# Chapter 16
# INSTRUMENTAL VARIABLE ESTIMATION

The causal inference methods described so far in this book rely on a key untestable assumption: all variables needed to adjust for confounding and selection bias have been identified and correctly measured. If this assumption is incorrect—and it will always be to a certain extent—there will be residual bias in our causal estimates.

It turns out that there exist other methods that can validly estimate causal effects under an alternative set of assumptions that do not require measuring all adjustment factors. Instrumental variable estimation is one of those methods. Economists and other social scientists reading this book can breathe now. We are finally going to describe a very common method in their fields, a method that is unlike any other we have discussed so far.

## 16.1 The three instrumental conditions

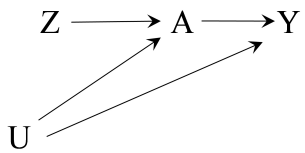


Figure 16.1

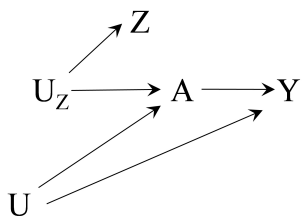


Figure 16.2

The causal diagram in Figure 16.1 depicts a randomized trial: $Z$ is the randomization assignment indicator (1: treatment, 0: placebo), $A$ is an indicator for receiving treatment (1: yes, 0: no) because not all participants adhere to their assignment, $Y$ the outcome, and $U$ all factors (some unmeasured) that affect both the outcome and the adherence. Because participants and their doctors do not know whether the pill they are given is treatment or placebo, they are said to be "blinded " and the study is referred to as a *double-blind placebo-controlled* randomized trial.

Suppose we want to consistently estimate the average causal effect of $A$ on $Y$. Whether we use IP weighting, standardization, g-estimation, stratification, or matching, we need to correctly measure, and adjust for, variables that block the backdoor path $A \leftarrow U \rightarrow Y$, i.e., we need to ensure conditional exchangeability of the treated and the untreated. Unfortunately, all these methods will result in biased effect estimates if some of the necessary variables are unmeasured, imperfectly measured, or misspecified in the model.

Instrumental variable (IV) methods are different: they may be used to attempt to identify the average causal effect of $A$ on $Y$ in this randomized trial, even if we did not measure the variables normally required to adjust for the confounding caused by $U$. To perform their magic, IV methods need an instrumental variable $Z$, or an *instrument*. A variable $Z$ is an instrument because it meets three instrumental conditions:

(i) $Z$ is associated with $A$

(ii) $Z$ does not affect $Y$ except through its potential effect on $A$

(iii) $Z$ and $Y$ do not share causes

See Technical Point 16.1 for a more rigorous definition of these conditions, which we will use in the other technical points.

Condition (ii) would not be guaranteed if, for example, participants were inadvertently unblinded by side effects of treatment.

In the double-blind randomized trial described above, the randomization indicator $Z$ is an instrument. Condition (i) is met because trial participants are more likely to receive treatment if they were assigned to treatment, condition (ii) is expected by the double-blind design, and condition (iii) is expected by the random assignment of $Z$.

Technical Point 16.1

**The instrumental conditions, formally.** Instrumental condition (i), sometimes referred to as the *relevance* condition, is non-null association between $Z$ and $A$, or $Z \perp\!\!\!\not\perp A$ does not hold. Condition (i) is expected to hold in randomized experiments because treatment assignment is expected to influence the treatment received.

Instrumental condition (ii), commonly known as the *exclusion restriction*, is the condition of "no direct effect of $Z$ on $Y$." At the individual level, condition (ii) is $Y_i^{z,a} = Y_i^{z',a} = Y_i^a$ for all $z, z'$, all $a$, all individuals $i$. However, for some results presented in this chapter, only the population level condition (ii) is needed, i.e., $\mathrm{E}\left[Y^{z,a}\right] = \mathrm{E}\left[Y^{z',a}\right]$. Both versions of condition (ii) are expected to hold in double-blind randomized experiments because assignment is not expected to influence the outcome (e.g., through behavioral changes) if assignment is unknown to all individuals. Condition (ii) is trivially true for surrogate instruments.

Instrumental condition (iii) can be written as *marginal exchangeability* $Y^{a,z} \perp\!\!\!\perp Z$ for all $a, z$, which holds in the SWIGs corresponding to Figures 16.1, 16.2, and 16.3. Together with condition (ii) at the individual level, it implies $Y^a \perp\!\!\!\perp Z$. A stronger condition (iii) is joint exchangeability, or $\{Y^{z,a}; a \in [0,1], z \in [0,1]\} \perp\!\!\!\perp Z$ for dichotomous treatment and instrument. See Technical Point 2.1 for a discussion on different types of exchangeability and Technical Point 16.2 for a description of results that require each version of exchangeability. Both versions of condition (iii) are expected to hold in randomized experiments because $Z$ is randomly assigned.

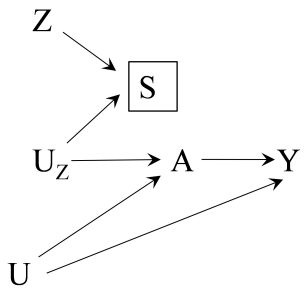


Figure 16.3

Figure 16.1 depicts a special case in which the instrument $Z$ has a causal effect on the treatment $A$. We then refer to $Z$ as a *causal instrument*. Sometimes the causal instrument is unmeasured and we use a measured proxy or *surrogate instrument* $Z$ that is associated with the unmeasured causal instrument $U_Z$. A surrogate instrument does not have a causal effect on treatment $A$, but meets the instrumental conditions with the $Z$-$A$ association (i) now resulting from the cause $U_Z$ shared by $Z$ and $A$, and with condition (iii) modified as "$Z$ and $Y$ do not share causes except for $U_Z$". Both causal and surrogate instruments can be used for IV estimation, with some caveats described in Section 16.4. As a curiosity, Figure 16.3 depicts an example of an unusual surrogate instrument $Z$ in a selected population: the $Z$-$A$ association arises from conditioning on a common effect $S$ of the unmeasured causal instrument $U_Z$ and the surrogate instrument $Z$.

In previous chapters we have estimated the effect of smoking cessation on weight change using various causal inference methods applied to observational data. To estimate this effect using IV methods, we need an instrument $Z$. Since there is no randomization indicator in an observational study, consider the following candidate for an instrument: the price of cigarettes. It can be argued that this variable meets the three instrumental conditions if (i) cigarette price affects the decision to quit smoking, (ii) cigarette price affects weight change only through its effect on smoking cessation, and (iii) no common causes of cigarette price and weight change exist. Fine Point 16.1 reviews some proposed instruments in observational studies.

Condition (i) is met if the candidate instrument $Z$ "price in state of birth" is associated with smoking cessation $A$ through the unmeasured variable $U_Z$ "price in place of residence".

To fix ideas, let us propose an instrument $Z$ that takes value 1 when the average price of a pack of cigarettes in the U.S. state where the individual was born was greater than \$1.50, and takes value 0 otherwise. Unfortunately, we cannot determine whether our variable $Z$ is actually an instrument. Of the three instrumental conditions, only condition (i) is empirically verifiable. To verify this condition we need to confirm that the proposed instrument $Z$ and the treatment $A$ are associated, i.e., that $\Pr[A = 1|Z = 1] - \Pr[A = 1|Z = 0] > 0$. The probability of quitting smoking is 25.8% among those with $Z = 1$ and 19.5% among those with $Z = 0$; the risk difference $\Pr[A = 1|Z = 1] -$

Fine Point 16.1

**Candidate instruments in observational studies.** Many variables have been proposed as instruments in observational studies and it is not possible to review all of them here. Three commonly used categories of candidate instruments are

- Genetic factors: The proposed instrument $Z$ is a genetic variant $Z$ that is associated with treatment $A$ and that, supposedly, is only related with the outcome $Y$ through $A$. For example, when estimating the effects of alcohol intake on the risk of coronary heart disease, $Z$ can be a polymorphism associated with alcohol metabolism (say, ALDH2 in Asian populations). Causal inference from observational data via IV estimation using genetic variants is part of the framework known as *Mendelian randomization* (Katan 1986, Davey Smith and Ebrahim 2004, Didelez and Sheehan 2007, VanderWeele et al. 2014).
- Preference: The proposed instrument $Z$ is a measure of the physician's (or a care provider's) preference for one treatment over the other. The idea is that a physician's preference influences the prescribed treatment $A$ without having a direct effect on the outcome $Y$. For example, when estimating the effect of prescribing COX-2 selective versus non-selective nonsteroidal anti-inflammatory drugs on gastrointestinal bleeding, $U_Z$ can be the physician's prescribing preference for drug class (COX-2 selective or non-selective). Because $U_Z$ is unmeasured, investigators replace it in the analysis by a (measured) surrogate instrument $Z$, such as "last prescription issued by the physician before current prescription" (Korn and Baumrind 1998, Earle et al. 2001, Brookhart and Schneeweiss 2007).
- Access: The proposed instrument $Z$ is a measure of access to the treatment. The idea is that access impacts the use of treatment $A$ but does not directly affect the outcome $Y$. For example, physical distance or travel time to a facility has been proposed as an instrument for treatments available at such facilities (McClellan et al. 1994, Card 1995, Baiocchi et al. 2010). Another example: calendar period has been proposed as an instrument for a treatment whose accessibility varies over time (Hoover et al. 1994, Detels et al. 1998). In the main text we use "price of the treatment", another measure of access, as a candidate instrument.

