economics. As explained by Angrist and Krueger (2001) and Angrist and Pischke (2009, 2010), the ways in which IVs are used has changed in the past 40 years. Because it has been hard to achieve consensus that particular no-net-direct-effect assumptions are credible, IVs that arise from naturally occurring variation have become more popular. Genuine "gifts of nature," such as variation in the weather and natural boundaries, have become the most prized sources of IVs (see Dunning 2012 and Rosenzweig and Wolpin 2000 for lists of such instruments).

Not all economists see this shift in IV estimation techniques toward naturally occurring IVs as necessarily a step in the right direction. Rosenzweig and Wolpin (2000) offer one of the most cogent critiques (but see also Deaton 2010; Heckman 2000, 2005; Heckman and Urzua 2010). Rosenzweig and Wolpin make three main points. First, the variation on which naturally occurring IVs capitalize is often poorly explained and/or does not reflect the variation that maintained theories posit should be important. As a result, IV estimates from natural experiments have a black-box character that lessens their appeal as informative estimates for the development of theory or policy guidance. Second, the random variation created by a naturally occurring experiment does not necessarily ensure that an IV has no net direct effect on the outcome. Other causes of the outcome can respond to the natural event that generates the IV as well. Third, naturally occurring IVs typically do not estimate structural parameters

[18] Complications (1), (2), (3), and (4) are all closely related. Situation (4) can be considered a less extreme version of the three-part predicament depicted in (1)–(3).

[19] There are no clear guidelines on how large an association between an IV and a treatment variable must be before analysis can proceed safely. Most everyone agrees that an IV is too weak if it does not yield a test statistic that rejects a null hypotheses of no association between Z and D. However, a main point of this literature is that the converse is not true. If a dataset is large enough, the small association between Z and D generated by a weak IV can still yield a test statistic that rejects a null hypotheses of no association. Even so, the problems in the main text are not vitiated, especially the explosion of the bias generated by a small violation of the identifying assumption.

of fundamental interest, which can and should be defined in advance of estimation based on criteria other than whether a naturally occurring IV is available. Rosenzweig and Wolpin contend that natural experiments tend to lead analysts to ignore these issues because natural experiments appear falsely infallible. Reflecting on the same set of issues, Deaton (2010:432) writes, "The general lesson is once again the ultimate futility of trying to avoid thinking about how and why things work."

We will explain these points further in this chapter, after introducing IV estimation in the presence of causal effect heterogeneity. (We will also then revisit this critique in the next chapter on mechanisms.) The new IV literature, to be discussed next, addresses complications of the constant coefficient assumption implied by the stipulated constant value of δ in Equations (9.1) and (9.9). The issues are similar to those presented for regression estimators in Chapter 6, in that heterogeneity invalidates traditional causal inference from IV estimates. But IV estimators do identify specific narrow slices of average causal effects that may be of distinct interest, and as a result they represent a narrowly targeted estimation strategy with considerable appeal.

9.3 Instrumental Variable Estimators in the Presence of Individual-Level Heterogeneity

Following the adoption of a counterfactual perspective, a group of econometricians and statisticians has clarified what IVs identify when individual-level causal effects are heterogeneous. In this section, we will emphasize the work that has developed the connections between traditional IV estimators and potential-outcome-defined treatment effects (Angrist and Imbens 1995; Angrist, Imbens, and Rubin 1996; Imbens and Angrist 1994; Imbens and Rubin 1997). The key innovation here is the definition of a new treatment effect parameter: the local average treatment effect (LATE). We will also discuss other important work that has further clarified these issues, and some of this literature is more general than the LATE literature that we introduce first (see Heckman 1996, 1997, 2000, 2010; Heckman, Tobias, and Vytlacil 2003; Heckman, Urzua, and Vytlacil 2006; Heckman and Vytlacil 1999, 2000, 2005; Manski 2003; Vytlacil 2002).[20]

9.3.1 IV Estimation as LATE Estimation

Consider the following motivation of the Wald estimator in Equation (9.5), and recall the definition of Y presented in Equation (4.5):

$$\begin{aligned} Y &= Y^0 + (Y^1 - Y^0)D \\ &= Y^0 + \delta D \\ &= \mu^0 + \delta D + \upsilon^0, \end{aligned} \tag{9.16}$$

[20] Given the rapidity of these developments in the IV literature, some disagreement on the origins of these ideas pervades the literature. Heckman and Robb (1985, 1986) did provide extensive analysis of what IV estimators identify in the presence of heterogeneity. Subsequent work by others, as we discuss in this section, has clarified and extended these ideas, even while Heckman and his colleagues continued to refine their ideas.

where $\mu^0 \equiv E[Y^0]$ and $\upsilon^0 \equiv Y^0 - E[Y^0]$. Note that δ is now defined as $Y^1 - Y^0$, unlike its structural representation in Equations (9.1) and (9.9) where δ was implicitly assumed to be constant across all individuals.

To understand when an IV estimator can be interpreted as an average causal effect estimator, Imbens and Angrist (1994) developed a framework to classify individuals into those who respond positively to an instrument, those who remain unaffected by an instrument, and those who rebel against an instrument. Their innovation was to define potential treatment assignment variables, $D^{Z=z}$, for each state z of the instrument Z. When D and Z are binary variables, there are four possible groups of individuals in the population.[21] These can be summarized by a four-category latent variable C for compliance status:

$$\begin{aligned}
\text{Compliers } (C=c)&: D^{Z=0}=0 \text{ and } D^{Z=1}=1,\\
\text{Defiers } (C=d)&: D^{Z=0}=1 \text{ and } D^{Z=1}=0,\\
\text{Always takers } (C=a)&: D^{Z=0}=1 \text{ and } D^{Z=1}=1,\\
\text{Never takers } (C=n)&: D^{Z=0}=0 \text{ and } D^{Z=1}=0.
\end{aligned}$$

Consider the private schooling example presented earlier in IV Demonstration 1 (see page 294). Students who would enroll in private schools only if offered the voucher are *compliers* ($C=c$). Students who would enroll in private schools only if not offered the voucher are *defiers* ($C=d$). Students who would always enroll in private schools, regardless of whether they are offered the voucher, are *always takers* ($C=a$). And, finally, students who would never enroll in private schools are *never takers* ($C=n$).

Analogous to the definition of the observed outcome, Y, the observed treatment indicator variable D can then be defined as

$$\begin{aligned}
D &= D^{Z=0} + (D^{Z=1} - D^{Z=0})Z \\
&= D^{Z=0} + \kappa Z,
\end{aligned} \tag{9.17}$$

where $\kappa \equiv D^{Z=1} - D^{Z=0}$.[22] The parameter κ in Equation (9.17) is the individual-level causal effect of the instrument on D, and it varies across individuals if $D^{Z=1} - D^{Z=0}$ varies across individuals (i.e., if observed treatment status varies as the instrument is switched from "off" to "on" for each individual). If the instrument represents encouragement to take the treatment, such as the randomly assigned school voucher in IV Demonstration 1, then κ can be interpreted as the individual-level compliance inducement effect of the instrument. Accordingly, $\kappa = 1$ for compliers and $\kappa = -1$ for defiers. For always takers and never takers, $\kappa = 0$ because none of these individuals respond to the instrument.

