Chapter 19

More advanced topics in causal inference

Chapter 18 described causal inference strategies for non-experimental data that assume ignorability of the treatment assignment mechanism. It is reasonable to be concerned about this assumption, however. After all, when are we really confident that we have measured *all* confounders? This chapter explores several alternative causal inference strategies that rely on slightly different sets of assumptions that may be more plausible in certain (less commonly observed) settings. These studies all require specific study designs or data requirements. The chapter concludes with a look at another approach to dealing with dissatisfaction with the ignorability assumption—methods that assess the sensitivity of our treatment effect estimates to violations of ignorability.

19.1 Estimating causal effects indirectly using instrumental variables

In some situations when the argument for ignorability of the treatment assignment seems weak, there may exist another variable which does appear to be randomly assigned (or conditionally randomly assigned). If this variable, called the *instrument*, is predictive of the treatment, we *may* be able to use it to isolate a particular kind of targeted causal estimand.

Example: a randomized-encouragement design

Suppose we wanted to estimate the effect of watching an educational television program (this time the program is Sesame Street) on letter recognition. We might consider implementing a randomized experiment where the participants are preschool children, the treatment of interest is watching Sesame Street, the control condition is not watching, and the outcome is the score on a test of letter recognition. Actually the researchers in this study recorded four viewing categories: (1) rarely watched, (2) watched once or twice a week, (3) watched 3-5 times a week, and (4) watched more than 5 times a week on average. Since there is not a category for "never watched," for the purposes of this illustration we treat the lowest viewing category ("rarely watched") as if it were equivalent to "never watched." [1]

In any case, it is not possible for an experimenter to force children to watch a TV show or to refrain from watching (the experiment took place while Sesame Street was on the air). Thus *watching* cannot be randomized. Instead, when this study was actually performed, what was randomized was *encouragement* to watch the show—this is called a *randomized encouragement design*.

A simple comparison of outcomes across randomized groups in this study will yield an estimate of the effect of *encouraging* these children to watch the show, not an estimate of the effect of actually viewing the show. This estimand is often referred to as the *intention-to-treat* (ITT) effect. However, we may be able to take advantage of the randomization to estimate a causal effect of watching for at least some of the people in the study by using the randomized

[1] Data and code for this example appear in the folder `Sesame`.

encouragement as an *instrument* that randomly induces variation in the treatment variable of interest.

Conceptual framework

We begin by describing a conceptual framework that will help us to understand the estimand targeted by this approach as well as the required assumptions. This requires conceptualizing a categorization of the children who participated in the study by their compliance behavior (that is, the extent to which their viewing behavior matched what they were encouraged to do).

A critical feature of this type of study is that only some of the children's viewing patterns were affected by the encouragement. Those children whose viewing patterns could be altered by encouragement are the only participants in the study for whom we can conceptualize counterfactuals with regard to viewing behavior since under different experimental conditions they might have been observed either viewing or not viewing. Therefore, a comparison of the potential outcomes for these children makes sense. Following the conventions of the statistics literature on instrumental variables, we shall label these children "compliers." These are the only children for whom we will make inferences about the effect of watching Sesame Street and the effect for them is referred to as the *complier average causal effect* (CACE).

What other types of children exist in the study? We know that there were children who were encouraged to watch but did not; we might plausibly assume that these children also would not have watched if not encouraged. A child who would not watch regardless of his assignment is labeled a "never taker." We cannot directly estimate the effect of viewing for these children since in this context they would never be observed watching the show. Similarly, for the children who watched Sesame Street even though not encouraged, we might plausibly assume that if they had been encouraged they would have watched as well. Again these children cannot shed light on the effect of viewing, since all of them are regular viewers of the program so there is no way of making comparisons between outcomes viewing conditions. We shall label those children who would watch whether encouraged or not "always takers." The assumptions made above in motivating the profiles of the never takers and always takers ruled out children who would always do the opposite of what they were told, so-called "defiers"; this assumption will be formalized below.

Assumptions for instrumental variables estimation

Instrumental variables analyses rely on several key assumptions, one combination of which we discuss in this section in the context of a simple example with binary treatment and binary instrument:

- Ignorability of the instrument,
- Monotonicity,
- Nonzero association between instrument and treatment variable,
- Exclusion restriction.

We discuss each in more detail below. The model additionally assumes no interference between units (the stable unit treatment value assumption) as with most other causal analyses, an issue we have already discussed at the end of Section 17.3.

Ignorability of the instrument. The first assumption in the list above is *ignorability of the instrument* with respect to the potential outcomes (both for the primary outcome of interest and the treatment variable). Formally we can write this as $y(0), y(1) \perp z$. This is trivially satisfied in a randomized experiment (assuming the randomization was pristine) assuming, as always, that any design features are reflected in the analysis (for example, block

indicators would need to be included in the analysis of data arising from a randomized block design). In the absence of a randomized experiment (or natural experiment) this property may be more difficult to satisfy and often requires conditioning on other pre-treatment variables (potential confounders). To the extent that this assumption is difficult to justify, any advantage over a traditional observational study (relying on ignorability of the treatment assignment mechanism) is generally lost.

Monotonicity. In defining never takers and always takers, we assumed that there were no children who would watch if they were not encouraged but who would *not* watch if they *were* encouraged; that is, we assumed that there were no "defiers." Formally this is called the *monotonicity assumption*, and it will not necessarily hold in practice, though there are many situations in which it is defensible. In particular, in some studies it would be impossible for study participants in the non-encouraged group to gain access to the treatment of interest. In those situations neither defiers nor always takers are possible.

Nonzero association between instrument and treatment variable. To demonstrate how we can use the instrument to obtain a causal estimate of the treatment effect in our example, first consider that about 90% of those encouraged watched the show regularly; by comparison, only 55% of those not encouraged watched the show regularly. Therefore, if we are interested in the effect of actually viewing the show, we should focus on the 35% of the treatment population who decided to watch the show because they were encouraged but who otherwise would not have watched the show. If the instrument (encouragement) did not affect regular watching, then we could not proceed. Although a nonzero association between the instrument and the treatment is an assumption of the model, fortunately this assumption is empirically verifiable.

Exclusion restriction. To estimate the effect of viewing for those children whose viewing behavior would have been affected by the encouragement (the induced watchers), we must make another important assumption, called the *exclusion restriction*. This assumption says for those children whose behavior would not have been changed by the encouragement (never takers and always takers) there is no effect of encouragement on outcomes. So for the never takers (children who would not have watched either way), for instance, we assume encouragement to watch did not affect their outcomes. And for the always takers (children who would have watched either way), we assume encouragement to watch did not affect their outcomes. Technically, the assumptions regarding always takers and never takers represent distinct exclusion restrictions. In this simple framework, however, the analysis suffers if either assumption is violated. Using more complicated estimation strategies, it can be helpful to consider these assumptions separately as it may be possible to weaken one or the other or both.

It is not difficult to tell a story that violates the exclusion restriction. Consider, for instance, the conscientious parents who do not let their children watch television and are concerned with providing their children with a good start educationally. The materials used to encourage them to have their children watch Sesame Street for its educational benefits might instead have motivated them to purchase other types of educational materials for their children or to read to them more often.

To illustrate the instrumental variables approach, however, we proceed as if the exclusion restriction were true (or at least approximately true). In this case, if we think about individual-level causal effects, the answer becomes relatively straightforward.