$\Pr[A = 1|Z = 0]$ is therefore 6%. When, as in this case, $Z$ and $A$ are weakly associated, $Z$ is often referred as a *weak instrument* (more on weak instruments in Section 16.5).

On the other hand, conditions (ii) and (iii) cannot be empirically verified. To verify condition (ii), we would need to prove that $Z$ can only cause the outcome $Y$ through the treatment $A$. We cannot prove it by conditioning on $A$, which is a collider on the pathway $Z \longleftarrow U_Z \rightarrow A \longleftarrow U \rightarrow Y$ in Figure 16.2, because that would induce an association between $Z$ and $Y$ even if condition (ii) held true. And we cannot, of course, prove that condition (iii) holds because we can never rule out confounding for the effect of any variable. We can only assume that conditions (ii) and (iii) hold. IV estimation, like all methods we have studied so far, is based on untestable assumptions.

Conditions (ii) and (iii) can sometimes be empirically falsified by using data on instrument, treatment, and outcome. However, falsification tests only reject the conditions for a small subset of violations. For most violations, the test has no statistical power, even for an arbitrarily large sample size (Balke and Pearl 1997, Bonet 2001, Glymour et al. 2012).

In observational studies we cannot prove that our proposed instrument $Z$ is truly an instrument. We refer to $Z$ as a proposed or *candidate instrument* because we can never guarantee that the structures represented in Figures 16.1 and 16.2 are the ones that actually occur. The best we can do is to use subject-matter knowledge to build a case for why the proposed instrument $Z$ may be reasonably assumed to meet conditions (ii) and (iii); this is similar to how we use subject-matter knowledge to justify the identifying assumptions of the methods described in previous chapters.

But let us provisionally assume that $Z$ is an instrument. Now what? Can we now see the magic of IV estimation in action? Can we consistently estimate the average causal effect of $A$ on $Y$ without having to identify and measure

Technical Point 16.2

**Bounds: Partial identification of causal effects.** For a dichotomous outcome $Y$, the average causal effect $\Pr\left[Y^{a=1}=1\right]-\Pr\left[Y^{a=0}=1\right]$ can take values between $-1$ (if all individuals develop the outcome unless they were treated) and $1$ (if no individuals develop the outcome unless treated). The bounds of the average causal effect are $(-1,1)$. The distance between these bounds can be cut in half by using the data: because for each individual we know the value of either her counterfactual outcome $Y^{a=1}$ (if the individual was actually treated) or $Y^{a=0}$ (if the individual was actually untreated), we can compute the causal effect after assigning the most extreme values possible to each individual's unknown counterfactual outcome. This will result in bounds of the average causal effect that are narrower but still include the null value $0$. For a continuous outcome $Y$, deriving bounds requires the specification of the minimum and maximum values for the outcome; the width of the bounds will vary depending on the chosen values.

The bounds for $\Pr\left[Y^{a=1}=1\right]-\Pr\left[Y^{a=0}=1\right]$ can be further narrowed if there is a variable $Z$ that meets instrumental condition (ii) at the population level (see Technical Point 16.1) and marginal exchangeability (iii) (Robins 1989; Manski 1990). The width of these so-called *natural bounds*, $\Pr[A=1|Z=0]+\Pr[A=0|Z=1]$, is narrower than that of the bounds identified from the data alone. Sometimes narrower bounds—the so-called *sharp bounds*—can be achieved when marginal exchangeability is replaced by joint exchangeability (Balke and Pearl 1997; Richardson and Robins 2014).

The conditions necessary to achieve the sharp bounds can also be derived from the SWIGs under joint interventions on $z$ and $a$ corresponding to any of the causal diagrams depicted in Figures 16.1, 16.2, and 16.3. Richardson and Robins (2010, 2014) showed that the conditions $Y^{a,z}\perp\!\!\!\perp(Z,A)\,|U$ and $Z\perp\!\!\!\perp U$, together with a population level condition (ii) within levels of $U$, i.e., $\mathrm{E}\left[Y^{z,a}|U\right]=\mathrm{E}\left[Y^{z',a}|U\right]$, are sufficient to obtain the sharp bounds. These conditions, which hold for all three SWIGs, imply $Z\perp\!\!\!\perp U$, $Y\perp\!\!\!\perp Z|U,A$, and that $\mathrm{E}\left[Y^{z,a}\right]$ is given by the g-formula $\int\mathrm{E}\left[Y|A=a,U=u\right]dF(u)$ ignoring $Z$, which reflects that $Z$ has no direct effect on $Y$ within levels of $U$. Dawid (2003) proved that these latter conditions lead to the sharp bounds. Under further assumptions, Richardson and Robins derived yet narrower bounds. See also Richardson, Evans, and Robins (2011).

Unfortunately, all these partial identification methods (i.e., methods for bounding the effect) are often relatively uninformative because the bounds are wide. Swanson et al (2018) review partial identification methods for binary instruments, treatments, and outcomes. Swanson et al. (2015a) describe a real-world application of several partial identification methods and discuss their relative advantages and disadvantages.

There is a way to decrease the width of the bounds: making parametric assumptions about the form of the effect of $A$ on $Y$. Under sufficiently strong assumptions described in Section 16.2, the upper and lower bounds converge into a single number and the average causal effect is point identified.

the confounders? Sadly, the answer is no without further assumptions. An instrument by itself does not allow us to identify the average causal effect of smoking cessation $A$ on weight change $Y$, but only identifies certain upper and lower bounds. Typically, the bounds are very wide and often include the null value (see Technical Point 16.2).

In our example, these bounds are not very helpful. They would only confirm what we already knew: smoking cessation can result in weight gain, weight loss, or no weight change. Unfortunately, that is all an instrument can offer unless one is willing to make additional unverifiable assumptions. Sections 16.3 and 16.4 review additional conditions under which the IV estimand is the average causal effect. Before that, we review the methods to do so.

## 16.2 The usual IV estimand

When a dichotomous variable $Z$ is an instrument, i.e., meets the three instrumental conditions (i)-(iii), and an additional condition (iv) described in the

next section holds, then the average causal effect of treatment on the additive scale $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ is identified and equals

$$\frac{\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]}{\mathrm{E}[A|Z=1] - \mathrm{E}[A|Z=0]},$$

which is the *usual IV estimand* for a dichotomous instrument. (Note $\mathrm{E}[A|Z=1] = \Pr[A=1|Z=1]$ for a dichotomous treatment). Technical Point 16.3 provides a proof of this result in terms of an additive structural mean model, but you might want to wait until the next section before reading it.

We will focus on dichotomous instruments, which are the commonest ones. For a continuous instrument $Z$, the usual IV estimand is $\frac{Cov(Y,Z)}{Cov(A,Z)}$, where $Cov$ means covariance.

To intuitively understand the usual IV estimand, consider again the randomized trial from the previous section. The numerator of the IV estimand is the average causal effect of assignment $Z$ on $Y$—the intention-to-treat effect—and the denominator is the average causal effect of assignment $Z$ on $A$—a measure of adherence to, or compliance with, the assigned treatment. When there is perfect adherence, the denominator is equal to 1, and the effect of $A$ on $Y$ equals the effect of $Z$ on $Y$. As adherence worsens, the denominator starts to get closer to 0, and the effect of $A$ on $Y$ becomes greater than the effect of $Z$ on $Y$. The lower the adherence, the greater the difference between the effect of $A$ on $Y$—the IV estimand—and the effect of $Z$ on $Y$.

In randomized experiments, the IV estimand is the ratio of two effects of $Z$: the effect of $Z$ on $Y$ and the effect of $Z$ on $A$. Each of these effects can be consistently estimated without adjustment because $Z$ is randomly assigned.