[21] These four groups are considered principal strata in the framework of Frangakis and Rubin (2002).

[22] Note also that D has now been counterfactually defined with reference to D^Z. Accordingly, the definition of Y in Equation (9.16) is conditional on the definition of D in Equation (9.17). Furthermore, although there is little benefit in doing so, it bears noting that this definition of D could be structured analogously to the definition of Y in Equation (9.16) so that the last line would become $D = \zeta + \kappa Z + \iota$, where $\zeta \equiv E[D^{Z=0}]$ and $\iota \equiv D^{Z=0} - E[D^{Z=0}]$. The parameter ζ would then be the expected probability of being in the treatment if all individuals were assigned to the state "instrument switched off," and ι would be the individual-level departure from this expected value, taking on values $1 - E[D^{Z=0}]$ and $-E[D^{Z=0}]$ to balance the right-hand side of Equation (9.17) so that D is equal to either 1 or 0.

Given these definitions of potential outcome variables and potential treatment variables, a valid instrument Z for the causal effect of D on Y must satisfy three assumptions in order to identify a LATE:

$$\text{Independence assumption: } (Y^1, Y^0, D^{Z=1}, D^{Z=0}) \perp\!\!\!\perp Z, \tag{9.18}$$

$$\text{Nonzero effect of instrument assumption: } \kappa \neq 0 \text{ for all } i, \tag{9.19}$$

$$\text{Monotonicity assumption: either } \kappa \geq 0 \text{ for all } i \text{ or } \kappa \leq 0 \text{ for all } i. \tag{9.20}$$

The independence assumption in Equation (9.18) is analogous to the assumption that $\text{Cov}(Z, \varepsilon) = 0$ in the traditional IV literature; see the earlier discussion of Equation (9.13).[23] It stipulates that the instrument must be independent of the potential outcomes and potential treatments. Knowing the value of the instrument for individual i must not yield any information about the potential outcome of individual i under either treatment state. Moreover, knowing the realized value of the instrument for individual i must not yield any information about the probability of being in the treatment under alternative hypothetical values of the instrument. This latter point may well appear confusing, but it is exactly analogous to the independence assumption of potential outcomes from observed treatment status, discussed earlier for Equation (2.6). A valid instrument predicts observed treatment status (D), but it does not predict potential treatment status ($D^{Z=z}$).

The assumptions in Equations (9.19) and (9.20) are assumptions about individual responses to shifts in the instrument. The assumption of a nonzero effect of Z on D is a stipulation that the instrument must predict treatment assignment for at least some individuals. There must be at least some compliers or some defiers in the population of interest. The monotonicity assumption then further specifies that the effect of Z on D must be either weakly positive or weakly negative for all individuals i. Thus, there may be either defiers or compliers in the population but not both.[24]

If these three assumptions obtain, then an instrument Z identifies the LATE: the average treatment effect for the subset of the population whose treatment selection is induced by the instrument.[25] If $\kappa \geq 0$ for all i, then the Wald estimator from Equation

[23] The stable unit treatment value assumption (SUTVA) must continue to hold, and now it must apply to potential treatments as well. In addition, as we noted earlier, our presentation here follows Imbens and Angrist (1994), Angrist and Imbens (1995), and Angrist et al. (1996). As we note later, IVs can be defined in slightly different (in some cases more general) ways. But for now, we restrict attention to the LATE literature, in which assumptions such as complete independence of the instrument are utilized.

[24] Manski defines this assumption as a monotone treatment selection (MTS) assumption in order to distinguish it from his monotone treatment response (MTR) assumption (see Manski 1997; Manski and Pepper 2000). Vytlacil (2002) establishes its connections to the index structure model laid out in Heckman and Vytlacil (1999), as we discuss in Section 9.3.4. Heckman (2000, 2010), Heckman and Vytlacil (2005, 2007), Heckman et al. (2006), and Heckman and Urzua (2010) provide a broad accounting of the relationships between alternative IV estimators.

[25] The LATE is often referred to as the "complier average causal effect" to signify that the descriptor local is really defined by the restricted applicability of the LATE estimate to the average causal effect of compliers. The following example does not use the compliance language explicitly, except insofar as those who respond to the voucher are labeled compliers. The LATE literature that we cite provides the explicit connections between LATE and the large literature on noncompliance in randomized experiments (see in particular Imbens and Rubin 1997 and citations therein).

(9.5) converges to a particular LATE:

$$\hat{\delta}_{\text{IV,WALD}} \xrightarrow{p} E[\delta|C=c], \tag{9.21}$$

which is equal to $E[Y^1-Y^0|D^{Z=1}=1, D^{Z=0}=0]$ and is therefore the average treatment effect among compliers. In contrast, if $\kappa \leq 0$ for all i, then the Wald estimator from Equation (9.5) converges to the opposite LATE:

$$\hat{\delta}_{\text{IV,WALD}} \xrightarrow{p} E[\delta|C=d], \tag{9.22}$$

which is equal to $E[Y^1-Y^0|D^{Z=1}=0, D^{Z=0}=1]$ and is therefore the average treatment effect among defiers. In either case, the treatment effects of always takers and never takers are not informed in any way by the IV estimate.

In the next section, we will explain the claims in Equations (9.21) and (9.22) in considerable detail through a more elaborate version of IV Demonstration 1. But the basic intuition is straightforward. A valid IV is nothing more than an exogenous dimension across which the treatment and outcome variables are analyzed jointly. For a binary instrument, this dimension is a simple contrast, which is the ratio presented earlier; see Equations (9.5) and (9.8):

$$\frac{E_N[y_i|z_i=1]-E_N[y_i|z_i=0]}{E_N[d_i|z_i=1]-E_N[d_i|z_i=0]}.$$

The numerator is the naive estimate of the effect of Z on Y, and the denominator is the naive estimate of the effect of Z on D.

To give a causal interpretation to this ratio of differences across the third dimension indexed by Z, a model of individual treatment response must be specified and then used to interpret the causal effect estimate. The model of individual response adopted for a LATE analysis with a binary IV is the fourfold typology of compliance, captured by the latent variable C defined earlier for always takers, never takers, compliers, and defiers. Within this model, the always takers and never takers do not respond to the instrument (i.e., they did not choose to take part in the "experiment" created by the IV, and thus their treatment assignment is not determined by Z). This means that they are distributed in the same proportion within alternative values of the instrument Z. And, as a result, differences in the average value of Y, when examined across Z, are not a function of the outcomes of always takers and never takers.[26]

In contrast, defiers and compliers contribute all of the variation that generates the IV estimate because only their behavior is responsive to the instrument. For this reason, any differences in the average value of Y, when examined across Z, must result from treatment effects for those who move into and out of the causal states represented by D. If compliers are present but defiers are not, then the estimate offered by the ratio is interpretable as the average treatment effect for compliers. If defiers are present but compliers are not, then the estimate offered by the ratio is interpretable as the average treatment effect for defiers. If both compliers and defiers are present, then the estimate generated by the ratio does not have a well-defined causal interpretation. In the following demonstration, we will consider the most common case in the

[26] In a sense, the outcomes of always takers and never takers represent a type of background noise that is ignored by the IV estimator. More precisely, always takers and never takers have a distribution of outcomes, but the distribution of these outcomes is balanced across the values of the instrument.