Derivation of instrumental variables estimation with complete data (including unobserved potential outcomes)

Figure 19.1 illustrates with hypothetical data displaying for each study participant not only the observed data (encouragement and viewing status as well as observed outcome test score)

Unit, i	Potential viewing outcomes, T_i^0	T_i^1		Encouragement indicator, z_i	Potential test outcomes, y_i^0	y_i^1	Encouragement effect, $y_i^1 - y_i^0$
1	**0**	1	(complier)	0	**67**	76	9
2	**0**	1	(complier)	0	**72**	80	8
3	**0**	1	(complier)	0	**74**	81	7
4	**0**	1	(complier)	0	**68**	78	10
5	**0**	0	(never taker)	0	**68**	68	0
6	**0**	0	(never taker)	0	**70**	70	0
7	**1**	1	(always taker)	0	**76**	76	0
8	**1**	1	(always taker)	0	**74**	74	0
9	**1**	1	(always taker)	0	**80**	80	0
10	**1**	1	(always taker)	0	**82**	82	0
11	0	**1**	(complier)	1	67	**76**	9
12	0	**1**	(complier)	1	72	**80**	8
13	0	**1**	(complier)	1	74	**81**	7
14	0	**1**	(complier)	1	68	**78**	10
15	0	**0**	(never taker)	1	68	**68**	0
16	0	**0**	(never taker)	1	70	**70**	0
17	1	**1**	(always taker)	1	76	**76**	0
18	1	**1**	(always taker)	1	74	**74**	0
19	1	**1**	(always taker)	1	80	**80**	0
20	1	**1**	(always taker)	1	82	**82**	0

Figure 19.1 *Hypothetical complete data in a randomized encouragement design. Units have been ordered for convenience. For each unit, the students are encouraged to watch Sesame Street ($z_i = 1$) or not ($z_i = 0$). This reveals which of the potential viewing outcomes (T_i^0, T_i^1) and which of the potential test outcomes (y_i^0, y_i^1) we get to observe. The observed outcomes are displayed in boldface. Here, potential outcomes are what we would observe under either encouragement option. The exclusion restriction forces the potential outcomes to be the same for those whose viewing would not be affected by the encouragement. The effect of watching for the "compliers" is equivalent to the intent-to-treat effect (encouragement effect over the whole sample) divided by the proportion induced to view; thus,* $3.4/0.4 = 8.5$. *The researcher cannot calculate this effect directly because he cannot see both potential outcomes. Moreover, true compliance class is known for only some individuals and even then only under the monotonicity assumption.*

but also the unobserved categorization, c_i, into always taker, never taker, or complier, based on potential watching behavior as well as the counterfactual test outcomes (the potential outcome corresponding to the treatment not received). Here, potential outcomes are the outcomes we would have observed under either *encouragement* option. Because of the exclusion restriction, for the always takers and the never takers the potential outcomes are the same no matter the encouragement (really they need not be *exactly* the same, just distributionally the same, but this simplifies the exposition).

The true intent-to-treat effect for these 20 observations is then an average of the effects for the 8 induced watchers, along with 12 zeroes corresponding to the encouragement effects for the always takers and never takers:

$$\begin{aligned}
\text{ITT} &= \frac{9+8+7+10+9+8+7+10+0+\cdots+0}{20} \\
&= 8.5 * \frac{8}{20} + 0 * \frac{12}{20} \\
&= 8.5 * 0.4 \\
&= 3.4.
\end{aligned}$$

The effect of watching Sesame Street for the compliers is 8.5 points on the letter recognition test. This is algebraically equivalent to the intent-to-treat effect (3.4) divided by the proportion of compliers ($8/20 = 0.40$).

Deconstructing the complier average causal effect

To better understand the implications of the assumptions we have made in order to identify this effect, let us back up for a minute and consider a more general formulation. We start by conceptualizing the intention-to-treat effect as a weighted average of four different ITT effects—one for each of the four types of compliance classifications. Informally we can write

$$\begin{aligned}
\text{ITT} = \; & \text{ITT}_{c=\text{complier}} \Pr(c = \text{complier}) + \text{ITT}_{c=\text{never taker}} \Pr(c = \text{never taker}) + \\
& + \text{ITT}_{c=\text{always taker}} \Pr(c = \text{always taker}) + \text{ITT}_{c=\text{defier}} \Pr(c = \text{defier}),
\end{aligned}$$

where $\text{ITT}_{c=\text{complier}}$ denotes the effect of the *instrument* on the outcomes for the compliers, and the other ITT effects in the expression are similarly defined.

The exclusion restriction sets $\text{ITT}_{c=\text{never taker}}$ and $\text{ITT}_{c=\text{always taker}}$ to 0 and the monotonicity assumption sets $\Pr(c = \text{defier})$ to 0, simplifying the expression to,

$$\text{ITT} = \text{ITT}_{c=\text{complier}} \Pr(c = \text{complier}),$$

which is easily rearranged to look like our complier average causal effect estimand,

$$\text{ITT}_{c=\text{complier}} = \text{CACE} = \frac{\text{ITT}}{\Pr(c = \text{complier})}.$$

Exercise XXXX explores the correspondence between $\Pr(c = \text{complier})$ and $\text{E}(T(1) - T(0))$ that allows this expression to be rewritten as,

$$\text{ITT}_{c=\text{complier}} = \text{CACE} = \frac{\text{ITT}}{\text{E}(T(z=1) - T(z=0))}, \tag{19.1}$$

where the expression in the denominator can be conceived of as the ITT effect of the instrument on the treatment variable, T. This expression provides another way of understanding the assumption that the instrument must have an effect on the treatment; that is, the denominator of the CACE estimand cannot be zero.

Violations of the ignorability assumption. The ignorability assumption is arguably the most essential to the instrumental variables framework presented here because it is required to unbiasedly estimate both the numerator and denominator of the estimand in (19.1). Violations of this assumption could lead to either positive or negative bias. This bias will be exacerbated in situations when the estimated proportion of compliers is small (because dividing the biased ITT estimate by a small number that is less than 1 is equivalent to multiplying it by a big number; see Exercise XXX for further exploration of this bias).

Violations of the exclusion restriction. What happens when the exclusion restriction is violated? Consider a scenario in which the effect of the instrument on the never takers is α. In this case our expanded ITT formula reduces to

$$\text{ITT} = \text{CACE} * \Pr(c = \text{complier}) + \alpha * \Pr(c = \text{never taker}),$$

and straightforward algebraic manipulations reveal that

$$\frac{\text{ITT}}{\Pr(c = \text{complier})} = \text{CACE} + \alpha \frac{\Pr(c = \text{never taker})}{\Pr(c = \text{complier})},$$

with the term on the far right representing the bias. This bias increases with the size of the effect of the instrument on the never takers, α. This effect is either shrunk or magnified depending on the ratio of never takers to compliers. Whenever the ratio of compliers to never takers is less the bias captured in the never taker effect, α, will be amplified. This revelation provides further caution against use of instruments that are not strongly predictive of the treatment as such studies will be highly vulnerable to violations of the exclusion restriction.

Violations of the monotonicity assumption. If the monotonicity assumption is violated, then $\Pr(c = \text{defier}) \neq 0$ and consequently the equivalence between $\Pr(c = \text{complier})$ and $\text{E}(T(1) - T(0))$ is lost. In this case the calculations to derive the bias are slightly more complicated so we omit the steps here; we present the result that the bias in this case looks like,

$$\text{bias} = \delta \left(\text{ITT}_{c=\text{complier}} + \text{ITT}_{c=\text{defier}} \right),$$

where

$$\delta = \frac{\Pr(c = \text{defier})}{\Pr(c = \text{complier}) - \Pr(c = \text{defier})}.$$

Somewhat oddly, this bias is driven by the *difference* between the intent-to-treat effects for the compliers and defiers. The ITT effect for a defier represents $y(T = 0) - y(T = 1)$ (since defiers, by definition, are the people who choose the opposite of their treatment assignment); whereas for compliers, who always do what they are told, the ITT effect represents $y(T = 1) - y(T = 0)$. Thus we can reframe this bias cancellation property by noting that the bias will disappear if the effect of the treatment (watching) on the outcome is the same for compliers and defiers. Any difference between the causal effect of the treatment on the outcomes however will be magnified if the proportion of defiers is high relative to the proportion of compliers. Again this points to the potential dangers of a weak instrument when there is a chance that an assumption will be violated.

Local average treatment effect (LATE) versus intent-to-treat effect (ITT)

As we have discussed, the instrumental variables strategy here does not estimate an overall causal effect of watching Sesame Street across everyone in the study, or even an effect for all those treated. The complier average causal effect (CACE) estimate applies only to those children whose treatment receipt is dictated by their randomized instrument assignment and is a special case of what is commonly called a *local average treatment effect* (LATE) by economists.

Some researchers argue that intent-to-treat effects are more interesting from a policy perspective because they accurately reflect that not all targeted individuals will participate in the intended program. However, the intent-to-treat effect only parallels a true policy effect if in the subsequent policy implementation the compliance rate remains unchanged. We recommend estimating both the intent-to-treat effect and the complier average causal effect to maximize what we can learn about the intervention.

Instrumental variables estimate: Sesame Street

We can calculate an estimate of the effect of watching Sesame Street for the compliers with the actual data using the same principles.

We first estimate the percentage of children actually induced to watch Sesame Street by the intervention, which is the coefficient on the instrument (`encouraged`), in the following regression:

```
fit_1a <- lm(watched ~ encouraged, data=sesame)
```
R code

The estimated coefficient of `encouraged` here is 0.36 (which, in this regression with a single binary predictor, is simply the proportion of compliers in the data).