The IV estimand bypasses the need to adjust for the confounders by inflating the effect of assignment (the numerator). The magnitude of the inflation increases as adherence decreases, i.e., as the $Z$-$A$ risk difference (the denominator) gets closer to zero. The same rationale applies to the instruments used in observational studies, where the denominator of the IV estimator may equal either the causal effect of the causal instrument $Z$ on $A$ (Figure 16.1), or the noncausal association between the surrogate instrument $Z$ and the treatment $A$ (Figures 16.2 and 16.3).

Also known as the Wald estimator (Wald 1940).

The standard IV estimator is calculated as the ratio of the estimates of the numerator and the denominator of the usual IV estimand. In our smoking cessation example with a dichotomous instrument $Z$ (1: state with high cigarette price, 0: otherwise), the numerator estimate $\widehat{\mathrm{E}}[Y|Z=1] - \widehat{\mathrm{E}}[Y|Z=0]$ equals $2.686 - 2.536 = 0.1503$ and the denominator $\widehat{\mathrm{E}}[A|Z=1] - \widehat{\mathrm{E}}[A|Z=0]$ equals $0.2578 - 0.1951 = 0.0627$. Therefore, the usual IV estimate is the ratio $0.1503/0.0627 = 2.4$ kg. Under the three instrumental conditions (i)-(iii) plus condition (iv) from next section, this is an estimate of the average causal effect of smoking cessation on weight gain in the population.

CODE: Program 16.1
For simplicity, we exclude individuals with missing outcome or instrument. In practice, we could use IP weighting to adjust for possible selection bias before using IV estimation.

We estimated the numerator and denominator of the IV estimand by simply calculating the four sample averages $\widehat{\mathrm{E}}[A|Z=1]$, $\widehat{\mathrm{E}}[A|Z=0]$, $\widehat{\mathrm{E}}[Y|Z=1]$, and $\widehat{\mathrm{E}}[Y|Z=0]$. Equivalently, we could have fit two (saturated) linear models to estimate the differences in the denominator and the numerator. The model for the denominator would be $\mathrm{E}[A|Z] = \alpha_0 + \alpha_1 Z$, and the model for the numerator $\mathrm{E}[Y|Z] = \beta_0 + \beta_1 Z$.

An alternative method to calculate the standard IV estimator is the *two-stage-least-squares estimator*. The procedure is as follows. First, fit the first-stage treatment model $\mathrm{E}[A|Z] = \alpha_0 + \alpha_1 Z$, and generate the predicted values $\widehat{\mathrm{E}}[A|Z]$ for each individual. Second, fit the second-stage outcome model $\mathrm{E}[Y|Z] = \beta_0 + \beta_1 \widehat{\mathrm{E}}[A|Z]$. The parameter estimate $\widehat{\beta}_1$ will always be numerically equivalent to the standard IV estimate. Thus, in our example, the two-stage-least-squares estimate was again 2.4 kg.

CODE: Program 16.2

The 2.4 point estimate has a very large 95% confidence interval: −36.5 to 41.3. This is expected for our proposed instrument because the $Z$-$A$ association is weak and there is much uncertainty in the first-stage model. A commonly

used rule of thumb is to declare an instrument as weak if the F-statistic from the first-stage model is less than 10 (it was a meager 0.8 in our example). We will revisit the problems raised by weak instruments in Section 16.5.

Some of the assumptions implicit in regarding the two-stage-least-squares estimator as identifying the causal effect of treatment can be made more explicit by using additive or multiplicative structural mean models, like the ones described in Technical Points 16.3 and 16.4, for IV estimation. The parameters of structural mean models can be estimated via g-estimation. In addition, in the presence of measured common causes $L$ of the instrument and the outcome that therefore must be adjusted for in the analysis, the trade-offs involved in the choice between two-stage-least-squares linear models and structural mean models can be similar to those involved in the choice between outcome regression and structural nested models for non-IV estimation (see Chapters 14 and 15).

CODE: Program 16.3

Anyway, the above estimators are only valid when the usual IV estimand can be interpreted as the average causal effect of treatment $A$ on the outcome $Y$. For that to be true, a fourth identifying condition needs to hold in addition to the three instrumental conditions.

## 16.3 A fourth identifying condition: homogeneity

The three instrumental conditions (i)-(iii) are insufficient to ensure that the IV estimand is the average causal effect of treatment $A$ on $Y$. A fourth condition, *effect homogeneity* (iv), is needed. Here we describe four possible homogeneity conditions (iv) in order of (historical) appearance.

The most extreme, and oldest, version of homogeneity condition (iv) is constant effect of treatment $A$ on outcome $Y$ across individuals. In our example, this condition would hold if smoking cessation made every individual in the population gain (or lose) the same amount of weight, say, exactly 2.4 kg. A constant effect is equivalent to additive rank preservation which, as we discussed in Section 14.4, is scientifically implausible for most treatments and outcomes—and impossible for dichotomous outcomes, except under the sharp null or universal harm (or benefit). In our example, we expect that, after quitting smoking, some individuals will gain a lot of weight, some will gain little, and others may even lose some weight. Therefore, we are not generally willing to accept the homogeneity assumption of constant effect as a reasonable condition (iv).

Yet additive rank preservation was implicitly assumed in many early IV analyses using the two-stage-least-squares estimator.

A second, less extreme homogeneity condition (iv) for dichotomous $Z$ and $A$ is equality of the average causal effect of $A$ on $Y$ across levels of $Z$ in both the treated and in the untreated, i.e., $\mathrm{E}\left[Y^{a=1} - Y^{a=0} | Z = 1, A = a\right] = \mathrm{E}\left[Y^{a=1} - Y^{a=0} | Z = 0, A = a\right]$ for $a = 0, 1$. This additive homogeneity condition (iv) was the one used in the mathematical proof of Technical Point 16.3. An alternative homogeneity condition on the multiplicative scale is discussed in Technical Point 16.4. (This multiplicative homogeneity condition leads to an IV estimand that is different from the usual IV estimand.)

Even when condition (iii) $Y^a \perp\!\!\!\perp Z$ holds—as in the SWIGs for Figures 16.1, 16.2, 16.3—$Y^a \perp\!\!\!\perp Z | A$ does not generally hold. Therefore the treatment effect may depend on $Z$, i.e., the less extreme homogeneity condition may not hold.

The above homogeneity condition is expressed in terms that are not naturally intuitive. How can subject-matter experts provide arguments in support of a constant average causal effect within levels of the proposed instrument $Z$ and the treatment $A$ in any particular study? More natural—even if still untestable—homogeneity conditions (iv) would be stated in terms of effect modification by possibly known (even if unmeasured) confounders $U$. One

Technical Point 16.3

**Additive structural mean models and IV estimation.** Consider the following saturated, additive structural mean model for a dichotomous treatment $A$ and an instrument $Z$ as depicted in Figures 16.1, 16.2, or 16.3:

$$\mathrm{E}\left[Y^{a=1} - Y^{a=0} | A = 1, Z\right] = \beta_0 + \beta_1 Z$$

This model can also be written as $\mathrm{E}\left[Y - Y^{a=0} | A, Z\right] = A\left(\beta_0 + \beta_1 Z\right)$. The parameter $\beta_0$ is the average causal effect of treatment among the treated individuals with $Z = 0$, and $\beta_0 + \beta_1$ is the average causal effect of treatment among the treated individuals with $Z = 1$. Thus $\beta_1$ quantifies additive effect modification by $Z$.

If we a priori assume that there is no additive effect modification by $Z$, then $\beta_1 = 0$ and $\beta_0$ is exactly the usual IV estimand (Robins 1994). That is, the usual IV estimand is the parameter of an additive structural mean model for the effect of treatment on the treated under no effect modification by $Z$.

The proof is simple. When $Z$ is an instrument, condition (ii) holds, which implies $\mathrm{E}\left[Y^{a=0} | Z = 1\right] = \mathrm{E}\left[Y^{a=0} | Z = 0\right]$. Under the above structural model, this conditional mean independence can be rewritten as $\mathrm{E}\left[Y - A\left(\beta_0 + \beta_1\right) | Z = 1\right] = \mathrm{E}\left[Y - A\beta_0 | Z = 0\right]$. Solving the above equation with $\beta_1 = 0$ we have

$$\beta_0 = \frac{\mathrm{E}\left[Y | Z = 1\right] - \mathrm{E}\left[Y | Z = 0\right]}{\mathrm{E}\left[A | Z = 1\right] - \mathrm{E}\left[A | Z = 0\right]}$$

You may wonder why we a priori set $\beta_1 = 0$. The reason is that we have an equation with two unknowns ($\beta_0$ and $\beta_1$) and that equation exhausts the constraints on the data distribution implied by the three instrumental conditions. Since we need an additional constraint, which by definition will be untestable, we arbitrarily choose $\beta_1 = 0$ (rather than, say, $\beta_1 = 2$). This is what we mean when we say that an instrument is insufficient to identify the average causal effect.