LATE literature for which the monotonicity condition holds in the direction such that compliers exist in the population but defiers do not.

IV Demonstration 2

Recall the setup for IV Demonstration 1 (see page 294). In an attempt to determine whether private high schools outperform public high schools, a state education department assembles a dataset on a random sample of 10,000 ninth graders, $\{y_i, d_i, z_i\}_{i=1}^{10,000}$, where Y is a standardized test, and D is equal to 1 for students who attend private high schools and 0 for students who attend public high schools. Likewise, Z is equal to 1 for those who win a lottery for a \$3,000 school voucher and 0 for those who do not.

As noted for IV Demonstration 1, a bivariate regression of the values of y_i on d_i yielded a treatment effect estimate of 9.67; see discussion of Equation (9.7). An alternative IV estimate with Z as an instrument for D yielded an estimate of 5.5; see Equation (9.8). The hypothetical state officials relied on the IV estimate rather than on the regression estimate because they recognized that private school students have more advantaged social backgrounds. And they assumed that the randomization of the voucher lottery, in combination with the relationship between D and Z shown in Table 9.1, established the voucher lottery outcome as a valid IV.

We stated just after our discussion of IV Demonstration 1 that the estimate of 5.5 is properly interpreted as a particular LATE: the average causal effect for the subset of all students who would attend a private school if given a voucher but would not attend a private school in the absence of a voucher. To explain the reasoning behind this conclusion, we will first explain how the observed values for D and Y are generated as a consequence of variation in Z. Then we will introduce potential outcomes and use the treatment response model in order to explain why the IV estimate is interpretable as the average treatment effect for compliers.

We reported the frequency distribution of D and Z in Table 9.1. The same information is presented again in Table 9.2, but now also as probability values (where, for example, the term $\Pr_N[.,.]$ in the upper left cell is equal to $\Pr_N[d_i = 0, z_i = 0]$ by plugging the row and column headings into the joint probability statement). We also now report the expectations of the outcome variable Y, conditional on D and Z.

First, consider how the least squares and IV estimates are calculated. The coefficient of 9.67 on D in Equation (9.7) is equal to the naive estimator, $E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$. One can calculate these two conditional expectations from the elements of Table 9.2 by forming weighted averages within columns:

$$E_N[y_i|d_i = 1] = \frac{.1}{.1 + .02}60 + \frac{.02}{.1 + .02}58 = 59.667,$$
$$E_N[y_i|d_i = 0] = \frac{.8}{.8 + .08}50 + \frac{.08}{.8 + .08}50 = 50.0.$$

As shown earlier in Equation (9.8), the IV estimate of 5.5 is the ratio of two specific contrasts:

$$\frac{E_N[y_i|z_i = 1] - E_N[y_i|z_i = 0]}{E_N[d_i|z_i = 1] - E_N[d_i|z_i = 0]} = \frac{51.6 - 51.111}{.2 - .111} = 5.5. \tag{9.23}$$

Table 9.2 The Joint Probability Distribution and Conditional Expectations of the Test Score for Voucher Winner by School Sector for IV Demonstrations 1 and 2

		Public school $d_i = 0$	Private school $d_i = 1$
Voucher loser	$z_i = 0$	$N = 8000$	$N = 1000$
		$\Pr_N[.,.] = .8$	$\Pr_N[.,.] = .1$
		$E_N[y_i \mid .,.] = 50$	$E_N[y_i \mid .,.] = 60$
Voucher winner	$z_i = 1$	$N = 800$	$N = 200$
		$\Pr_N[.,.] = .08$	$\Pr_N[.,.] = .02$
		$E_N[y_i \mid .,.] = 50$	$E_N[y_i \mid .,.] = 58$

Both of these contrasts are calculated within the rows of Table 9.2 rather than the columns. The contrast in the numerator is the naive estimate of the effect of Z on Y. It is calculated as the difference between

$$E_N[y_i \mid z_i = 1] = \frac{.08}{.08 + .02}50 + \frac{.02}{.08 + .02}58 = 51.6 \quad \text{and}$$
$$E_N[y_i \mid z_i = 0] = \frac{.8}{.8 + .1}50 + \frac{.1}{.8 + .1}60 = 51.111.$$

The contrast in the denominator is the naive estimate of the effect of Z on D. It is calculated as the difference between

$$E_N[d_i = 1 \mid z_i = 1] = \frac{.02}{.08 + .02} = .2 \quad \text{and}$$
$$E_N[d_i = 1 \mid z_i = 0] = \frac{.1}{.8 + .1} = .111.$$

Thus, calculating the IV estimate in this case is quite simple and does not require any consideration of the underlying potential outcomes or potential treatments.

But, to interpret the IV estimate when causal effect heterogeneity is present, the potential outcome and potential treatment framework are needed. Consider the three identifying assumptions in Equations (9.18) through (9.20). Because the voucher lottery is completely random, the voucher instrument Z is independent of the potential outcome and potential treatment variables. Also, as just shown, Z predicts D, thereby sustaining the nonzero effect assumption. Thus, the first two assumptions are satisfied as explained for IV Demonstration 1.

Only the monotonicity assumption in Equation (9.20) requires a new justification. Fortunately, there is no evidence in the school choice literature that students and their parents rebel against vouchers, changing their behavior to avoid private schooling only

Table 9.3 The Distribution of Never Takers, Compliers, and Always Takers for IV Demonstration 2

		Public school $d_i = 0$	Private school $d_i = 1$
Voucher loser	$z_i = 0$	7200 Never takers 800 Compliers	1000 Always takers
Voucher winner	$z_i = 1$	800 Never takers	111 Always takers 89 Compliers

when offered a voucher.[27] Accordingly, it seems reasonable to assume that there are no defiers in this hypothetical population and hence that the monotonicity assumption obtains.

The joint implications of independence and monotonicity for the four groups of individuals in the cells of Table 9.2 should be clear. Monotonicity allows us to stipulate that defiers do not exist, while independence ensures that the same distribution of never takers, always takers, and compliers is present among those who win the voucher lottery and those who do not. As a result, the proportion of always takers can be estimated consistently from the first row of Table 9.2 and the proportion of never takers can be estimated consistently from the second row of Table 9.2 as

$$\frac{\Pr_N[d_i = 1, z_i = 0]}{\Pr_N[z_i = 0]} \xrightarrow{p} \Pr[C = a], \tag{9.24}$$

$$\frac{\Pr_N[d_i = 0, z_i = 1]}{\Pr_N[z_i = 1]} \xrightarrow{p} \Pr[C = n]. \tag{9.25}$$

For this example, the proportion of always takers is $1,000/9,000 = .111$, and the proportion of never takers is $800/1,000 = .8$. Because no defiers exist in the population with regard to this instrument, these two estimated proportions can be subtracted from 1 in order to obtain the proportion of compliers:

$$1 - \frac{\Pr_N[d_i = 1, z_i = 0]}{\Pr_N[z_i = 0]} - \frac{\Pr_N[d_i = 0, z_i = 1]}{\Pr_N[z_i = 1]} \xrightarrow{p} \Pr[C = c]. \tag{9.26}$$

For this example, the proportion of compliers in the population is $1 - .111 - .8 = .089$. Applying this distribution of always takers, never takers, and compliers (and a bit of rounding) to the frequencies from Table 9.2 yields the joint frequency distribution presented in Table 9.3.