We then compute the intent-to-treat estimate, obtained in this case using the regression of the outcome on the instrument:

```
fit_1b <- lm(y ~ encouraged, data=sesame)
```
R code

The estimated coefficient of `encouraged` in this regression is 2.9, which we then "inflate" by dividing by the percentage of children affected by the intervention:

```
iv_est <- coef(fit_1a)[,"encouraged"]/coef(fit_1b)[,"encouraged"]
```
R code

The estimated effect of regularly viewing Sesame Street is thus $2.9/0.36 = 7.9$ points on the letter recognition test. This type of ratio estimator is sometimes called a *Wald estimator*.

19.2 Instrumental variables in a regression framework

Instrumental variables models and estimators can also be derived by focusing on the mean structure of the regression model, allowing us to more easily extend the basic concepts discussed in the previous section. A general instrumental variables model with (potentially continuous) instrument, z, and (potentially continuous) treatment, T, can be written as,

$$\begin{aligned} y_i &= \beta_0 + \beta_1 T_i + \epsilon_i \\ T_i &= \gamma_0 + \gamma_1 z_i + \nu_i. \end{aligned} \tag{19.2}$$

The assumptions can now be expressed in a slightly different way. The first set of assumptions is that z_i is uncorrelated with both ϵ_i and ν_i; these translate informally into the ignorability assumption and exclusion restriction (expressed informally as, "the instrument only affects the outcome through its effect on the treatment"). The requirement that the correlation between z_i and T_i must be nonzero is explicit in both formulations. We next address how this framework identifies the causal effect of T on y.

Identifiability with instrumental variables

Generally speaking, *identifiability* refers to whether the data contain sufficient information for unique estimation of a given parameter or set of parameters in a particular model.

What if we did not impose the exclusion restriction for our basic model? The model (ignoring covariate information, and switching to mathematical notation for simplicity and generalizability) can be written as,

$$\begin{aligned} y &= \beta_0 + \beta_1 T + \beta_2 z + \text{error} \\ T &= \gamma_0 + \gamma_1 z + \text{error}, \end{aligned} \tag{19.3}$$

where y is the response variable, z is the instrument, and T is the treatment of interest. Our goal is to estimate β_1, the treatment effect. The difficulty is that T has not been randomly assigned; it is observational and, in general, can be correlated with the error in the first equation; thus we cannot simply estimate β_1 by fitting a regression of y on T and z.

However, as described in the previous section, we can estimate β_1 using instrumental variables. We derive the estimate here algebraically, in order to highlight the assumptions needed for identifiability.

Substituting the formula for T into the formula for y yields,

$$\begin{aligned} y &= \beta_0 + \beta_1 T + \beta_2 z + \text{error} \\ &= \beta_0 + \beta_1(\gamma_0 + \gamma_1 z) + \beta_2 z + \text{error} \\ &= (\beta_0 + \beta_1\gamma_0) + (\beta_1\gamma_1 + \beta_2)z + \text{error}. \end{aligned} \tag{19.4}$$

We now show how to estimate β_1, the causal effect of interest, using the slope of this regression, along with the regressions (19.3) and the exclusion restriction.

The first step is to express (19.4) in the form,

$$y = \delta_0 + \delta_1 z + \text{error}.$$

From this equation we need δ_1, which can be estimated from a simple regression of y on z. We can now solve for β_1 in the following equation:

$$\delta_1 = \beta_1\gamma_1 + \beta_2,$$

which we can rearrange to get,

$$\beta_1 = (\delta_1 - \beta_2)/\gamma_1. \tag{19.5}$$

We can directly estimate the denominator of this expression, γ_1, from the regression of T on z in (19.3)—this is not a problem since we are assuming that the instrument, z, is randomized.

The only challenge that remains in estimating β_1 from (19.5) is to estimate β_2, which in general cannot simply be estimated from the top equation of (19.3) since, as already noted, the error in that equation can be correlated with T. However, under the exclusion restriction, we know that β_2 is zero, and so $\beta_1 = \delta_1/\gamma_1$, leaving us with the standard instrumental variables estimate.

Other models. There are other ways to achieve identifiability in this two-equation setting. Approaches such as selection correction models rely on functional form specifications to identify the causal effects even in the absence of an instrument. For example, a probit specification could be used for the regression of T on z. The resulting estimates of treatment effects are often unstable if a true instrument is not included as well.

Two-stage least squares: Sesame Street

The Wald estimate discussed in the previous section can be used with this formulation of the model as well. We now describe a more general estimation strategy, *two-stage least squares*.

To illustrate we return to our Sesame Street example. The first step is to regress the "treatment" variable—an indicator for regular watching (`watched`)—on the randomized instrument, encouragement to watch (`encouraged`). Then we plug predicted values of `watched` into the equation predicting the letter recognition outcome, `y`:

R code
```
fit_2a <- lm(watched ~ encouraged, data=sesame)
sesame$watched_hat_2 <- fit_2a$fitted
fit_2b <- lm(y ~ watched_hat_2, data=sesame)
```

The result is,

R output
```
             coef.est coef.se
(Intercept)      20.6     3.9
watched_hat_2     7.9     4.9
  n = 240, k = 2
  residual sd = 13.3, R-Squared = 0.01
```

where the coefficient on `watched_hat_2` is the estimate of the causal effect of watching Sesame Street on letter recognition for those induced to watch by the experiment. This two-stage estimation strategy is especially useful for more complicated versions of the model, for instance, when multiple instruments are included.

This second-stage regression does not give the correct standard error, however, as we discuss on page 395.

Adjusting for covariates in an instrumental variables framework

It turns out that the randomization for this particular experiment took place within sites and settings; it is therefore appropriate to adjust for these covariates in estimating the treatment effect. Additionally, pre-test scores are available that are highly predictive of post-test scores. Our preferred model would adjust for all of these predictors. We can calculate the same ratio (intent-to-treat effect divided by effect of encouragement on viewing) as before using models that include these additional predictors but pulling out only the coefficients on `encouraged` for the ratio.

Here we equivalently perform this analysis using two-stage least squares:

R code

```
fit_3a <- lm(watched ~ encouraged + pretest + as.factor(site) + setting,
  data=sesame)
watched_hat_3 <- fit_3a$fitted
fit_3b <- lm(y ~ watched_hat_3 + pretest + as.factor(site) + setting, data=sesame)
display(fit_3b)
```

yielding,

R output

```
                 coef.est coef.se
(Intercept)           1.2     4.8
watched_hat_3        14.0     4.0
pretest               0.7     0.1
as.factor(site)2      8.4     1.8
as.factor(site)3     -3.9     1.8
as.factor(site)4      0.9     2.5
as.factor(site)5      2.8     2.9
setting               1.6     1.5
  n = 240, k = 8
  residual sd = 9.7, R-Squared = 0.49
```

The estimated effect of watching Sesame Street on the compliers is about 14 points on the letter recognition test. Again, we do not trust this standard error and will discuss later how to appropriately adjust it for the two stages of estimation.

Since the randomization took place within each combination of site (five categories) and setting (two categories), it would be appropriate to interact these variables in our equations. Moreover, it would probably be interesting to estimate variation of effects across sites and settings. However, for simplicity of illustration (and also due to the complication that one site $\times$ setting combination has no observations) we only include main effects for this discussion.

Standard errors for instrumental variables estimates

The second step of two-stage regression yields the instrumental variables estimate, but the standard-error calculation is complicated because we cannot simply look at the second regression in isolation. We show here how to adjust the standard error to account for the uncertainty in both stages of the model. We illustrate with the model we have just fitted.

The regression of compliance on treatment and other covariates (model `fit_3a`) is unchanged. We then regress the outcome on predicted compliance and covariates, this time

saving the predictor matrix from this second-stage regression, which we do using the `x=TRUE` option:

R code
```
fit_3b <- lm(y ~ watched_hat_3 + pretest + as.factor(site) + setting, x=TRUE,
  data=sesame)
```

We next compute the standard deviation of the adjusted residuals, $r_i^{\text{adj}} = y_i - X_i^{\text{adj}}\hat{\beta}$, where X^{adj} is the predictor matrix from `fit_3b` but with the column of predicted treatment values replaced by observed treatment values:

R code
```
X_adj <- fit_3b$x
X_adj[,"watched_hat_3"] <- watched
residual_sd_adj <- sd(y - X_adj %*% coef(fit_3b))
```

Finally, we compute the adjusted standard error for the two-stage regression estimate by taking the standard error from `fit_3b` and scaling by the adjusted residual standard deviation, divided by the residual standard deviation from `fit_3b` itself:

R code
```
se_adj <-se.coef(fit_3b)["watched_hat_3"]*residual_sd_adj / sigma.hat(fit_3b)
```

So the adjusted standard errors are calculated as the square roots of the diagonal elements of $(X^tX)^{-1}\hat{\sigma}^2_{\text{TSLS}}$ rather than $(X^tX)^{-1}\hat{\sigma}^2$, where $\hat{\sigma}$ is the residual standard deviation from `fit_3b` and $\hat{\sigma}_{\text{TSLS}}$ is calculated using the residuals from an equation predicting the outcome from `watched` (not `watched_hat_3`) using the two-stage least squares estimate of the coefficient, not the coefficient that would have been obtained in a least squares regression of the outcome on `watched`).