Therefore, to conclude that the average causal effect of treatment in the treated $\beta_0 = \mathrm{E}\left[Y^{a=1} - Y^{a=0} | A = 1, Z = z\right] = \mathrm{E}\left[Y^{a=1} - Y^{a=0} | A = 1\right]$ equals the average causal effect in the study population $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$—and thus that the usual IV estimand is $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$—we must assume that the effects of treatment in the treated and in the untreated are identical, which is an additional untestable assumption.

Hence, under the additional assumption $\beta_1 = 0$, $\beta_0 = \mathrm{E}\left[Y^{a=1} - Y^{a=0} | A = 1, Z = z\right] = \mathrm{E}\left[Y^{a=1} - Y^{a=0} | A = 1\right]$ for any $z$ is the average causal effect of treatment in the treated.

To conclude that $\beta_0$ is the average causal effect in the study population $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$—and thus that $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ is the usual IV estimand—we must assume that the effects of treatment are identical in the treated *and* in the untreated, i.e., the parameter for $Z$ is also 0 in the structural model for $A = 0$. This is an additional untestable assumption.

Hernán and Robins (2006b) showed that, if $U$ is an additive effect modifier, then it would not be reasonable for us to believe that the previous additive homogeneity condition (iv) holds.

Wang and Tchetgen Tchetgen (2018) proposed the homogeneity condition "$t(U)$ is a constant" and proved that it was a special case of the general condition. See a proof in Technical Point 16.5

such condition is that $U$ is not an additive effect modifier, i.e., that the average causal effect of $A$ on $Y$ is the same at every level of the unmeasured confounder $U$ or $\mathrm{E}\left[Y^{a=1} | U\right] - \mathrm{E}\left[Y^{a=0} | U\right] = \mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$. This third homogeneity condition (iv) is often implausible because some unmeasured confounders may also be effect modifiers. For example, the magnitude of weight gain after smoking cessation may vary with prior intensity of smoking, which may itself be an unmeasured confounder for the effect of smoking cessation on weight gain.

Another type of homogeneity condition (iv) is that the $Z$-$A$ association on the additive scale is constant across levels of the unmeasured confounders $U$, i.e., $\mathrm{E}\left[A | Z = 1, U\right] - \mathrm{E}\left[A | Z = 0, U\right] = \mathrm{E}\left[A | Z = 1\right] - \mathrm{E}\left[A | Z = 0\right]$. Both this condition and the earlier condition $\mathrm{E}\left[Y^{a=1} | U\right] - \mathrm{E}\left[Y^{a=0} | U\right] = \mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ are special cases of the following (much more) general condition: the modification by $U$ of the effect of the treatment $A$ on the outcome $Y$, $e\left(U\right) \equiv \mathrm{E}\left[Y^{a=1} - Y^{a=0} | U\right]$, is uncorrelated with the modification by $U$ of the $Z$-$A$ association on the additive scale, $t\left(U\right) \equiv \mathrm{E}\left[A | Z = 1, U\right] - \mathrm{E}\left[A | Z = 0, U\right]$,

Technical Point 16.4

**Multiplicative structural mean models and IV estimation.** Consider the following saturated, multiplicative (log-linear) structural mean model for a dichotomous treatment $A$

$$\frac{\mathrm{E}\left[Y^{a=1}|A=1,Z\right]}{\mathrm{E}\left[Y^{a=0}|A=1,Z\right]}=\exp\left(\beta_0+\beta_1 Z\right),$$

which can also be written as $\mathrm{E}\left[Y|A,Z\right]=\mathrm{E}\left[Y^{a=0}|A,Z\right]\exp\left[A\left(\beta_0+\beta_1 Z\right)\right]$. For a dichotomous $Y$, $\exp\left(\beta_0\right)$ is the causal risk ratio in the treated individuals with $Z=0$ and $\exp\left(\beta_0+\beta_1\right)$ is the causal risk ratio in the treated with $Z=1$. Thus $\beta_1$ quantifies multiplicative effect modification by $Z$. If we a priori assume that $\beta_1=0$—and additionally assume no multiplicative effect modification by $Z$ in the untreated—then the causal effect on the multiplicative (risk ratio) scale is $\mathrm{E}\left[Y^{a=1}\right]/\mathrm{E}\left[Y^{a=0}\right]=\exp\left(\beta_0\right)$, and the causal effect on the additive (risk difference) scale is

$$\mathrm{E}\left[Y^{a=1}\right]-\mathrm{E}\left[Y^{a=0}\right]=\mathrm{E}\left[Y|A=0\right]\left(1-\mathrm{E}\left[A\right]\right)\left[\exp\left(\beta_0\right)-1\right]+\mathrm{E}\left[Y|A=1\right]\mathrm{E}\left[A\right]\left[1-\exp\left(-\beta_0\right)\right]$$

The proof, which relies on the instrumental conditions, can be found in Robins (1989) and Hernán and Robins (2006b).

That is, if we assume a multiplicative structural mean model with no multiplicative effect modification by $Z$ in the treated and in the untreated, then the average causal effect $\mathrm{E}\left[Y^{a=1}\right]-\mathrm{E}\left[Y^{a=0}\right]$ remains identified, but no longer equals the usual IV estimator. As a consequence, our estimate of $\mathrm{E}\left[Y^{a=1}\right]-\mathrm{E}\left[Y^{a=0}\right]$ will depend on whether we assume no additive or multiplicative effect modification by $Z$. Unfortunately, it is not possible to determine which, if either, assumption is true even if we had an infinite sample size (Robins 1994) because, when considering saturated additive or multiplicative structural mean models, we have more unknown parameters to estimate than equations to estimate them with. That is precisely why we need to make modeling assumptions such as homogeneity.

i.e., $Cov\left[e\left(U\right),t\left(U\right)\right]=0$. The previous two conditions are special cases of $Cov\left[e\left(U\right),t\left(U\right)\right]=0$ because they can be expressed as "$e(U)$ is a constant" and "$t(U)$ is a constant", respectively, and the covariance of any variable with a constant is always 0.

Because of the perceived implausibility of the homogeneity conditions in many settings, the possibility that IV methods can validly estimate the average causal effect of treatment seems questionable. There are two approaches that bypass the homogeneity conditions.

One approach is the introduction of baseline covariates in the models for IV estimation. To do so, it is safer to use structural mean models, which impose fewer parametric assumptions than two-stage-linear-squares estimators. The inclusion of covariates in a structural mean model allows the treatment effect in the treated to vary with $Z$ by imposing constraints on how the treatment effect varies within levels of the covariates. See Section 16.5. and Technical Point 16.6 for more details on structural mean models with covariates.

Also, models can be used to incorporate multiple proposed instruments simultaneously, to handle continuous treatments, and to estimate causal risk ratios when the outcome is dichotomous (see Palmer et al. 2011 for a review).

Another approach is to use an alternative condition (iv) that does not require effect homogeneity. When combined with the three instrumental conditions (i)-(iii), this alternative condition allows us to endow the usual IV estimand with a causal interpretation, even though it does not suffice to identify the average causal effect in the population. We review this alternative condition (iv) in the next section.

Technical Point 16.5

**Proof of the general homogeneity condition.** We wish to show that $\mathrm{E}[Y^{a=1} - Y^{a=0}] = \frac{\mathrm{E}[Y|Z=1]-E[Y|Z=0]}{\mathrm{E}[A|Z=1]-\mathrm{E}[A|Z=0]}$ under the causal diagram in Figure 16.1 and the general homogeneity condition of zero covariance $Cov\left(e\left(U\right), t\left(U\right)\right) = 0$, where $e\left(U\right)$ and $t\left(U\right)$ are defined in the main text.