For Table 9.3, notice the symmetry across rows that is generated by the independence of the instrument: $1,000/7,200 \approx 111/800$ (subject to rounding) and $800/9,000 \approx 89/1,000$ (again, subject to rounding). Of course, there is one major difference between

[27] Some parents might reason that private schooling is no longer as attractive if it is to be flooded with an army of voucher-funded children. Thus, defiers might emerge if the vouchers were widely distributed, and the monotonicity condition would then fail. In this case, however, SUTVA would also fail, necessitating deeper analysis in any case.

the two rows: The compliers are in private schools among voucher winners but in public schools among voucher losers.

Before continuing, two important points should be noted. First, it is important to recognize that the calculations that give rise to the distribution of never takers, always takers, and compliers in Table 9.3 are not determined solely by the data. In a deeper sense, they are entailed by maintenance of the monotonicity assumption. In the absence of that assumption, an unspecified number of defiers would be in the table as well, making the calculation of these proportions impossible.

Second, not all students in the dataset can be individually identified as always takers, never takers, or compliers. Consider the private school students for this example. Of these $1,200$ students, $1,000$ students are known to be always takers, as they have observed values of $d_i = 1$ and $z_i = 0$. The remaining 200 private school students are observationally equivalent, with observed values of $d_i = 1$ and $z_i = 1$. We know, based on the maintenance of the monotonicity and independence assumptions, that these 200 students include 111 always takers and 89 compliers. But it is impossible to determine which of these 200 students are among the 111 always takers and which are among the 89 compliers. The same pattern prevails for public school students. Here, we can definitively identify 800 students as never takers, but the $8,000$ public school students who are voucher losers cannot be definitively partitioned into the specific $7,200$ never takers and the 800 compliers.

Now, consider why the IV estimator yields 5.5 for this example, and then why 5.5 is interpretable as the average effect of private schooling for those who are induced to enroll in private schools because they have won vouchers. As noted already, because Z is independent of Y^1 and Y^0, the same proportion of always takers and never takers is present among both voucher winners and voucher losers. The difference in the expectation of Y across the two rows of Table 9.2 must arise from (1) the existence of compliers only in public schools in the first row and only in private schools in the second row and (2) the existence of a nonzero average treatment effect for compliers.

To see this claim more formally, recall Equation (9.21), in which this particular LATE is defined. By the linearity of expectations and the definition of an individual-level causal effect as a linear difference between y_i^1 and y_i^0, the average causal effect among compliers, $E[\delta|C = c]$, is equal to the expectation $E[Y^1|C = c]$ minus the expectation $E[Y^0|C = c]$. To obtain a consistent estimate of the average causal effect for compliers, it is sufficient to obtain consistent estimates of $E[Y^1|C = c]$ and $E[Y^0|C = c]$ separately and then to subtract the latter from the former.

Fortunately, this strategy is feasible because the contribution of these two conditional expectations to the observed data can be written out in two equations and then solved. In particular, $E[Y^1|C = c]$ and $E[Y^0|C = c]$ contribute to the expectations of the observed outcome Y, conditional on D and Z, in the following two equations:

$$\begin{aligned} E[Y|D = 1, Z = 1] = &\frac{\Pr[C = c]}{\Pr[C = c] + \Pr[C = a]} E[Y^1|C = c] \\ &+ \frac{\Pr[C = a]}{\Pr[C = c] + \Pr[C = a]} E[Y^1|C = a], \end{aligned} \tag{9.27}$$

$$E[Y|D=0,Z=0] = \frac{\Pr[C=c]}{\Pr[C=c]+\Pr[C=n]}E[Y^0|C=c] + \frac{\Pr[C=n]}{\Pr[C=c]+\Pr[C=n]}E[Y^0|C=n]. \tag{9.28}$$

These two equations are population-level decompositions of the conditional expectations for the observed data that correspond to the two cells of the diagonal of Table 9.2. These are the only two cells in which compliers are present, and thus the only two cells in which the observed data are affected by the outcomes of compliers.

How can we plug values into Equations (9.27) and (9.28) in order to solve for $E[Y^1|C=c]$ and $E[Y^0|C=c]$ and thereby obtain all of the ingredients of a consistent estimate of $E[\delta|C=c]$? We have already shown from applying the convergence assertions in Equations (9.24)–(9.26) that the terms $\Pr[C=c]$, $\Pr[C=a]$, and $\Pr[C=n]$ can be consistently estimated. And, in fact, these are given earlier for the example data as .089, .8, and .111. Thus, to solve these equations for $E[Y^1|C=c]$ and $E[Y^0|C=c]$, the only remaining pieces that need to be estimated are $E[Y^1|C=a]$ and $E[Y^0|C=n]$, which are the average outcome under the treatment for the always takers and the average outcome under the control for the never takers. Fortunately, the independence and monotonicity assumptions guarantee that voucher losers in private schools represent a random sample of always takers. Thus, $E[Y^1|C=a]$ is estimated consistently by $E_N[y_i|d_i=1,z_i=0]$, which is 60 for this example (see the upper right-hand cell in Table 9.2). Similarly, because voucher winners in public schools represent a random sample of never takers, $E[Y^0|C=n]$ is estimated consistently by $E_N[y_i|d_i=0,z_i=1]$, which is equal to 50 for this example (see the lower left-hand cell in Table 9.2). Plugging all of these values into Equations (9.27) and (9.28) then yields

$$58 = \frac{.089}{.089+.111}E[Y^1|C=c] + \frac{.111}{.089+.111}60, \tag{9.29}$$

$$50 = \frac{.089}{.089+.8}E[Y^0|C=c] + \frac{.8}{.089+.8}50. \tag{9.30}$$

Solving Equation (9.29) for $E[Y^1|C=c]$ results in 55.5, whereas solving Equation (9.30) for $E[Y^0|C=c]$ results in 50. The difference between these values is 5.5, which is the average causal effect for the subset of all students who would attend a private school if given a voucher but would not attend a private school in the absence of a voucher.[28] The value of 5.5 yields no information whatsoever about the effect of private schooling for the always takers and the never takers.[29]

[28] For completeness, consider how the naive estimate and the LATE would differ if all remained the same except the stipulated value of 50 for $E_N[d_i=0,z_i=0]$ in Table 9.2. If $E_N[d_i=0,z_i=0]$ were instead 50.25, then the naive estimate would be 9.44 and the LATE estimate would be 3.00. And, if $E_N[d_i=0,z_i=0]$ were instead 50.5, then the naive estimate would be 9.21 and the LATE estimate would be .5. Thus, for the example in the main text, compliers on average do no worse in public schools than never takers. But, for these two variants of the example, compliers on average do slightly better in public schools than never takers. As a result, the calculations in Equation (9.29) remain the same, but the values of Equation (9.30) change such that $E[Y^0|C=c]$ is equal to 52.5 and 55.0, respectively. The LATE estimate is therefore smaller in both cases because the performance of compliers in public schools is higher (whereas the performance of compliers in private schools remains the same).