The resulting standard error for our example is 3.9, which is actually a bit smaller than the unadjusted standard error (which is not unusual for these corrections).

Performing two-stage least squares automatically using the tsls function

We have illustrated the key concepts in our instrumental variables discussion using basic R commands with which you were already familiar so that the steps were transparent. There is, however, a package available in R called `sem` that has a function, `tsls()`, that automates this process. Or use the ivreg() in AER package

To calculate the effect of regularly watching Sesame Street on post-treatment letter recognition scores using encouragement as an instrument, we specify both equations:

R code
```
iv1 <- tsls(postlet ~ watched, ~ encour, data=sesame)
display(iv1)
```

where in the second equation it is assumed that the "treatment" (in econometric parlance, the *endogenous* variable) for which `encour` is an instrument is whatever predictor from the first equation that is not specified as a predictor in the second. Fitting and displaying the two-stage least squares model yields,

R output
```
             Estimate  Std. Error
(Intercept)   20.6       3.7
watched        7.9       4.6
```

To incorporate other pre-treatment variables as controls, we must include them in both equations; for example,

R code
```
iv2 <- tsls(postlet ~ watched + prelet + as.factor(site) + setting,
  ~ encour + prelet + as.factor(site) + setting, data=sesame)
display(iv2)
```

yielding,

R output

```
                  Estimate  Std. Error
(Intercept)          1.2       4.6
watched             14.0       3.9
prelet               0.7       0.1
as.factor(site)2     8.4       1.8
as.factor(site)3    -3.9       1.7
as.factor(site)4     0.9       2.4
as.factor(site)5     2.8       2.8
setting              1.6       1.4
```

The point estimate of the treatment effect calculated this way is the same as with the preceding step-by-step procedure, and standard errors are also provided.

More than one treatment variable; more than one instrument

In the experiment discussed in Section 18.6, the children randomly assigned to the intervention group received several services ("treatments") that the children in the control group did not receive, most notably, access to high-quality child care and home visits from trained professionals. Children assigned to the intervention group did not make full use of these services. Simply conceptualized, some children participated in the child care while some did not, and some children received home visits while others did not. Can we use the randomization to treatment or control groups as an instrument for these two treatments? The answer is no.

Similar arguments as those used in Section 19.2 can be given to demonstrate that a single instrument cannot be used to identify more than one treatment variable (see Exercise XXX). In fact, as a general rule, we need to use at least as many instruments as treatment variables in order for all the causal estimates to be identifiable. A model in which the number of instruments is equal to the number of treatment variables is called a just-identified model. A model in which the number of instruments exceeds the number of treatment variables is called an overidentified model. We discuss issues with overidentified models in the section below on weak instruments.

Continuous treatment variables or instruments

When using two-stage least squares, the models we have discussed can easily be extended to accommodate continuous treatment variables and instruments, although at the cost of complicating the interpretation of the causal effects.

Researchers must be careful, however, in the context of binary instruments and continuous treatment variables. A binary instrument cannot without further assumptions identify a continuous treatment or dosage effect. If we map this back to a randomized experiment, the randomization assigns someone only to be encouraged or not. This is equivalent to a setting with many different treatments (one at each dosage level) but only one instrument—therefore causal effects for all these treatments are not identifiable without further assumptions. To identify such dosage effects without making further parametric assumptions, one would need to randomly assign encouragement levels corresponding to the different dosages or levels of participation.

Have we really avoided the ignorability assumption?

We have motivated instrumental variables using the cleanest setting, within a controlled, randomized experiment. The drawback of illustrating instrumental variables using this example is that it de-emphasizes one of the most important assumptions of the instrumental variables model, *ignorability of the instrument*. In the context of a randomized experiment,

this assumption should be trivially satisfied (assuming the randomization was pristine). However, in practice an instrumental variables strategy may also be reasonable in the context of a *natural experiment*, that is, an observational study context in which a "randomized" variable or instrument appears to have occurred naturally. Examples of this include:

- The weather in New York as an instrument for estimating the effect of supply of fish on their price,
- The sex mix of the first two children in a family (in an analysis of people who have at least two children) as an instrument for estimating the effect of number of children on labor supply.

In these examples we have traded one ignorability assumption (ignorability of the treatment variable) for another (ignorability of the instrument) that we believe to be more plausible. Additionally, we must assume monotonicity and the exclusion restriction.

In the absence of a controlled randomized experiment or a natural experiment, instrumental variables strategies are more suspect. While we cannot confirm the existence of ignorability in the absence of planned or naturally occurring randomization we may be able to rule it out by performing the types of balance checks we did in the propensity score matching context. Or we may believe that ignorability holds only conditional on observed covariates in which case we should adjust for these in our analysis. Broadly speaking, if the ignorability assumption is not highly plausible the expected gains from performing an instrumental variables analysis are not likely to outweigh the potential for bias.

Plausibility of exclusion restriction

One way to assess the plausibility of the exclusion restriction is to calculate an estimate within a sample that would not be expected to be affected by the instrument. For instance, researchers estimated the effect of military service on earnings (and other outcomes) using, as an instrument, the lottery number for young men eligible for the draft during the Vietnam War. This number was assigned randomly and strongly affected the probability of military service. It was hoped that the lottery number would only have an effect on earnings for those who served in the military only because they were drafted (as determined by a low enough lottery number). Satisfaction of the exclusion restriction is not certain, however, because, for instance, men with low lottery numbers may have altered their educational plans so as to avoid or postpone military service. So the researchers also ran their instrumental variables model for a sample of men who were assigned numbers so late that the war ended before they ever had to serve. This showed no clear relation between lottery number and earnings, which provides some support for the exclusion restriction.

Strength of the relationship between the instrument and the treatment: weak instruments

Our assumptions require that the instrument have non-zero correlation with the treatment variable. This is the only assumption that can be directly tested. Thus the "first stage" model (the model that predicts the treatment using the instrument) should be examined closely to assess the strength of the instrument. As discussed in Section 19.1, a weak instrument can exacerbate the bias that can result from failure to satisfy the ignorability or monotonicity assumptions or the exclusion restriction. The weaker the instrument, the smaller the proportion of compliers, the greater the potential for bias if one of these assumptions is violated.

In addition there is the concern that the compliers—the group about which the data is informative regarding causal effects—may not be particularly representative of the full population of interest. This is a form of sampling bias that is subtle because we cannot

directly observe whether individuals are compliers; we can only use statistical methods to infer characteristics of compliers and noncompliers in the sample and the population.

It is also important to ensure that the association between the instrument and treatment is *positive*. If the association is not in the expected direction, this might be the result of a mixture of two different mechanisms, the expected process and one operating in the opposite direction, which could in turn imply a violation of the monotonicity assumption.

Structural equation models

A goal in many areas of social science is to infer causal relations among many variables, a generally difficult problem (as discussed in Section 17.7). *Structural equation modeling* is a family of methods of multivariate data analysis that are sometimes used for causal inference. The same statistical model is also used to estimate latent factors in noncausal regression settings with multiple inputs and sometimes multiple outcome variables.

In the causal setting, structural equation modeling relies on conditional independence assumptions in order to identify causal effects, and the resulting inferences can be sensitive to strong parametric assumptions such as linear relationships and multivariate normality of errors. Instrumental variables can be considered as a special case of a structural equation model where certain dependencies are set to be exactly zero. As we have just discussed, even in a relatively simple instrumental variables model, the assumptions needed to identify causal effects are difficult to satisfy and largely untestable. A structural equation model that tries to estimate many causal effects at once multiplies the number of assumptions required with each desired effect so that it quickly becomes difficult to justify all of them. Therefore we do not discuss the use of structural equation models for causal inference in any greater detail here. We have no objection to complicated models, but we are cautious about attempting to estimate complex causal structures from observational data.