To do so, note that $\mathrm{E}\left[e\left(U\right)\right] = \mathrm{E}[Y^{a=1} - Y^{a=0}]$. Further, $\mathrm{E}\left[t\left(U\right)\right] = \mathrm{E}\left[A|Z=1\right] - \mathrm{E}\left[A|Z=0\right]$ because $U \perp\!\!\!\perp Z$. Hence the zero covariance condition implies $\mathrm{E}[e\left(U\right)t\left(U\right)] / \left\{\mathrm{E}\left[A|Z=1\right] - \mathrm{E}\left[A|Z=0\right]\right\} = \mathrm{E}[Y^{a=1} - Y^{a=0}]$. It remains to show that $\mathrm{E}\left[Y|Z=1\right] - \mathrm{E}\left[Y|Z=0\right] = \mathrm{E}[e\left(U\right)t\left(U\right)]$. To do so, write $Y = A\left(Y^{a=1} - Y^{a=0}\right) + Y^{a=0}$. Because $Y^a \perp\!\!\!\perp (A, Z)\,|U$ and $U \perp\!\!\!\perp Z$ in Figure 16.1, we have $\mathrm{E}\left[Y|Z\right]$ equal to

$= \sum_u \sum_{a=\{0,1\}} \mathrm{E}\left[Y|A=a, Z, U=u\right] \Pr\left(A=a|Z, U=u\right) f\left(u|Z\right)$

$= \sum_u \left\{\mathrm{E}[Y^{a=1} - Y^{a=0}|U=u] \Pr\left(A=1|Z, U=u\right) + \mathrm{E}\left[Y^{a=0}|U=u\right]\right\} f\left(u\right).$

Thus, $\mathrm{E}\left[Y|Z=1\right] - \mathrm{E}\left[Y|Z=0\right] = \mathrm{E}\left[\left\{\mathrm{E}[Y^{a=1} - Y^{a=0}|U]\right\}\left\{\Pr\left(A=1|Z=1, U\right) - \Pr\left(A=1|Z=0, U\right)\right\}\right]$ as required.

## 16.4 An alternative fourth condition: monotonicity

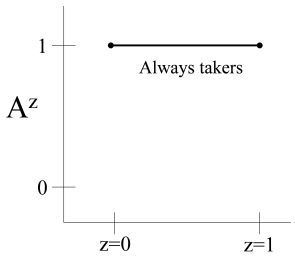


Figure 16.4

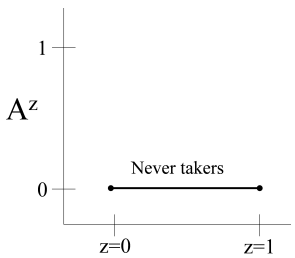


Figure 16.5

Consider again the double-blind randomized trial with randomization indicator $Z$, treatment $A$, and outcome $Y$. For each individual in the trial, the counterfactual variable $A^{z=1}$ is the value of treatment—1 or 0—that an individual would have taken if he had been assigned to receive treatment ($z = 1$). The counterfactual variable $A^{z=0}$ is analogously defined as the treatment value if the individual had been assigned to receive no treatment ($z = 0$).

If we knew the values of the two counterfactual treatment variables $A^{z=1}$ and $A^{z=0}$ for each individual, we could classify all individuals in the study population into four disjoint subpopulations:

1. *Always-takers*: Individuals who will always take treatment, regardless of the treatment group they were assigned to. That is, individuals with both $A^{z=1} = 1$ and $A^{z=0} = 1$.

2. *Never-takers*: Individuals who will never take treatment, regardless of the treatment group they were assigned to. That is, individuals with both $A^{z=1} = 0$ and $A^{z=0} = 0$.

3. *Compliers* or cooperative: Individuals who will take treatment when assigned to treatment, and no treatment when assigned to no treatment. That is, individuals with $A^{z=1} = 1$ and $A^{z=0} = 0$.

4. *Defiers* or contrarians: Individuals who will take no treatment when assigned to treatment, and treatment when assigned to no treatment. That is, individuals with $A^{z=1} = 0$ and $A^{z=0} = 1$.

Note that these subpopulations—often referred as *compliance types* or *principal strata*—are not generally identified. If we observe that an individual was assigned to $Z = 1$ and took treatment $A = 1$, we do not know whether she is a complier or an always-taker. If we observe that an individual was assigned to $Z = 1$ and took treatment $A = 0$, we do not know whether he is a defier or a never-taker.

When no defiers exist, we say that there is monotonicity because the instrument $Z$ either does not change treatment $A$—as shown in Figure 16.4 for always-takers and Figure 16.5 for never-takers—or increases the value of treatment $A$—as shown in Figure 16.6 for compliers. For defiers, the instrument

Technical Point 16.6

**More general structural mean models.** Consider an additive structural mean model that allows for continuous and/or multivariate treatments $A$, instruments $Z$, and pre-instrument covariates $V$. Such model assumes

$$\mathrm{E}\left[Y - Y^{a=0}|Z, A, V\right] = \gamma\left(Z, A, V; \beta\right)$$

where $\gamma\left(Z, A, V; \beta\right)$ is a known function, $\beta$ is an unknown (possibly vector-valued) parameter, and $\gamma\left(Z, A = 0, V; \beta\right) = 0$. That is, an additive structural mean model is a model for the average causal effect of treatment level $A$ compared with treatment level 0 among the subset of individuals at level $Z$ of the instrument and level $V$ of the confounders whose observed treatment is precisely $A$. The parameters of this model can be identified via g-estimation under the conditional counterfactual mean independence assumption $\mathrm{E}\left[Y^{a=0}|Z = 1, V\right] = \mathrm{E}\left[Y^{a=0}|Z = 0, V\right]$.

Analogously, a general multiplicative structural mean model assumes

$$\mathrm{E}\left[Y|Z, A, V\right] = \mathrm{E}\left[Y^{a=0}|Z, A, V\right] \exp\left[\gamma\left(Z, A, V; \beta\right)\right]$$

where $\gamma\left(Z, A, V; \beta\right)$ is a known function, $\beta$ is an unknown parameter vector, and $\gamma\left(Z, A = 0, V; \beta\right) = 0$. The parameters of this model can also be identified via g-estimation under analogous conditions. Identification conditions and efficient estimators for structural mean models were discussed by Robins (1994) and reviewed by Vansteelandt and Goetghebeur (2003). More generally, g-estimation of nested additive and multiplicative structural mean models can extend IV methods for time-fixed treatments and confounders to settings with time-varying treatments and confounders.

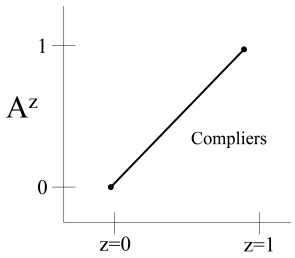


Figure 16.6

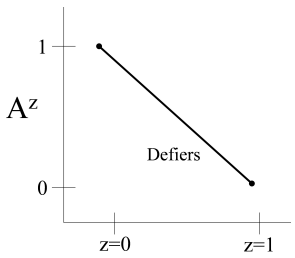


Figure 16.7

$Z$ would decrease the value of treatment $A$—as shown in Figure 16.7. More generally, monotonicity holds when $A^{z=1} \geq A^{z=0}$ for all individuals.

Now let us replace any of the homogeneity conditions from the last section by the monotonicity condition, which will become our new condition (iv). Then the usual IV estimand does not equal the average causal effect of treatment $\mathrm{E}\left[Y^{a=1}\right] - \mathrm{E}\left[Y^{a=0}\right]$ any more. Rather, under monotonicity (iv), the usual IV estimand equals the average causal effect of treatment in the compliers, that is

$$\mathrm{E}\left[Y^{a=1} - Y^{a=0}|A^{z=1} = 1, A^{z=0} = 0\right].$$

Technical Point 16.6 shows a proof for this equality under the assumption that $Z$ was effectively randomly assigned. As a sketch of the proof, the equality between the usual IV estimand and the effect in the compliers holds because the effect of assignment $Z$ on $Y$—the numerator of the IV estimand—is a weighted average of the effect of $Z$ in each of the four principal strata. However, the effect of $Z$ on $Y$ is exactly zero in always-takers and never-takers because the effect of $Z$ is entirely mediated through $A$ and the value of $A$ in those subpopulations is fixed, regardless of the value of $Z$ they are assigned to. Also, no defiers exist under monotonicity (iv). Therefore the numerator of the IV estimand is the effect of $Z$ on $Y$ in the compliers—which is the same as the effect of $A$ on $Y$ in the compliers—times the proportion of compliers in the population, which is precisely the denominator of the usual IV estimand.

In observational studies, the usual IV estimand can also be used to estimate the effect in the compliers in the absence of defiers. Technically, there are no compliers or defiers in observational studies because the proposed instrument $Z$ is not treatment assignment, but the term compliers refers to individuals with $(A^{z=1} = 1, A^{z=0} = 0)$ and the term defiers to those with $(A^{z=1} = 0, A^{z=0} = 1)$. In our smoking cessation example, the compliers are the individuals who would quit smoking in a state with high cigarette price and who would not quit smoking in a state with low price. Conversely, the defiers are the individuals

who would not quit smoking in a state with high cigarette price and who would quit smoking in a state with low price. If no defiers exist and the causal instrument is dichotomous (see below and Technical Point 16.6), then 2.4 kg is the IV effect estimate in the compliers.