[29] We can estimate $E[Y^1|C=a]$ and $E[Y^0|C=n]$ consistently with $E_N[y_i|d_i=1,z_i=0]$ and $E_N[y_i|d_i=0,z_i=1]$. But we have no way to effectively estimate their counterfactual analogs: the

Of course, the Wald estimate is also 5.5, as shown in Equations (9.8) and (9.23). And thus, in one sense, the Wald estimator can be thought of as a quick alternative method for calculating all of the steps just presented to solve exactly for $E[\delta|C=c]$. Even so, this correspondence does not explain when and how the Wald estimator can be interpreted as the average causal effect for compliers. For this example, the correspondence arises precisely because we have assumed that there are no defiers in the population, based on the substance of the application and the treatment response model that we adopted. As a result, the Wald estimate can be interpreted as a consistent estimate of the average effect of private schooling for those who comply with the instrument because it is equal to that value under the assumed model of treatment response we are willing to adopt.[30]

LATE estimators have been criticized because the identified effect is defined by the instrument under consideration. As a result, different instruments define different average treatment effects for the same group of treated individuals. And, when this is possible, the meanings of the labels for the latent compliance variable C depend on the instrument, such that some individuals can be never takers for one instrument and compliers for another. Deaton writes:

> Without explicit prior consideration of the effect of the instrument choice on the parameter being estimated, such a procedure is effectively the opposite of standard statistical practice in which a parameter of interest is defined first, followed by an estimator that delivers that parameter. Instead, we have a procedure in which the choice of the instrument … is implicitly allowed to determine the parameter of interest. This goes beyond the old story of looking for an object where the light is strong enough to see; rather, we have at least some control over the light but choose to let if fall where it may and then proclaim that whatever it illuminates is what we were looking for all along. (Deaton 2010:429)

Although from one perspective the instrument-dependent nature of the LATE is a weakness, from another perspective it is the most attractive feature of the LATE. For IV Demonstration 2, the IV estimate does not provide any information about the average effect for individuals who would attend private schooling anyway (i.e., the always takers) or for those who would still not attend the private schools if given a voucher (i.e., the never takers). Instead, the IV estimate is an estimate of a narrowly defined average effect only among those induced to take the treatment by the voucher policy intervention. But, for IV Demonstration 2, this is precisely what should be of interest to the state officials. If the policy question is "What is the effect of vouchers

mean outcome in public schools for always takers and the mean outcome in privates schools for never takers.

[30] In other words, the Wald estimate of 5.5 is also a quick method for calculating an entirely different causal effect under a different set of assumptions. If monotonicity cannot be defended, then the IV estimate can be given a traditional structural interpretation, under the assumption that the causal effect is constant for all individuals. In this sense, because the Wald estimate has more than one possible causal interpretation, merely understanding how it is calculated does not furnish an explanation for how it can be interpreted.

on school performance?" then they presumably care most about the average effect for compliers.

The limited power of the LATE interpretation of an IV estimate is thus, in some contexts, beneficial because of its targeted clarity. Moreover, when supplemented by a range of additional IV estimates (i.e., different voucher sizes and so on), complementary LATE estimates may collectively represent an extremely useful set of parameters that describe variation in the causal effect of interest for different groups of individuals exposed to the cause for alternative (but related) reasons. Before summarizing the marginal treatment effect literature that more completely specifies the interrelationships among all types of average causal effect estimators, we first lay out the implications of the LATE perspective for traditional IV estimation and consider the relevant graphs for representing IVs that identify LATEs.

9.3.2 Implications of the LATE Perspective for Traditional IV Estimation

The LATE literature specifies a set of assumptions under which it is permissible to give IV estimates an average causal effect interpretation using the potential outcome model. In this sense, the new framework is mostly a set of guidelines for how to interpret IV estimates. As such, the LATE perspective has direct implications for traditional IV estimation, as introduced earlier in this chapter.

Monotonicity and Assumptions of Homogeneous Response

An important implication of the LATE framework is that many conventional IV estimates lack a justifiable average causal effect interpretation if the IV does not satisfy a monotonicity condition. In the presence of causal effect heterogeneity and in the absence of monotonicity of response to the instrument, a conventional IV estimator yields a parameter estimate that has no clear interpretation, as it is likely an unidentifiable mixture of the treatment effects of compliers and defiers.

For IV Demonstration 2, we showed that the estimate of 5.5 is applicable to students whose families would change their child's enrollment choice from a public school to a private school for a $3,000 voucher. Can an assumption be introduced that allows the estimate of 5.5 to be be interpreted as informative about other students who do (or who would) attend private schools?

Two variants of the same homogeneity assumption allow for such extrapolated inference: the assumption that the causal effect is a structural effect that is (1) constant across all members of the population or (2) constant across all members of the population who typically take the treatment.[31] In its stronger form (1), the assumption simply asserts that the causal effect estimate is equally valid for all members of the population, regardless of whether or not the group of students whose enrollment status would change in response to the voucher is representative of the population of students as a whole. In its weaker form (2), the assumption pushes the assumed constancy of the effect only half as far, stipulating that the IV estimate is valid as an estimate

[31] These assumptions are known as constant-coefficient assumptions, homogeneous response assumptions, or shifted outcome assumptions (see Angrist and Pischke 2009; Manski 1995, 2003).

of the ATT only. For IV Demonstration 2, the weaker variant of the homogeneity assumption is equivalent to asserting that the IV estimate provides information only about the achievement gains obtained by private school students. Although weaker, the homogeneity assumption (2) is still quite strong, in that all individuals in private schools are considered homogeneous with respect to the size of the treatment effect. In examples such as IV Demonstration 2, it is clear that there are two distinct groups within the treated: always takers and compliers. And there is little reason to expect that both groups respond in exactly the same way to private schooling. Thus, for examples such as this one, Manski (1995:44) argues that this homogeneity assumption "strains credibility" because there is almost certainly patterned heterogeneity in the effect among treated individuals.

One traditional way to bolster a homogeneity assumption is to condition on variables in a vector X that can account for all such heterogeneity and then assert a conditional homogeneity of response assumption. To do so, the Wald estimator must be abandoned in favor of a two-stage least squares (2SLS) estimator. As shown in any econometrics textbook (e.g., Greene 2000; Wooldridge 2010), the endogenous regressors D and X are embedded in an encompassing $\mathbf{X}$ 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 for the constant and one for the treatment variable D). Then, a matrix $\mathbf{Z}$ is constructed that is equivalent to X, except that the column in $\mathbf{X}$ that includes the treatment variable D is replaced with its instrument Z. The 2SLS estimator is then

$$\hat{\delta}_{\text{IV,2SLS}} \equiv (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}, \tag{9.31}$$

where $\mathbf{y}$ is an $n \times 1$ vector containing the outcomes y_i. The strategy is to attempt to condition out all of the systematic variability in the observed response and then to simultaneously use the instrument Z to identify a pure net structural effect that can be regarded as an invariant constant.