19.3 Regression discontinuity: known assignment mechanism but no overlap

We motivated this chapter by saying that often the assumption of ignorability is not plausible, where it does not make sense to assume that treatment assignment depends only on observed pre-treatment predictors. However we can design observational studies for which the assignment mechanism is entirely known (as with a randomized experiment) but that involves no explicit randomization. As an example, suppose we wanted to evaluate the effect of an afterschool tutoring program for fourth graders, but the school district in which the program would be implemented felt that it was important to give priority when providing these services to the children who were struggling the most academically. One way to balance the desire for equity but also provide a reasonable design for evaluating the effectiveness of the program would be to assign children to receive the program or not based on the students' scores on a pre-test designed to measure the relevant skills. For instance, if there were funding for 100 students to participate and there were 1000 students enrolled in fourth grade in the district the program administrators would simply choose the 100 students with the *lowest* scores on the pre-test to participate in the tutoring program. Those with higher scores would not be allowed access.

Suppose that the 100 lowest pre-test scores were all below 60 and the rest of the scores were above 60. In the purest form of this design, the assignment to the treatment, T, is completely deterministic, specifically: $\Pr(T = 1 \,|\, \text{pre-test} < 60) = 1$ and $\Pr(T = 1 \,|\, \text{pre-test} > 60) = 0$. Another way of framing this property is that the pre-test score is our only confounding covariate. This sounds like a randomized block experiment, doesn't it? A randomized block experiment with one blocking variable also has a known assignment mechanism with one confounding covariate. What's the difference?

In a randomized block experiment, observations are randomly assigned to treatment

groups within subclasses defined by the blocking groups. Since the probability of assignment depends only on this variable, it is our only confounder. Within each block the potential outcomes will be balanced (in distribution) across treatment groups. If the probability of assignment to the treatment varies substantially across blocks the blocking variable (our only confounder) will be imbalanced across treatment groups. However, there is a requirement with a randomized block design that the probability of assignment to treatment can never be 0 or 1. The implication of this design feature is that there will always be overlap across treatment groups with respect to this crucial confounder. See Exercise 19.** for a numerical example of these features of randomized block and regression discontinuity designs.

Contrast this scenario with that of the regression discontinuity design. By construction the pure form of this design leads to a situation in which there is *absolutely no overlap* across treatment groups with respect to our only confounder, the "assignment variable" or "forcing variable," the pre-test score; there must also be substantial imbalance in these distributions due to the lack of overlap. These features stand in stark contrast with the scenarios in the previous chapter in which we worked so hard to assure balance and overlap. With a pure regression discontinuity design there is no hope of achieving either balance or overlap. How then do we proceed to make causal inferences?

One way to conceptualize making progress is by considering the students whose scores lie just below and just above the cutoff of 60. For instance, if the standard deviation on the test were 15, would we really think that students who received scores of 57 were different in important ways than those who received scores of 63? More precisely, would we expect that students in this range of pre-test scores would have different potential test score outcomes? If not, then maybe we could consider the data from students with scores in this range to be nearly equivalent to data that were generated by a randomized experiment. In practice we might not trust this assumption too closely and would do well to adjust for our pre-test as well. Although we know that it can be dangerous to use models for data with limited (in this case no) overlap, if we limit our analysis to a narrow enough range of the data, a smooth model can make sense. At the same time, when we limit the range of pre-tests we limit the number and type of students about whom we can make inferences. We could use a broader window of scores, but the wider the range of pre-test scores considered, the more we will have to worry about the model extrapolations necessary to perform causal inference. So a tension exists between the range of values we are willing to consider for the assignment variable and the size of the inferential group.

Example: The effect of an educational program on test scores in Chile

We illustrate some of the key features of the regression discontinuity design using an example of data that arose from a discontinuity that could be said to be naturally occurring, in that no one prospectively designed this study to address the research question at hand; rather, researchers retrospectively realized that data that were generated from a policy that mimics such a design. In the regression discontinuity framework, these data can be used to estimate the effect of of an educational program in Chile on subsequent school-level test scores.

The Chilean government introduced the "900 schools program" (P-900) in 1990 in an effort to increase the performance of struggling public schools by providing resources in four different areas. In 1990 and 1991 the focus was on infrastructure improvements and instructional materials. Starting in 1992 the focus shifted to teacher training and after-school tutoring. In an effort to target the neediest schools, the government provided these resources only to schools whose mean fourth grade test scores (averaged both across students and across separate reading and math tests) were lower than a set cutoff.

There are three features of this program assignment that make this regression discontinuity design slightly more complicated than the simple example presented above. First, the cutoff

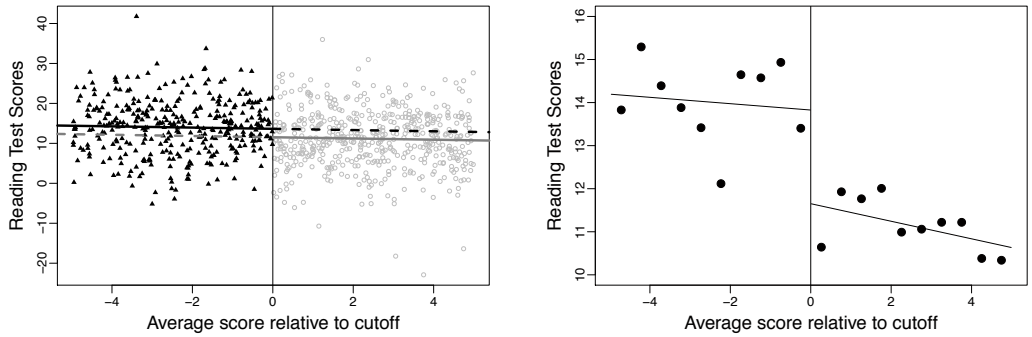

Figure 19.2 *Example of a regression discontinuity analysis: the intent-to-treat effect of exposure to P-900 on the change in reading test scores between 1988 and 1992. (a) School-level data and regression lines for the relationship between the forcing variable (average test score relative to the cutoff) and the outcome (change in reading test scores between 1988 and 1992) close to the cutoff. The variability in the test scores obscures the difference in average test scores at the discontinuity. (b) Binned averages of the data in the left graph, along with a regression fit that includes an interaction between the forcing variable and the indicator for eligibility (being above or below the cutoff).*

varied by region of the country. Therefore our assignment, or forcing, variable will always reflect the positive or negative distance of each school from the cutoff for their region.

Second, the precise cutoff was not recorded and must be deduced from the data. Here we use one of the cutoff rules introduced by the researchers who first analyzed these data using a regression discontinuity design.

The third challenge is that assignment rules do not seem to have been strictly followed: there are schools that did not receive the treatment while having average test scores that fall below the average test scores of the treated schools. In fact, while fewer than 2% of the schools above the derived cutoff received the program, only about 68% of those below the cutoff received it. This is an example of a "fuzzy" regression discontinuity. We will discuss further options for analyzing such designs later, but for now will just conceptualize the estimand as a sort of intention-to-treat effect. That is, we estimate the effect of being on one side of the cutoff or the other, regardless of whether the P-900 resources were actually received. We shall refer to this characteristic as "eligibility."

Figure 19.2a displays school-level data for the forcing variable (distance between average test score and the cutoff for that school's region) and the outcome variable (difference between average test score in 1992 and average test score in 1988). By definition the cutoff is at zero. The data are restricted to schools close to the cutoff. Schools that are eligible are plotted as triangles and schools that should not be eligible are displayed as open circles. A simple regression fit to the data is overlaid with solid lines appearing only where there is support in the data (that is, below the cutoff for those schools that were eligible and above the cutoff for those schools that were not eligible).

The wide variation in test scores makes it difficult to determine whether a discontinuity exists. Figure 19.2b provides a less noisy representation of this relationship that bins schools and plots the mean in each bin. The regression overlay in this plot includes an interaction between the *forcing variable* and the causal predictor of interest, eligibility for P-900.

Regression discontinuity analysis. If we wish to make comparisons between similar schools we should probably restrict our data analysis to a narrow range, for example, the schools with test scores within 5 points of the cutoff. We could then fit a model of the form,

$$y_i = \beta_0 + \tau * \text{eligible}_i + \beta_1 * \text{rule2}_i + \text{error}_i,$$

where eligible$_i$ is the indicator for being below the cutoff, rule2$_i$ represents the assignment rule, and the outcome y_i is the change in reading test scores between 1988 and 1992. It also makes sense when analyzing such gain scores to adjust for pre-test, that is, the reading test score in 1988.