The replacement of homogeneity by monotonicity was welcomed in the mid-1990s as the salvation of IV methods. While homogeneity is often an implausible condition (iv), monotonicity appeared credible in many settings. IV methods under monotonicity (iv) cannot identify the average causal effect in the population, only in the subpopulation of compliers, but that seemed a price worth paying in order to keep powerful IV methods in our toolbox. However, the estimation of the average causal effect of treatment in the compliers under monotonicity (iv) has been criticized on several grounds.

The "compliers average causal effect" (CACE) is a local average treatment effect (LATE) in a subpopulation, not the global average causal effect in the entire population. Greenland (2000) refers to compliers as cooperative, and to defiers as non-cooperative, to prevent confusion with the concept of (observed) compliance in randomized trials.

First, the relevance of the effect in the compliers is questionable. The subpopulation of compliers is not identified and, even though the proportion of compliers in the population can be calculated (it is the denominator of the usual IV estimand, see Technical Point 16.7), it varies from instrument to instrument and from study to study. Therefore, causal inferences about the effect in the compliers are difficult to use by decision makers. Should they prioritize the administration of treatment $A = 1$ to the entire population because treatment has been estimated to be beneficial among the compliers, which happen to be 6% of the population in our example but could be a smaller or larger group in the real world? What if treatment is not as beneficial in always-takers and never-takers, the majority of the population? Unfortunately, the decision maker cannot know who is included in the 6%. Rather than arguing that the effect of the compliers is of primary interest, it may be better to accept that interest in this estimand is not the result of its practical relevance, but rather of the (often erroneous) perception that it is easy to identify.

Deaton (2010) on the CACE: "This goes beyond the old story of looking for an object where the light is strong enough to see; rather, we have control over the light, but choose to let it fall where it may and then proclaim that whatever it illuminates is what we were looking for all along."

Second, monotonicity is not always a reasonable assumption in observational studies. The absence of defiers seems a safe assumption in randomized trials: we do not expect that some individuals will provide consent for participation in a trial with the perverse intention to do exactly the opposite of what they are asked to do. Further, monotonicity is ensured by design in trials in which those assigned to no treatment are prevented from receiving treatment, i.e., there are no always-takers or defiers. In that scenario, the effect in the compliers is actually the effect in the treated.

A mitigating factor is that, under strong assumptions, investigators can characterize the compliers in terms of their distribution of the observed variables (Angrist and Pischke 2009, Baiocchi et al 2014).

However, monotonicity is harder to justify for some instruments proposed in observational studies. Consider the proposed instrument "physician preference" to estimate the treatment effect in patients attending a clinic where two physicians with different preferences work. The first physician usually prefers to prescribe the treatment, but she makes exceptions for her patients with diabetes (because of some known contraindications). The second usually prefers to not prescribe the treatment, but he makes exceptions for his more physically active patients (because of some perceived benefits). Any patient who was both physically active and diabetic would have been treated contrary to both of these physicians' preferences, and therefore would be labeled as a defier. That is, monotonicity is unlikely to hold when the decision to treat is the result of weighing multiple criteria or dimensions of encouragement that include both risks and benefits. In these settings, the proportion of defiers may not be negligible.

The example to the right was proposed by Swanson and Hernán (2014). Also Swanson et al (2015b) showed empirically the existence of defiers in an observational setting.

The situation is even more complicated for the surrogate instruments $Z$ represented by Figures 16.2 and 16.3. If the causal instrument $U_Z$ is continuous (e.g., the true, unmeasured physician's preference), then the standard IV estimand using a dichotomous surrogate instrument $Z$ (e.g., some mea-

sured surrogate of preference) is not the effect in a particular subpopulation of compliers. Rather, the standard IV estimand identifies a particular weighted average of the effect in all individuals in the population, which makes it difficult to interpret. Therefore the interpretation of the IV estimand as the effect in the compliers is questionable when the proposed dichotomous instrument is not causal, even if monotonicity held for the continuous causal instrument $U_Z$ (see Technical Point 16.7 for details).

Definition of monotonicity for a continuous causal instrument $U_Z$: $A^{u_z}$ is a non-decreasing function of $u_z$ on the support of $U_Z$ (Angrist and Imbens 1995, Heckman and Vytlacil 1999).

Last, but definitely not least important, the partitioning of the population into four subpopulations or principal strata may not be justifiable. In many realistic settings, the subpopulation of compliers is an ill-defined subset of the population. For example, using the proposed instrument "physician preference" in settings with multiple physicians, all physicians with the same preference level *who could have seen a patient* would have to treat the patient in the exact same way. This is not only an unrealistic assumption, but also essentially impossible to define in many observational studies in which it is unknown which physicians could have seen a patient. A stable partitioning into compliers, defiers, always takers and never takers also requires deterministic counterfactuals (not generally required to estimate average causal effects), no interference (e.g., I may be an always-taker, but decide not to take treatment when my friend doesn't), absence of multiple versions of treatment and other forms of heterogeneity (a complier in one setting, or for a particular instrument, may not be a complier in another setting).

Swanson et al (2015b) discuss the difficulties to define monotonicity, and introduce the concept of global and local monotonicity in observational studies.

In summary, if the effect in the compliers is considered to be of interest, relying on monotonicity (iv) seems a promising approach in double-blind randomized trials with two arms and all-or-nothing compliance, especially when one of the arms will exhibit full adherence by design. However, caution is needed when using this approach in more complex settings and observational studies, even if the proposed instrument were really an instrument.

Sommer and Zeger (1991), Imbens and Rubin (1997), and Greenland (2000) describe examples of full compliance in the control group.

## 16.5 The three instrumental conditions revisited

The previous sections have discussed the relative advantages and disadvantages of choosing monotonicity or homogeneity as the condition (iv). Our discussion implicitly assumed that the proposed instrument $Z$ was in fact an instrument. However, in observational studies, the proposed instrument $Z$ will fail to be a valid instrument if it violates either of the instrumental conditions (ii) or (iii), and will be a weak instrument if it only barely meets condition (i).

In all these cases, the use of IV estimation may result in substantial bias even if condition (iv) held perfectly. We now discuss each of the three instrumental conditions.

Condition (i), a $Z$-$A$ association, is empirically verifiable. Before declaring $Z$ as their proposed instrument, investigators will check that $Z$ is associated with treatment $A$. However, when the $Z$-$A$ association is weak as in our smoking cessation example, the instrument is said to be weak (see Fine Point 16.2). Three serious problems arise when the proposed instrument is weak.

In the context of linear models, Martens et al. (2006) showed that instruments are guaranteed to be weak in the presence of strong confounding, because a strong $A$-$U$ association leaves little residual variation for a strong $A$-$U_Z$, or $A$-$Z$, association.

First, weak instruments yield effect estimates with wide 95% confidence intervals, as in our smoking cessation example in Section 16.2. Second, weak instruments amplify bias due to violations of conditions (ii) and (iii). A proposed instrument $Z$ which is weakly associated with treatment $A$ yields a small denominator of the IV estimator. Therefore, violations of conditions (ii) and (iii) that affect the numerator of the IV estimator (e.g., unmeasured con-

Fine Point 16.2

**Defining weak instruments** There are two related, but different, definitions of weak instrument in the literature:

1. An instrument is (substantively) weak if the true value of the $Z$-$A$ association—the denominator of the IV estimand—is "small."
2. An instrument is (statistically) weak if the F-statistic associated to the observed $Z$-$A$ association is "small," typically meaning less than $10$.

In our smoking cessation example, the proposed instrument met both definitions: the risk difference was only $6\%$ and the F-statistic was a meager $0.8$.

The first definition, based on the true value of the $Z$-$A$ association, reminds us that, even if we had an infinite sample, the IV estimator greatly amplifies any biases in the numerator when using a proposed weak instrument (the second problem of weak instruments in the main text). The second definition, based on the statistical properties of the $Z$-$A$ association, reminds us that, even if we had a perfect instrument $Z$, the IV estimator can be biased in finite samples (the third problem of weak instruments in the main text).

founding for the instrument, a direct effect of the instrument) will be greatly exaggerated. In our example, any bias affecting the numerator of the IV estimator would be multiplied by approximately 15.9 (1/0.0627). Third, even with a valid instrument and a large sample size, weak instruments introduce bias in the standard IV estimator.

Bound, Jaeger and Baker (1995) documented this bias. Their paper was followed by many others that investigated the shortcomings of weak instruments.