Is this strategy a feasible solution? Probably not. If good measures of all of the necessary variables in X are available, a simple OLS estimator probably would have been feasible in the first place. Rarely would all possible necessary variables be available, except for a single variable that has a net additive constant effect on the outcome that can then be estimated consistently by an available IV.

Other Challenges for Interpretation

IV estimates are hard, and sometimes impossible, to interpret as LATE estimates when the instrument measures something other than an incentive to which individuals can consciously respond by complying or defying. Instruments based on exogenous field variation (as championed in Angrist and Krueger 2001 but criticized in Rosenzweig and Wolpin 2000 and Deaton 2010) can be particularly hard to interpret, because the shifts in costs and benefits that the natural variation is supposed to induce generally remain unspecified, thereby weakening a main link in the narrative that explains why some individuals take the treatment in response to the instrument.

Moreover, if two or more IVs are available, then the traditional econometric literature suggests that they should both be used to "overidentify" the model and obtain a more precise treatment effect estimate by the 2SLS estimator in Equation (9.31).

Overidentified models, in which more than one instrument is used to identify the same treatment effect, generate a mixture-of-LATEs challenge.

Consider the Catholic school example discussed earlier. Suppose that two IVs are used: the share of the county that identifies as Catholic and a student's religious identification (as in Neal 1997). Even if these potential IVs have no net direct effects on test scores (and, further, that the weak instrument problem discussed earlier is not applicable), can a theoretically meaningful LATE interpretation be given to the effect that the two instruments in this example jointly identify? For the religious identification instrument, the implied LATE is the average effect of Catholic schooling among those who are likely to attend Catholic schools only because they are Catholic. When overidentified with the IV represented by the share of the local population that is Catholic, this first LATE is mixed in with a second LATE: the average effect of Catholic schooling among those who attend Catholic schools only because of the small difference in tuition that a high proportion of Catholics in the local population tends to generate. As a result, the overidentified causal effect that is estimated by the 2SLS estimator would be an average across two very different groups of hypothetical individuals, both of which likely deserve separate attention.[32]

It is sometimes possible to deconstruct 2SLS estimates into component LATEs when multiple IVs are used, but explaining how to do so is beyond the scope of this book. We advise readers who are in this situation to first consult Angrist and Pischke (2009, section 4.5) and the work cited therein. Our position is that there is comparatively little value in estimating a single treatment effect parameter using a 2SLS model in these cases. Usually, if more than one LATE is identified, then these can and should be estimated separately.

In sum, if causal effect heterogeneity is present, then a constant-coefficient interpretation of an IV estimate is implausible. However, if the instrument satisfies a monotonicity condition and can be conceptualized as a proximate inducement to take the treatment, then IV estimates can be given LATE interpretations. These fine-grained interpretations can be very illuminating about particular groups of individuals, even though they may provide no information whatsoever about other groups of individuals in the population (including other individuals who typically choose to take the treatment). Thus, the limited nature of IV estimators when interpreted as LATE estimators shows both the potential strengths and weaknesses of IV estimation in general.

9.3.3 Graphs for IVs That Identify LATEs

In this section, we will explain how to represent instrumental variables that identify LATEs as observed variables in directed graphs. The primary complication is that compliance with the instrumental variable is a latent class variable, across which it

[32] There is also the possibility that, contra the justification of Hoxby (1996), the individuals who are induced to attend Catholic schooling because a greater share of the population is Catholic are not at all the same as those who are supposedly at the margin of a tuition-based cost calculation. There may be another mechanism that generates any observed association between Z and D, such as the possibility that the Catholic schools are simply better in these communities and hence are more attractive in general. This is one basic criticism that Rosenzweig and Wolpin (2000) level against all such naturally occurring IVs: There is often no evidence that the IV is inducing a group of individuals to take the treatment according to an assumed cost-benefit set of choices.

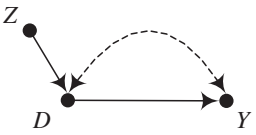


Figure 9.4 Instrumental variable identification of the causal effect of charter schools (D) on test scores (Y), where Z is the instrument.

must be assumed that heterogeneity of effects is present. Recall the charter schools example, as introduced in Section 1.3.2 and then as analyzed at length in Section 8.3. In that prior discussion, we showed how causal graphs can be used to represent complex patterns of self-selection and heterogeneity using a latent class variable. In particular, reconsider Figure 8.6(b), which we used to explain why back-door conditioning for the effect of charter schooling D on educational achievement Y was infeasible because of back-door paths through G.[33]

For simplicity, suppose now that the parental background confounder P is also unobserved. As a result, the analyst is left with no way to even begin to enact a back-door conditioning strategy for the effect of charter schools. Suppose, instead, that an instrumental variable Z is observed, as in Figure 9.4.

Are there any plausible instrumental variables for charter school attendance? The geographic distance between the student's residence and the charter school site is similar to the sorts of potential IVs used in the traditional economics literature. The rationale would be that the location of the treatment site is arbitrary but has an effect on enrollment propensity because of the implicit costs of traveling to the site, which are typically borne by the parents, not by the school district. For the charter school example, it is unclear whether such an instrument would have any chance of satisfying the relevant assumptions, and it would depend crucially on the extent to which charter schools are located in arbitrary places. Our reading of the literature is that charter schools tend to be located nearer to students most likely to benefit from charter schooling, both because many have missions to serve disadvantaged students who are thought to benefit most from having a charter school opportunity and also because parents may then move to neighborhoods that are closer to the charter schools that they select for their children. It is possible that these threats to the identifying assumption could be mitigated by conditioning on other determinants of the location of charter schools within the district and also obtaining family residence data before students entered charter schools.

For the sake of methodological clarity in our presentation, we will use as our example a more convincing but unlikely instrumental variable, in the sense that it has never

[33] For a real example with a structure that is not too dissimilar from ours, Jin and Rubin (2009) eschew graphs and adopt an alternative principal stratification approach to represent latent classes for types of compliance as well as average effects within these classes. By omission, they demonstrate that graphs are not needed in order to offer a sensible representation of underlying heterogeneity and to focus on compliance types of particular interest. The utility of the principal stratification approach for compliance latent classes, first laid out in Frangakis and Rubin (2002), is unrelated to a more recent debate on its utility for interpreting direct and indirect causal effects (see Joffe 2011; Pearl 2011; VanderWeele 2008, 2011a).

yet become available and is unlikely to become available in the future. Suppose that in New York City conditional cash transfers are offered to families that send their children to charter schools. Suppose that this program is modeled on New York City's recent Opportunity NYC program, which was justified by the position that families should be given incentives to make decisions that promote their children's futures.

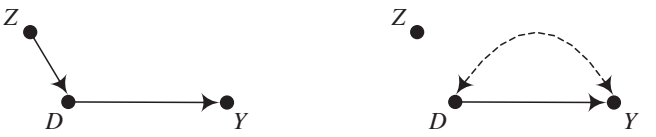
Suppose that for the new hypothetical program $3,000 in cash is offered each year to families for each child that they enroll in a charter school. Since charter schools do not charge tuition, families can spend the $3,000 per child however they see fit. Suppose further that, because of a budget constraint, cash transfers cannot be extended to all eligible families. For fairness, it is decided that families should be drawn at random from among all families resident in New York City with school-age children. Accordingly, a fixed number of letters is sent out notifying a set of winning families.