Here is the result of the regression with a subsetting condition set up to restrict the analysis to those within 5 points of the appropriate cutoff:

R output
```
lm(formula = dcas92 ~ eligible + rule2, subset=ss5, data=p900)
            coef.est coef.se
(Intercept) 11.57     0.51
eligible     2.11     0.93
rule2       -0.15     0.17
---
n = 883, k = 3
residual sd = 7.01, R-Squared = 0.04
```

The fit of this model is displayed by the lines in the left hand plot in Figure 19.2. It would make sense to control for pre-test in this regression, but that is not necessary for illustrating the general principles.

The intent-to-treat effect of being eligible for the P-900 program is estimated by the coefficient on `eligible` is an increase in reading test scores of about 2 points on average. The standard error is somewhat high relative to the magnitude of this coefficient. The large uncertainty in the coefficient for `eligible` is no surprise, given that we have restricted our analysis to the subset of data for which `rule2` lies in a fairly narrow range.

One downside of this analysis is that the estimated effect applies only for those schools close to each threshold. The upside is that the assumptions (formalized below) are more reasonable.

Regression fit to all the data. Alternatively, we could fit the model to the whole dataset which includes schools with average test scores that are roughly 20 points below their cutoff as well as schools with average test scores up to 40 points above their cutoff:

R output
```
lm(formula = dcas92 ~ eligible + rule2, data=p900)
            coef.est coef.se
(Intercept) 11.91     0.26
eligible     2.13     0.44
rule2       -0.24     0.02
---
n = 2604, k = 3
residual sd = 6.79, R-Squared = 0.16
```

The coefficient on `eligible` is estimated much more precisely, which makes sense given that we have more leverage on `rule2`. However it is unclear that we should trust a regression fit to all these data. Causal inference for the full sample implicitly requires a belief that model extrapolation is appropriate because it represents a comparison between a model for y^1 in neighborhoods of the covariate space where there are no treated observations and a model for y^0 in neighborhoods where there are no control observations. This will require a leap of faith, however perhaps a reasonable first step is to make sure that the model at least fits well for the observed data.

Regression with interactions. There is typically no reason to believe that the association between the forcing variable and the outcome is the same across the exposed and unexposed groups. Therefore it is typical to fit a model interacting `rule2` with the indicator for eligibility:

R output
```
lm(formula = dcas92 ~ eligible*rule2, subset=ss5, data=p900)
               coef.est coef.se
(Intercept)    11.67     0.60
```

```
eligible        2.17     0.95
rule2          -0.19     0.21
eligible:rule2  0.11     0.34
---
n = 883, k = 4
residual sd = 7.01, R-Squared = 0.04
```

This model, once again fit to schools close to the cutoff, corresponds to the regression fit in the right hand plot in Figure 19.2. This fit suggests an average causal effect at the margin that is just slightly higher than the effect estimated by the no-interaction model.

Comparison of the analysis near the threshold to the analysis with interactions using all the data. In this example, the analysis fit to the entire dataset (not shown here) yields an estimate of about half the size of the three estimates above (that is, close to 1) albeit with a much lower standard error. as the regression discontinuity analysis that focused on the region of near overlap. In general, however, the model fit just to the area proximal to the threshold may be considered more trustworthy.

Assumptions

What assumptions do we need to make believe that regression discontinuity estimates identify causal effects? If we simply used a difference in means to estimate the causal effect of being below the threshold, z, in a narrow range defined by the assignment variable, x we would need ignorability to hold in that range, that is $y(1), y(0) \perp z$ for $x \in (C-a, C+a)$, where C represents the threshold for assignment to treatment and control. Researchers in other disciplines sometimes express this as an assumption that no confounders vary discontinuously across the threshold.

In practice though we always want to condition on the assignment variable x by incorporating it as a predictor in our estimation model. This allows us to weaken the ignorability assumption to $y(1), y(0) \perp z \,|\, x$ for $x \in (C-a, C+a)$. However we then require that we can appropriately model $\mathrm{E}(y(0) \,|\, x$ and $\mathrm{E}(y(1) \,|\, x$ in this range, even without any common support to support this estimation from the data.

Special cases of these assumptions exist for estimands that we have previously discussed such as the effect of the treatment on the treated and the effect of the treatment on the controls. For instance if $\Pr(z = 1 \,|\, x < C) = 1$, then the effect of the treatment on the controls simply corresponds to an estimation range of $x < C$. But if we restrict our analyses to observations close to the threshold then we would expect all of these estimands to be similar (since the point of such a restriction is to create homogeneity across units).

Fuzzy regression discontinuity and connections to instrumental variables

What happens when the discontinuity is not starkly defined, that is, when treated and control observations fall on the opposite side of the threshold from where they were assigned? This is sometimes called a "fuzzy" regression discontinuity, as opposed to the "sharp" discontinuity discussed thus far. Our example with the P-900 program in Chile is an example of such a fuzzy regression discontinuity design.

As another (slightly more simple) example, consider a situation where the decision whether to promote children to the next grade is supposed to be made solely based upon results from a standardized test (or set of standardized tests). Theoretically this should create a situation with no overlap in these test scores across those children forced to repeat their grade and those promoted to the next grade (the treatment and control groups). In reality, though, students may not comply with their treatment assignment. Some children may be promoted despite failure to exceed the threshold based on any of several reasons, including parental pressure on school administrators, a teacher who advocates for the child,

or an exemption from the policy due to classification as learning disabled. Other students may be retained even though they earned a score above the threshold, due perhaps to social or behavioral issues.

This situation creates partial overlap between the treatment and control groups in terms of the variable that should be the sole confounding covariate, the forcing variable, promotion test scores. Given our concern about the lack of overlap, this may seem to be a preferable scenario. Unfortunately, this overlap arises from deviations from the stated assignment mechanism. If the reasons for these deviations are well defined (and measurable), then ignorability can be maintained by adjusting for the appropriate child, parent, or school characteristics. Similarly, if the reasons for these deviations are independent of the potential outcomes of interest, there is no need for concern. More plausibly, however, such inferences could be compromised by failure to adjust for important omitted confounders.

Connection to instrumental variables. An alternative strategy is to approach the causal effect estimation from an instrumental variables perspective. In certain scenarios we can extend our conceptualization of a sharp regression discontinuity design from that of a simple randomized experiment (for units within a narrow margin of the threshold) to that of a randomized experiment with noncompliance. In this framework the indicator for whether an observation falls above or below the threshold, z ("eligibility" in or Chile example) acts as the instrument. The causal variable of interest, T, is the indicator for whether the program or treatment was actually received (P-900 in the Chilean example). For this strategy to be successful all of the standard instrumental variables assumptions must hold. It is far from true that instrumental variables estimation will always clean up a fuzzy regression discontinuity. Let's walk through these assumptions in turn.

Ignorability of the instrument is a bit more complicated in the regression discontinuity setting (as compared to a randomized experiment at least) because of the necessity of either choosing an appropriately narrow bandwidth or specifying an appropriate model to properly adjust for the assignment variable.

The interpretation of the *exclusion restriction* is similar here to a traditional instrumental variables setting. Never takers (those who will never receive the treatment regardless of their placement with regard to the cutoff) and always takers (those who will always receive the treatment regardless of their placement with regard to the cutoff) must have the same outcomes regardless of whether then end up above or below the cutoff (that is, regardless of their value for the instrument). More colloquially, we can express this assumption as a requirement that the instrument (falling on either side of the cutoff) can only affect the outcome through its effect on the treatment.

To satisfy *monotonicity*, we need to exclude the possibility of defiers. That is there cannot be observations that would receive the treatment if they fall on the side of the cutoff that makes them eligible but would not receive it otherwise.

Finally, *the instrument must be predictive of the treatment variable.* If the cutoff that determines "eligibility" is unrelated to the probability that observations receive the treatment then we cannot make progress.

As a reminder, we emphasize that, as with standard instrumental variables estimation, if the required assumptions hold the procedure will estimate the effect of the treatment only for the compliers—that is, the people (or, more generally, experimental units) who *would have* adhered to their assignment no matter which side of the threshold they fell on—and only those compliers in the chosen bandwidth. In our grade retention example these are the students who would have been retained in grade had their test scores fallen short of the cutoff but who would have been promoted if their test scores had exceeded the threshold. Never takers are those students who would have been promoted no matter their test score and always takers are those who would be held back regardless. As with the ignorability assumption, the exclusion restriction may become more or less plausible as we vary the

margin around the threshold that defines the subset of students that we choose to consider for conducting our analyses.