To understand the nature of this third problem, consider a randomly generated dichotomous variable $Z$. In an infinite population, the denominator of the IV estimand will be exactly zero—there is a zero association between treatment $A$ and a completely random variable—and the IV estimate will be undefined. However, in a study with a finite sample, chance will lead to an association between the randomly generated $Z$ and the unmeasured confounders $U$—and therefore between $Z$ and treatment $A$—that is weak but not exactly zero. If we propose this random $Z$ as an instrument, the denominator of the IV estimator will be very small rather than zero. As a result the numerator will be incorrectly inflated, which will yield potentially very large bias. In fact, our proposed instrument "Price higher than \$1.50" behaves like a randomly generated variable. Had we decided to define $Z$ as price higher than \$1.60, \$1.70, \$1.80, or \$1.90, the IV estimate would have been 41.3, −40.9, −21.1, or −12.8 kg, respectively. In each case, the 95% confidence interval around the estimate was huge. Given how much bias and variability weak instruments may create, a strong proposed instrument that slightly violates conditions (ii) and (iii) may be preferable to a less invalid, but weaker, proposed instrument.

CODE: Program 16.4

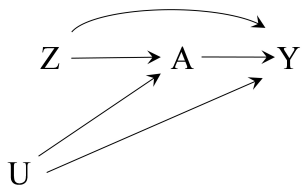


Figure 16.8

Condition (ii), the absence of a direct effect of the instrument on the outcome, cannot be verified from the data. A deviation from condition (ii) can be represented by a direct arrow from the instrument $Z$ to the outcome $Y$, as shown in Figure 16.8. This direct effect of the instrument that is not mediated through treatment $A$ will contribute to the numerator of the IV estimator, and it will be incorrectly inflated by the denominator as if it were part of the effect of treatment $A$.

Condition (ii) may be violated when a continuous or multi-valued treatment $A$ is replaced in the analysis by a coarser (e.g., dichotomized) version $A^*$. Figure 16.9 shows that, even if condition (ii) holds for the original treatment $A$, it does not have to hold for its dichotomized version $A^*$, because the path

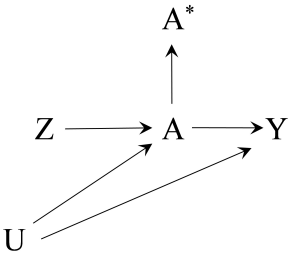


Figure 16.9

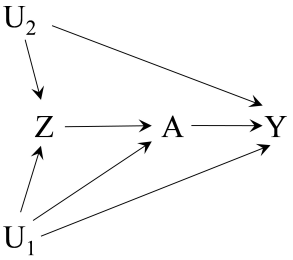


Figure 16.10

CODE: Program 16.5

$Z \rightarrow A \rightarrow Y$ represents a direct effect of the instrument $Z$ that is not mediated through the treatment $A^*$ whose effect is being estimated in the IV analysis. In practice, many treatments are replaced by coarser versions for simplicity of interpretation. Coarsening of treatment is problematic for IV estimation, but not necessarily for the methods discussed in previous chapters.

Condition (iii), no confounding for the effect of the instrument on the outcome, is also unverifiable. Figure 16.10 shows confounding due to common causes of the proposed instrument $Z$ and outcome $Y$, which may ($U_1$) or may not ($U_2$) be causes of treatment $A$. In observational studies, the possibility of confounding for the proposed instrument always exists (same as for any other variable not under the investigator's control). Confounding contributes to the numerator of the IV estimator and is incorrectly inflated by the denominator as if it were part of the effect of treatment $A$ on the outcome $Y$.

Sometimes condition (iii), and the other conditions too, can appear more plausible within levels of the measured covariates. Rather than making the unverifiable assumption that there is absolutely no confounding for the effect of $Z$ on $Y$, we might feel more comfortable making the unverifiable assumption that there is no unmeasured confounding for the effect of $Z$ on $Y$ within levels of the measured pre-instrument covariates $V$. We could then apply IV estimation repeatedly in each stratum of $V$, and pool the IV effect estimates under the assumption that the effect in the population (under homogeneity) or in the compliers (under monotonicity) is constant within levels of $V$. Alternatively we could include the variables $V$ as covariates in the two-stage modeling. In our example, this reduced the size of the effect estimate and increased its 95% confidence interval.

Another frequent strategy to support condition (iii) is to check for balanced distributions of the measured confounders across levels of the proposed instrument $Z$. The idea is that, if the measured confounders are balanced, it may be more likely that the unmeasured ones are balanced too. However, this practice may offer a false sense of security: even small imbalances can lead to counterintuitively large biases because of the bias amplification discussed above.

A violation of condition (iii) may occur even in the absence of confounding for the effect of $Z$ on $Y$. The formal version of condition (iii) requires exchangeability between individuals with different levels of the proposed instrument. Such exchangeability may be violated because of either confounding (see above) or selection bias. A surprisingly common way in which selection bias may be introduced in IV analyses is the exclusion of individuals with certain values of treatment $A$. For example, if individuals in the population may receive treatment levels $A = 0$, $A = 1$, or $A = 2$, an IV analysis restricted to individuals with $A = 1$ or $A = 2$ may yield a non-null effect estimate even if the true causal effect is null. This exclusion does not introduce bias in non-IV analyses whose goal is to estimate the effect of treatment $A = 1$ versus $A = 2$.

Swanson et al. (2015c) describe this selection bias in detail.

All the above problems related to conditions (i)-(iii) are exacerbated in IV analyses that use simultaneously multiple proposed instruments in an attempt to alleviate the weakness of a single proposed instrument. Unfortunately, the larger the number of proposed instruments, the more likely that some of them will violate one of the instrumental conditions.

## 16.6 Instrumental variable estimation versus other methods

IV estimation differs from all previously discussed methods in at least three aspects.

First, IV estimation replaces the assumption of conditional exchangeability by other assumptions. IP weighting and standardization require that the treated and the untreated are exchangeable conditional on the measured variables. In contrast, IV estimation can provide valid effect estimates, even if conditional exchangeability does not hold, when conditions (i)-(iv) hold. Therefore, the choice of method will depend on whether, in a particular research setting, it is easier to identify and measure the confounders of the effect of $A$ on $Y$ or to find an instrument $Z$ and expect that there is monotonicity or no relevant effect heterogeneity.

IV estimation is not the only method that ignores conditional exchangeability for identification of causal effects. Other approaches like *regression discontinuity analysis* (see Fine Point 16.3) and difference-in-differences (see Technical Point 7.3) do too.

Second, relatively minor violations of conditions (i)-(iv) for IV estimation may result in large biases. The foundation of IV estimation is that the denominator blows up the numerator. Therefore, when the conditions do not hold perfectly or the instrument is weak, there is potential for serious bias in either direction. As a result, an IV estimate may sometimes be more biased than an unadjusted estimate. In contrast, IP weighting and standardization tend to result in slightly biased estimates when their identifiability conditions are only slightly violated, and adjustment is less likely to introduce a large bias. The sensitivity of IV estimates to departures from its identifiability conditions highlights the importance of sensitivity analyses.

Baiocchi et al. (2014) review some approaches to quantify how sensitive IV estimates are to violations of key assumptions.

Third, the ideal setting for the applicability of standard IV estimation is more restrictive than that for other methods. As discussed in this chapter, standard IV estimation is better reserved for settings with lots of unmeasured confounding, a truly dichotomous and time-fixed treatment $A$, and a strong (and causal) proposed instrument $Z$, and in which either effect homogeneity or—if one is genuinely interested in the effect in the compliers—monotonicity is expected to hold. A consequence of these restrictions is that IV estimation is generally used to answer causal questions about point interventions. For this reason, IV estimation will not be a prominent method in Part III of this book, which is devoted to time-varying treatments and the contrast of complex treatment strategies that are sustained over time.

Transparency requires proper reporting of IV analyses. See some suggested guidelines by Brookhart et al (2010), Swanson and Hernán (2013), and Baiocchi et al. (2014).

Causal inference relies on transparency of assumptions and on triangulation of results from methods that depend on different sets of assumptions. IV estimation is therefore an attractive approach because it depends on a different set of assumptions than other methods. However, because of the wide 95% confidence intervals typical of IV estimates, the value added by using this approach will often be small. Also, users of IV estimation need to be critically aware of the limitations of the method. While this statement obviously applies to any causal inference method, the potentially counterintuitive direction and magnitude of bias in IV estimation requires especial attention.

Technical Point 16.7

**Monotonicity and the effect in the compliers.** Consider a dichotomous causal instrument $Z$, like the randomization indicator described in the text, and treatment $A$. Imbens and Angrist (1994) proved that the usual IV estimand equals the average causal effect in the compliers $\mathrm{E}\left[Y^{a=1} - Y^{a=0} | A^{z=1} - A^{z=0} = 1\right]$ under monotonicity (iv), i.e., when no defiers exist. Baker and Lindeman (1994) had a related proof for a binary outcome. See also Angrist, Imbens, and Rubin (1996), and the associated discussion, and Baker, Kramer, and Lindeman (2016). A proof follows.