It is later determined that 10 percent of students in charter schools received cash transfers. A dataset is then compiled with performance data on all students in the school district, and the cash transfer offer is coded as a variable Z, which is equal to 1 for those who were offered a cash transfer and 0 for those who were not. A quick analysis of the data shows that some families who received offers of cash transfers turned them down and chose to send their children to regular public schools. Moreover, it is then assumed that at least some of the charter school students who received cash transfers would have attended charter schools anyway, and they were simply lucky to have also received a cash transfer.

By the standards typical of IV applications, Z would be considered a valid instrument. It is randomly assigned in the population, and it is reasonable to assume that it has a direct causal effect on D because it is an effective incentive for charter school attendance. (We have also assumed in the setup that the data show that Z predicts D.) Again, the crucial assumption is that the entire association between Z and Y is attributable solely to the directed path, $Z \rightarrow D \rightarrow Y$. As we will discuss below in this section, this assumption is debatable in this case because the subsidy is cash and, without further restrictions, could be used by families of charter school students to purchase other goods that have effects on Y. Any such alternative uses of the cash transfer would open up additional causal pathways from Z to Y that are not intercepted by D. For now, however, we will provisionally accept this identification assumption.

What parameter does Z identify? Suppose that a monotonicity assumption is valid whereby the cash transfers do not create a disincentive for anyone to enter charter schools (i.e., defiers with respect to Z do not exist in the population). This assumption allows us to abandon the constant coefficient assumption and instead assert that Z identifies the following LATE: the average effect of charter schooling among those who enter charter schools in response to the offer of a conditional cash transfer.

Figure 9.5 shows one way to represent estimators of this type. For these two graphs, the population can be partitioned into two mutually exclusive groups, compliers and noncompliers (as explained in Section 9.3, and assuming defiers do not exist). Figure 9.5(a) is the graph for compliers. No back-door paths connect D to Y in this graph because compliers, by definition, decide to enter charter schools solely based on



(a) Compliers (b) Always takers and never takers

Figure 9.5 Instrumental variable identification of the causal effect of charter schools (D) on test scores (Y), where separate graphs are drawn for compliers and noncompliers.

whether they are offered the conditional cash transfer.[34] Analogous to the distribution calculated for Table 9.3, compliers are present in both regular public schools and charter schools.

Figure 9.5(b) is the graph for noncompliers. Always takers enter charter schools regardless of whether they receive the offer of a cash transfer, and never takers do not enter charter schools regardless of whether they receive the offer of a cash transfer. As a result, Z does not cause D for either always takers or never takers. The analyst can therefore place Z within the graph, but the causal effect $Z \rightarrow D$ must be omitted. Instead, the analyst includes $D \dashleftarrow\dashrightarrow Y$ to represent the unobserved joint causes of D and Y, as it is these factors that suggest why identification via an instrumental variable is needed.

Given that a central theme of this book is the power of graphs to represent causal relationships, we will conclude by addressing a final question: Can the clarity of Figure 9.5 be represented in a single causal graph, akin to the move from Figure 8.3 to Figure 8.4 in Section 8.3? Yes, but readers may not agree that the clarity is preserved.

Figure 9.6 is a combined representation of Figure 9.5, which now applies to the full population. The representation of compliance-based heterogeneity is accomplished by augmenting the graph with a latent class variable, C, which signifies whether an individual is a complier or not.[35] In particular, C takes on three values, one for compliers, one for always takers, and one for never takers (and we assume that defiers do not exist in the population). Most importantly, C interacts with Z in determining D, and then C interacts with D in determining Y.[36]

Now, to make the connection to the fully elaborated Figure 8.7 (see page 285), consider Figure 9.7, which includes all of the relevant back-door paths between D

[34] We thank Peter Steiner for pointing out to us that, contrary to figure A2 in Morgan and Winship (2012), Figure 9.5(a) should not include $D \dashleftarrow\dashrightarrow Y$. None of the common causes of D and Y among noncompliers have analogous effects for compliers because D is determined solely by Z for compliers.

[35] An alternative and more compact graph could be used for Figure 9.6. Because C is unobserved, one could simply declare that it is a member of the set of variables that generate the bidirected edge in Figure 9.6 (or as a member of the set of variables in V that will be introduced below for Figure 9.7). We give C its own pair of explicit causal effects on D and Y for clarity, even though it makes the graph more complex than it needs to be.

[36] It is possible that one could assume that C does not determine Y. This would be the case, for example, if one had reason to believe that the LATE is equal to the ATE, which would seem to be very unlikely in social science applications.

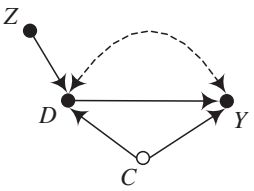


Figure 9.6 A combined graph for Figures 9.5(a)–(b), where Z is the instrument and compliance is represented as an unobserved latent class variable (C).

and Y represented as $D \dashleftarrow\dashrightarrow Y$ for Figure 9.6. Still, for simplicity we have replaced the path $I \rightarrow \text{Exp}(D \rightarrow Y) \rightarrow G$ from Figure 8.7 with the single variable V.[37] The conditional cash transfer is then represented as an instrumental variable Z, which has a sole causal effect in the graph on D because we have assumed that the cash transfer offer does not have other effects on Y. Finally, we then add the additional back-door path from D to Y through their new common cause C.

With the addition of $D \leftarrow C \rightarrow Y$ to the graph, two sources of confusion may arise for some readers. First, it remains true that we cannot use back-door conditioning to estimate the effect of D on Y because of the unblockable back-door path through $D \leftarrow V \rightarrow Y$. However, it is important to remember that the similarly structured back-door path $D \leftarrow C \rightarrow Y$ does not present any problems for an IV estimator because it is not a back-door path from Z to Y, nor part of a directed path that carries the effect of Z to Y. It only represents an additional unblocked back-door path from D to Y. Second, nothing in the causal graph itself explains why a resulting IV estimator delivers an average causal effect that applies only to compliers. To understand this point, it may be more helpful to draw two separate causal graphs, as in Figure 9.5. The single causal graph does not reveal that there is an implicit interaction between Z and C as causes of D. In particular, the instrument Z does not cause D for noncompliers, and C does not cause D for those who do not receive the offer of a cash transfer. Only the co-occurrence of Z and C switches some members of the population from $D = 0$ to $D = 1$.

Now, to conclude the discussion of the estimation of the charter school effect, consider two final points. It is likely that Figure 9.7 improperly omits the likely causal effect $P \rightarrow C$. The parental background variable P implicitly includes within it a variable for family income. Students from families with high incomes should be less likely to switch from regular public schools to charter schools because of an offer of a modest conditional cash transfer to their parents. Adding such a path, however, would not harm the feasibility of the IV estimator, since it does not generate an unblockable path from Z to Y. In fact, the effect $P \rightarrow C$ helps to explain who compliers likely are, because it suggests that they are more likely to be lower income families. In this sense, recognizing the likely presence of this effect helps to interpret the LATE that the IV identifies.[38] In addition, this type of effect reveals a distinct advantage of a

[37] This simplification is permissible under the assumption that an implicit error term e_V contains all of the information in the error terms e_I, e_{Exp}, and e_G in the causal graph in Figure 8.7.