Returning to the P-900 example. So can we use an instrumental variables to estimate the effect of actually receiving the P-900 program? In addition to the ignorability assumption that requires the schools close to the threshold to, in effect, be randomly assigned to being above or below that threshold, other assumptions would be required. To satisfy the exclusion restriction we would need to believe that lying above or below the threshold has no effect on outcomes except through its effect on whether schools received the P-900 program. Another way of framing the assumption is that schools that were "never takers" (here, those schools who would not receive the program no matter which side of the threshold they fell on) and "always takers" (here, schools that would receive the program no matter which side of the threshold they fell on) would not have had different test scores if they fell above or below the threshold. For a narrow range of schools this may be plausible.

The monotonicity assumption would require that there are no schools that would not receive P-900 if they fell in the eligible range of test scores but would receive it if they fell in the non-eligible range. We know that there is a strong correlation between the instrument (falling above or below the derived eligibility cutoff) and the treatment (receiving P-900).

If we perform an instrumental variables analysis on our subset of schools close to the threshold we obtain an estimate of the effect of participating in the P-900 program of about 3 points with a standard error of about 0.6. This result holds for the schools that would participate in the P-900 program if they were eligible (below the cutoff) and not otherwise (the compliers). But we do not know enough about education policy in Chile to feel completely secure in the assumptions required for these results to be valid, particularly with regard to the exclusion restriction.

19.4 Identification strategies that make use of variation within or between groups

Comparisons within groups using varying-intercept ("fixed effects") models

Another set of strategies capitalizes on situations in which repeated observations within groups can provide a means for adjusting for unobserved characteristics of these groups. If comparisons are made across the observations within a group, implicitly such comparisons hold constant all characteristics intrinsic to the group or individual that do not vary across members of the group.

For example, suppose you want to examine the effect of low birth weight on children's mortality and other health outcomes. One difficulty in establishing a causal effect is that children with low birth weight are also typically disadvantaged in genetic endowments and socioeconomic characteristics of the family, some of which may not be easy or possible to measure. Rather than trying to directly adjust for all of these characteristics, however, one could implicitly adjust for them by comparing outcomes across children within a pair of twins. So we may be able to consider birth weight to be randomly assigned (ignorable) *within* twin pairs. In essence then each twin acts as a counterfactual for his or her sibling. Theoretically, if there is enough variation in birth weight, within sets of twins, we can estimate the effect of birth weight on subsequent outcomes.

A regression model that is sometimes used to approximate this conceptual comparison simply allows for each group (here twin pair) to have its own intercept. This model can be expressed as,

$$y_{ij} = \beta_0 + \tau * \text{weight}_{ij} + \alpha_i + \epsilon_{ij}.$$

To fit this model one could regress outcomes on birth weight (the "treatment" variable) as well as one indicator variable for each pair of twins (keeping one pair as a baseline category

to avoid collinearity). This approach is computationally inefficient, however, and therefore differencing strategies are typically used in practice by software packages that target these models. In the case of groups (indexed by i) of size 2, for example, we would replace each pair of observations with a single differenced observation (for example, $y_{j2} - y_{j1}$) for each variable:

$$y_{i2} - y_{i1} = \tau * (\text{weight}_{i2} - \text{weight}_{i1}) + \epsilon_{i2} - \epsilon_{i1}$$

In examples with more than two observations per group a more general strategy would be to subtract the group specific means of each variable from in individual level value, as in,

$$y_{ij} - \bar{y}_i = \tau(z_{ij} - \bar{z}_i) + \epsilon_{ij} - \bar{\epsilon}_i$$

where z now denotes the treatment in this more general model. The `plm` package in `R` will fit this model (by using the "within" option) and will produce standard errors that adjust for the fact that the group-level means are estimated.

This varying-intercepts model is sometimes called a "fixed-effects model," a term we avoid because it is used in various ambiguous ways in the statistics literature. We discuss varying-intercept and varying-slopes models in more detail in our book on multilevel models. Two reasons why hierarchical or multilevel modeling can be useful for varying intercepts and slopes are: (1) when the number of observations per group is small, the simple unregularized linear regression estimates can be noisy, and (2) with a multilevel model, we can include group-level predictors and varying coefficients for group at the same time, which cannot be done using classical regression. For now, though, we will continue with the simple classical estimate to get across the general point of varying-intercepts models, with the understanding that you can go back later and fit the model better using multilevel regression, at which point you can also learn what pitfalls to avoid using that approach as well.

The causal predictor needs to vary within group. This study design can create comparisons across different levels of the treatment variable only if it varies *within group*. In examples where the treatment is dichotomous, it is possible for a substantial portion of the sample to exhibit no variation at all in the causal variable within groups. For instance, suppose a varying-intercepts model is used to estimate the effect of maternal employment on child cognitive outcomes by including indicators for each family (set of siblings) in a regression of outcomes on an indicator for whether the mother worked in the first year of the child's life. In some families the mother may not have varied her employment status across children. Therefore, no inferences about the effect of maternal employment status can be made for these families. We can only estimate effects for the type of family where the mother varied her employment choice across the children (for example, working after her first child was born but staying home from work after the second).

Assumptions. What assumptions need to hold in order for varying-intercept regression to identify a causal effect? The structural assumption required can be conceived of as an extension of the standard ignorability assumption that additionally conditions on the group indicators,

$$y^0, y^1 \perp z \,|\, \alpha.$$

We can map this to a situation in which we have a separate randomized experiment within each group. Another framing is that ignorability holds once we condition on (potentially unmeasured) group-specific confounders that are common to all members of the group.

In addition to the ignorability assumption the varying-intercepts model makes all the assumptions of a standard regression model. In the simple version of the model without covariates this may not translate to strong assumptions about linearity and additivity (depending on the form of the treatment variable). We also assume that the observations are (conditionally) independent of each other. While this may not seem plausible given the grouped structure of the data recall that we can now conceive of each group member's

observation as simply the distance of this value from the group mean. Therefore, if the mechanism driving the similarity in responses within a group was the shared mean, then subtracting the mean of the data should will eliminate the problem. More complicated correlation structures do exist but those are often difficult to identify and fix.

Covariates. The researcher might also want to adjust for pre-treatment covariates, x, to improve the plausibility of the ignorability assumption (to account for the fact that the treatment may not be strictly randomly assigned even within each group). Formally this extends the ignorability assumption to

$$y^0, y^1 \perp z \,|\, \alpha, x.$$

Conditioning on such confounders may come at the cost of additional parametric assumptions though this is generally worth the tradeoff.

In our birth weight example it is difficult to find child-specific predictors that vary across children within a pair but can still be considered pre-treatment. However, the haphazard nature of the "assignment" to birthweight may obviate the need for such controls; that is, ignorability seems plausible even without conditioning on other variables.

In the maternal employment example, however, we are less likely to believe that the choice to enter the labor market is independent of child potential outcomes (or, said another way, is independent of child and family characteristics that might also influence child developmental outcomes). Therefore the analysis would likely be stronger if researchers could include child covariates measured before the mother returned to work such as birthweight or number of days in hospital as well as immutable characteristics such as sex. As we discuss in more detail next, however, researchers should be wary of conditioning on confounding covariates that may have been affected by the level of the treatment variable.

Within-person controls

Perhaps the most common use of varying intercepts when estimating causal effects occurs in the context of panel data (data in which the same individuals are surveyed over time, typically with a fair amount of overlap in the questions asked at each time point). For instance, consider examining the effect of marriage on men's earnings by analyzing data that follows men over time and tracks marital status, earnings, and predictors (potential confounding covariates) such as race, educational attainment, number of children, and occupation. Researchers have justified the use of varying-intercepts regressions in this setting by arguing that it allows for within-person comparisons that implicitly adjust for unobserved characteristics of each individual that do not vary over time. However when these models also condition on confounding covariates that could be affected by the treatment, they end up, in effect, adjusting for post-treatment variables, which, as we know from Section 17.6, can lead to bias.

In the above example it is hard to rule out the possibility that educational attainment, number of children, and occupation were affected by whether or not the study participant had gotten married at a previous (or current) time point. To be on the safe side a prudent choice would be to omit any covariate from the model that has the potential to be affected by the treatment. Unfortunately this may render ignorability completely implausible. The references at the end of the chapter point to more advanced approaches that (under a set of assumptions related to ignorability) are able to adjust for such variables without incurring this bias.