The effect of treatment assignment (the intention-to-treat effect) can be written as the weighted average of the intention-to-treat effects in the four principal strata:

$$\begin{aligned}
\mathrm{E}\left[Y^{z=1} - Y^{z=0}\right] = {} & \mathrm{E}\left[Y^{z=1} - Y^{z=0} | A^{z=1} = 1, A^{z=0} = 1\right] \Pr\left[A^{z=1} = 1, A^{z=0} = 1\right] && \text{(always-takers)} \\
& + \mathrm{E}\left[Y^{z=1} - Y^{z=0} | A^{z=1} = 0, A^{z=0} = 0\right] \Pr\left[A^{z=1} = 0, A^{z=0} = 0\right] && \text{(never-takers)} \\
& + \mathrm{E}\left[Y^{z=1} - Y^{z=0} | A^{z=1} = 1, A^{z=0} = 0\right] \Pr\left[A^{z=1} = 1, A^{z=0} = 0\right] && \text{(compliers)} \\
& + \mathrm{E}\left[Y^{z=1} - Y^{z=0} | A^{z=1} = 0, A^{z=0} = 1\right] \Pr\left[A^{z=1} = 0, A^{z=0} = 1\right] && \text{(defiers)}
\end{aligned}$$

However, the intention-to-treat effect in both the always-takers and the never-takers is zero, because $Z$ does not affect $A$ in these two strata and, by individual-level condition (ii) of Technical Point 16.1, $Z$ has no independent effect on $Y$. If we assume that no defiers exist, then the above sum is simplified to

$$\mathrm{E}\left[Y^{z=1} - Y^{z=0}\right] = \mathrm{E}\left[Y^{z=1} - Y^{z=0} | A^{z=1} = 1, A^{z=0} = 0\right] \Pr\left[A^{z=1} = 1, A^{z=0} = 0\right] \qquad \text{(compliers)}.$$

But, in the compliers, the effect of $Z$ on $Y$ equals the effect of $A$ on $Y$ (because $Z = A$), that is $\mathrm{E}\left[Y^{z=1} - Y^{z=0} | A^{z=1} = 1, A^{z=0} = 0\right] = \mathrm{E}\left[Y^{a=1} - Y^{a=0} | A^{z=1} = 1, A^{z=0} = 0\right]$. Therefore, the effect in the compliers is

$$\mathrm{E}\left[Y^{a=1} - Y^{a=0} | A^{z=1} = 1, A^{z=0} = 0\right] = \frac{\mathrm{E}\left[Y^{z=1} - Y^{z=0}\right]}{\Pr\left[A^{z=1} = 1, A^{z=0} = 0\right]}$$

which is the usual IV estimand if we assume that $Z$ is randomly assigned, as random assignment implies $Z \perp\!\!\!\perp \{Y^{a,z}, A^z;\ z = 0, 1; a = 0, 1\}$. Under this joint independence and consistency, the intention-to-treat effect $\mathrm{E}\left[Y^{z=1} - Y^{z=0}\right]$ in the numerator equals $\mathrm{E}[Y|Z=1] - \mathrm{E}[Y|Z=0]$, and the proportion of compliers $\Pr\left[A^{z=1} = 1, A^{z=0} = 0\right]$ in the denominator equals $\Pr[A=1|Z=1] - \Pr[A=1|Z=0]$. To see why the latter equality holds, note that the proportion of always-takers $\Pr\left[A^{z=0} = 1\right] = \Pr[A=1|Z=0]$ and the proportion of never-takers $\Pr\left[A^{z=1} = 0\right] = \Pr[A=0|Z=1]$. Since, under monotonicity (iv), there are no defiers, the proportion of compliers $\Pr\left[A^{z=1} - A^{z=0} = 1\right]$ is the remainder $1 - \Pr[A=1|Z=0] - \Pr[A=0|Z=1] =$
$1 - \Pr[A=1|Z=0] - (1 - \Pr[A=1|Z=1]) = \Pr[A=1|Z=1] - \Pr[A=1|Z=0]$, which completes the proof.

The above proof only considers the setting depicted in Figure 16.1 in which the instrument $Z$ is causal. When, as depicted in Figures 16.2 and 16.3, data on a surrogate instrument $Z$—but not on the causal instrument $U_Z$—are available, Hernán and Robins (2006b) proved that the average causal effect in the compliers (defined according to $U_Z$) is also identified by the usual IV estimator. Their proof depends critically on two assumptions: that $Z$ is independent of $A$ and $Y$ given the causal instrument $U_Z$, and that $U_Z$ is binary. However, this independence assumption has often little substantive plausibility unless $U_Z$ is continuous. A corollary is that the interpretation of the IV estimand as the effect in the compliers is questionable in many applications of IV methods to observational data in which $Z$ is at best a surrogate for $U_Z$.

Fine Point 16.3

**Regression discontinuity design.** Suppose we are interested in the effect of a new antiviral treatment $A$ on oxygen levels $Y$, a continuous outcome measured 1 week later. The treatment is indicated for anyone who arrives at the hospital with a diagnosis of COVID-19. However, because the treatment is in short supply, the health authorities prohibit administering the treatment to people under age 65 to guarantee that everybody aged 65 years and older receives it. That is, the probability of receiving treatment $\Pr[A = 1|L < 65] = 0$ and $\Pr[A = 1|L \geq 65] = 1$ where $L$ is age. There is no positivity: the treated and the untreated do not have overlapping values of the confounder $L$.

In the absence of positivity, we need to make alternative assumptions to identify the causal effect. A reasonable assumption is that the conditional means of the counterfactual outcomes given $L$, $\mathrm{E}\left[Y^{a=1}|L\right]$ and $\mathrm{E}\left[Y^{a=0}|L\right]$, are continuous in $L$. In other words, if we could plot these means along the age axis (we can't because the means are counterfactual are thus unobserved), we would not observe any jumps in the lines. Under this continuity assumption, together with the exchangeability assumption that individuals close to both sides of the threshold are comparable, a discontinuity in the conditional mean of the observed mean given $L$, $\mathrm{E}[Y|L]$, around $L = 65$ could be interpreted as a consequence of the probability of treatment changing abruptly at age 65. Whether the mean of $Y$ jumps at the threshold can be empirically checked by plotting the observed data. (Strictly speaking, we only need continuity around the threshold $L = 65$ for our purposes.)

Therefore, under the continuity assumption, we could estimate an average causal effect of $A$ as the difference between the mean outcome $Y$ in individuals immediately above the threshold (say, those aged 65 years and 1 month) and the mean outcome in individuals immediately below the threshold (say, those aged 64 years and 11 months). If a bandwidth of 1 month around the threshold is too small (because too few individuals in the data are in that range), we would need to increase the bandwidth around the threshold. For example, we could use a bandwidth of 1 year by comparing individuals aged 64 versus individuals aged 65. The choice of the bandwidth is critical: wide intervals of age may introduce bias by comparing individuals who are not exchangeable. Once the bandwidth is fixed, we fit linear regression models on both sides of the threshold $L = 65$ to estimate the mean outcome on each side of the threshold. To help determine the bandwidth around the threshold, one can use data-adaptive procedures such as cross-validation (see Fine Point 18.2). Also, the regression model can include covariates if that is considered necessary to achieve conditional exchangeability.

The method described above is known as a *regression discontinuity design*, which was first proposed by Thistlewaite and Campbell (1960). It can be used when a single covariate $L$ is used to assign treatment, under the continuity assumption that the relation between $L$ and $Y$ is smooth (i.e., no jumps). A regression discontinuity design estimates the average causal effect of treatment $A$ on outcome $Y$ in the subset of the population with values of $L$ close to the threshold. This conditional effect may differ from the average causal effect in the population if $L$ is an effect modifier. Note that a regression discontinuity design will result in biased estimates of the conditional effect if treatments other than $A$ also change around the threshold (e.g., if health authorities also restrict the use of scarce intensive care units to people aged 65 and older) or if high-risk individuals aware of the threshold manipulate their own data (e.g., if at risk individuals aged 63 and 64 find a way to provide fake documentation that shows an older age).

More specifically, we have described here a *sharp regression discontinuity design* in which the probability of treatments jumps from $0$ to $1$ at the threshold. A *fuzzy regression discontinuity design* is an extension of the method that allows the jump in the probability of treatment from a value greater than 0 to a value less than 1. This extension, which relies on the monotonicity assumption, estimates the average causal effect in a subset of a subset of the population: the compliers with values of $L$ close to the threshold. For estimation details see Hahn, Todd and van der Klaauw (2001) and Imbens and Lemieux (2008).