[38] For situations such as these, Angrist and Fernandez-Val (2013) propose alternative methods for using multiple IVs and covariate information to put forward estimates of broader parameters based on identified LATEs (see also Aronow and Carnegie 2013).

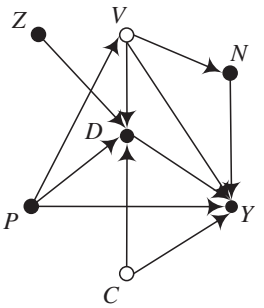


Figure 9.7 Identification of the LATE using an instrument (Z) for the charter school graph presented earlier in Figure 8.7. The unobserved variable V is a composite for the causal chain that generates self-selection in Figure 8.7 through information access and selection on the subjective evaluation of the individual-level causal effect.

single population-level graph that represents compliance as a node within it. Separate graphs for compliance classes do not permit a representation of the effect of P on compliance.

But, of course, not all additional causal effects will help to clarify the IV estimator. Suppose, for example, that Z generates an effect on N because the cash transfer is used to pay higher rent in another neighborhood for some families. As a result, a direct effect from Z to N is opened up. Conditioning on the observed variable N will block the new directed path $Z \rightarrow N \rightarrow Y$. But, because N is a collider on another path, $Z \rightarrow N \leftarrow V \rightarrow Y$, conditioning on N opens up this pathway by inducing a relationship between Z and V. Thus, conditioning away self-selection into neighborhoods then allows self-selection on the causal effect of charter schooling to confound the IV estimate of the LATE.

9.3.4 Local IVs and Marginal Treatment Effects

In this final section, we discuss one additional perspective on the identifying power of IVs in the presence of individual-level heterogeneity, which shows how a perfect instrument can help to identify a full pattern of causal effect heterogeneity. Heckman and Vytlacil (1999, 2000, 2005, 2007), building upon Heckman (1997), have shown that LATEs and many other average treatment effects can be seen as weighted averages of more fundamental marginal treatment effects.[39] Although the generality of their perspective is captivating, and the methodological content of the perspective unifies many strands of the literature in causal effect estimation, we summarize it only briefly here because the demands on data are quite substantial. The all-powerful IV that is needed to estimate a full schedule of marginal treatment effects will rarely be available to researchers.

The marginal treatment effect (MTE) perspective can be easily grasped with only a slight modification to the setup of IV Demonstration 2. Instead of 10 percent of students receiving a voucher that is exactly equal to $\$3,000$, suppose instead that

[39] See also Heckman et al. (2006) and Heckman and Urzua (2010).

these 10 percent receive a voucher that is a random draw from a uniform distribution with a minimum of $1 and a maximum equal to the tuition charged by the most expensive private school in the area.

For Heckman and his coauthors, the size of each student's voucher is a valid instrument Z, maintaining the same assumptions as we did for IV Demonstration 2 (i.e., Z is randomly assigned, Z has a nonzero effect on D, and the effect of Z on D is monotonic). The monotonicity assumption is a little more complex than before, but it stipulates that the true probability of taking the treatment is higher for all individuals with values of Z equal to z'' rather than z' if $z'' > z'$. This fits cleanly into the notation introduced earlier in this chapter by simply allowing Z to be many-valued.

Heckman and his coauthors then define two related concepts: a local instrumental variable (LIV) and an MTE. An LIV is the limiting case of a component binary IV drawn from Z in which z'' approaches z' for any two values of Z such that $z'' > z'$. Each LIV then defines a marginal treatment effect, which is the limiting form of a LATE, in which the IV is an LIV.

Consider the more elaborate version of IV Demonstration 2 just introduced here. One could form LIVs from Z by stratifying the data by the values of Z and then considering adjacent strata. Given a large enough sample for a large enough voucher program, LIVs could be constructed for each dollar increase in the voucher. Each LIV could then be used to estimate a LATE, and these LIV-identified LATEs could then be considered MTEs.

Heckman and Vytlacil (2005) show that most average causal effect estimates can be represented as weighted averages of MTEs, identified by LIVs. But the weighting schemes differ based on the parameter of interest, some of which, as was the case in our regression chapter, may have no inherent interest. Heckman and Vytlacil therefore propose a more general strategy. They argue that researchers should define the policy-relevant treatment effect (PRTE) based on an assessment of how a contemplated policy would affect treatment selection. Then, MTEs should be estimated with LIVs and weighted appropriately to obtain the PRTE that is of primary interest.

There is much to recommend in this approach, and in fact it should not be considered an approach relevant only to policy research. The approach is quite easily extended to targeted theory-relevant causal effects, for which one wishes to weight marginal causal effects according to a foundational theoretical model. But, in spite of this appeal, the entire approach may well come to represent a gold standard for what ought to be done rather than what actually can be done in practice. If it is generally recognized that IVs satisfying the LATE assumptions are hard to find, then those that satisfy LIV assumptions for all MTEs of interest must be harder still.[40]

[40] Partly for this reason, Carneiro, Heckman, and Vytlacil (2010) elaborate the PRTE perspective and focus attention on a limiting form of the PRTE that they label marginal policy-relevant treatment effects (MPRTEs). They argue that it is easier to estimate MPRTEs and that these may be all that are needed to evaluate proposed policy changes.

9.4 Conclusions

The impressive development of the IV literature in econometrics and statistics in the past two decades suggests a variety of recommendations for practice that differ from those found in the older IV literature:

1. Weak instruments yield estimates that are especially susceptible to finite sample bias. Consequently, natural experiments should be avoided if the implied IVs only very weakly predict the causal variable of interest. No matter how seductive their claims to satisfy identification assumptions may be, resulting point estimates and standard errors may be very misleading.

2. If individual-level causal effect heterogeneity is present, a potential IV should be used only if it satisfies a monotonicity condition. IV estimates should then be interpreted as LATE estimates defined by the instrument.

3. If individual-level causal effect heterogeneity is present, IVs should probably not be combined in a 2SLS model (except in the rare cases in which measures of all of the variables that account for the causal effect heterogeneity are available). Instead, IV estimates should be offered for those IVs that satisfy monotonicity conditions. These alternative estimates should then be interpreted as LATE estimates and reconciled with each other based on a narrative about why the causal effect varies for different types of individuals who are exposed to the cause (and/or different levels of the cause) for different reasons.

4. When possible, IVs should be used to examine general patterns of causal effect heterogeneity. Using IVs to estimate only the ATE or ATT is too narrow of a purpose because there is likely a good deal of variation in the treatment effect that is amenable to analysis with complementary IVs. The possibilities for this type of analysis are most clearly developed in the literature on the identification of MTEs using LIVs.

Properly handled, there is much to recommend in the IV estimation strategy for causal analysis. But, of course, IVs may not be available. We now turn to other techniques that may allow for the point identification and estimation of a causal effect when a complete model of causal exposure cannot be formulated because selection is determined by relevant unobserved variables.