Comparisons within and between groups: difference-in-differences estimation

Almost all causal strategies make use of comparisons across groups: one or more groups that were exposed to a treatment, and one or more groups that were not. *Difference-in-difference*

strategies additionally make use of another source of variation in outcomes, typically time, to help adjust for potential (observed and unobserved) differences across these groups. For example, consider estimating the effect of a newly introduced school busing program on housing prices in a school district where some neighborhoods were affected by the program and others were not. A simple comparison of housing prices for all houses across affected and unaffected areas several years after the busing program went into effect might seem tempting, however it would not be appropriate because these neighborhoods might be different in other ways that are related to housing prices. For instance the busing program might have been implemented in neighborhoods that were more disadvantaged than the other neighborhoods in the school district and thus already had lower housing prices. A simple before-after comparison of housing prices would also typically be inappropriate because other changes that occurred during this time period (for example, a recession or economic boom) might also be influencing housing prices.

A difference-in-differences approach would instead capitalize on both sources of variation: the difference in housing prices across time and the difference in prices across exposed and unexposed neighborhoods. Perhaps the most intuitive way to think about this model is that it makes comparisons between exposed and unexposed groups with respect to the differences in means over time (housing price trajectories). For example, suppose the housing prices in the unexposed areas increased on average by \$20 000 between the pre-intervention and post-intervention time periods but that they increased by only \$10 000 in the exposed areas over this same period of time. If we can assume that the growth in housing prices in the unexposed areas reflects what would have happened in the exposed areas had the busing program never occurred then we could say that the estimated effect of the busing for the houses in the exposed areas was a decrease in house price of \$10 000.

Regression framework

The calculation we just performed simply involved four means. We could estimate the same effect in a regression framework by including each combination of observation and time point as a separate row in our dataset and fitting the model,

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 P_i + \tau z_i P_i + \epsilon_i, \tag{19.6}$$

where z denotes the treatment and P denotes the time period; $P = 0$ references the pre-exposure time period and $P = 1$ references the post-exposure time period. In this model the difference-in-difference estimand is the coefficient on the interaction term. To prove this to yourself algebraically see Exercise 19.**. Intuitively this makes sense because the interaction term represents the difference in the slope coefficients on time across the treatment groups where the slope coefficients on time each represent a difference in means (across time points). Equivalently, we could say that the interaction term represents the difference in slope coefficients on treatment across the time periods.

In our housing example we saw prices on the same sample of houses at two different times. If we include each house twice in our analysis we know that our regression assumption of independent observations will be violated. A simple fix is to instead fit a change-score model:

$$d_i = \beta_0 + \tau P_i + \epsilon_i, \tag{19.7}$$

where the estimand, τ, has the same interpretation as before. This simplification will not work if our observational units are not the same across the two time points.

Why would we bother to fit a regression if this coefficient simply represents the difference between two mean differences? Why not do a simple calculation with four means? First, the regression approach allows us to easily estimate the standard error on our treatment effect estimate. Furthermore, this framework will extend easily to allow complications such as including covariates in the model.

Assumptions

We discussed the required structural assumption above colloquially as a requirement that the change in test scores for the control group needs to represent what would have happened to the treatment group had they not been exposed to the treatment. We can formalize the assumption by defining a new quantity that we refer to as a potential change (rather than a potential outcome). In particular, define $d^0 = y^0 - y_{P=0}$, to reflect the potential change in outcomes (appraised house price) between the pre-exposure and post-exposure time periods if the unit were never exposed to the treatment. Here $y_{P=0}$ denotes the baseline (pre-exposure time period) measure of the outcome variable, which in other contexts we would simply think of as an important confounding covariate.

Our use of the term "outcome" and the notation y to denote both post- and pre-exposure versions of this variable is admittedly a bit confusing. Until this point we have used y to refer to post-treatment outcomes, but here we use y to refer to the variable we think that the treatment might effect, even during time periods before the treatment can have affected it. The benefit of this choice is that it may help us to better understand the regression models used in this context in which the response will sometimes reflect a true post-exposure outcome and will sometimes reflect a pre-exposure version of the same variable. Also, in the notation for potential changes, this choice makes it clear that we aren't just subtracting the value of some important confounder—we are subtracting the exact same measure at a different time point.

Mapping this to our earlier informal discussion of the assumptions and estimand, we can say that if $d^0 \perp T$ we can identify the effect for the treated, $\mathrm{E}(d^1 - d^0 \,|\, T = 1)$.

To define the effect of the treatment on the controls we need to first define the potential change under treatment assignment, $d^1 = y^1 - y_{P=0}$. Then to identify effect of the treatment on the controls $\mathrm{E}(d^1 - d^0 \,|\, T = 0)$ we would need to assume $d^0 \perp T$. To estimate the average treatment effect $\mathrm{E}(d^1 - d^0)$ we would need to assume $d^0, d^1 \perp T$.

Adjusting for potential confounding covariates. Without introducing confounding covariates however we have no way of distinguishing between average treatment effects for the control group and average treatment effects for the treated. Said another way, our current assumptions define them to be the same (just as in a completely randomized experiment these effects are the same). We can loosen this restriction and create more plausible ignorability assumptions by conditioning on other confounding covariates, x. this also allows us to a create more plausible ignorability assumption $d^0, d^1 \perp T \,|\, x$.

If we think that there is heterogeneity in our treatment effect and we wanted to explicitly target the effect for the treated or the effect for the controls we could combine the difference-in-difference framework with a method (such as one of the ones discussed in the previous chapter) that allows for the targeting of such estimands. For instance, we could run a change score model on a reduced sample consisting of the treatment units and the control units that have been matched to them.

Comparison with standard ignorability assumptions. The assumption needed with this strategy appears to be weaker than the unconditional ignorability assumption because rather than assuming that potential outcomes are the same across treatment groups, one only has to assume that the potential *changes* in outcomes over time are the same across groups (for example, exposed and unexposed neighborhoods). Therefore we need only believe that the difference in housing prices over time would be the same across the two types of neighborhoods, not that the average post-program potential housing prices if exposed or unexposed would be the same.

In studies where we observe the same units at both time periods, however, the pre-exposure outcome can be thought of simply as an important confounding covariate. If we had access to such information we would certainly want to adjust for that covariate (for

instance through a regression model or propensity score approach). In that case we would be relying on a related assumption to identify the causal effect: $y^0, y^1 \perp T \,|\, y_{P=0}$. This says that we would expect that among those with the same housing price during the pre-exposure time period, there is no difference (distributionally) in potential outcomes across treated and untreated groups. The differences between this assumption and the difference-in-differences assumption are subtle indeed.

Different observations at each time point. Thus far we have been discussing a special case within the difference-in-differences framework in which the same set of units are observed at both time points. As we have just noted, in this setting the advantages of the difference-in-differences strategy are less apparent because an alternative model would be to include an indicator for treatment exposure but then simply regress on the pre-treatment version of the outcome variable.

One strength of the difference in difference framework is that it accommodates situation in which the observations are *not* the same across time periods simply by using the initial regression framework introduced in (19.6). This usage requires the additional assumption however that the units sampled across the two time points are both random samples from the same distribution.

When might this assumption be violated? Consider the earlier example with busing and housing prices. Suppose that instead of using appraisal values to obtain housing prices of every house in each neighborhood, the researchers instead used sale prices from houses that had actually sold in each time period. Then if the busing also had an impact on the ability of owners to sell their house, the sample of houses that sold in the post-exposure time period would not be representative of the sample of houses that sold in the pre-exposure time period and this assumption would be violated. The ability to condition on covariates could weaken this assumption though note the cautions below with regard to such extensions of the model.

Do not condition on post-treatment outcomes. Once again, to make the (new) ignorability assumption (as well as the additional assumption required when we have different observational units across time points) more plausible it may be desirable to condition on potential confounders. However for standard difference-in-differences models in which one of the differences is across time points this model will implicitly condition on post-treatment variables (similar to the varying-intercepts specification with panel data). If these predictors can be reasonably assumed to be unchanged by the treatment, then this may be reasonable. However, as discussed in Section 17.7, it is otherwise inappropriate to control for post-treatment variables.

When we are working with panel data, a better strategy presents itself. In this case, we can fit the model in (19.7) with the addition of pre-treatment covariates. This amounts to running a regression of the change in outcomes on the treatment variable and covariate values from the pre-exposure time period.

Regression discontinuity versus difference in differences. There may be a temptation to use a difference-in-difference approach in situations where the data arise from a prospective or retrospective regression discontinuity design. After all, don't such designs often lead to situations in which there is an exposed and an unexposed group and measures of a response of interest both before and after the treatment group is exposed? Why wouldn't this be ideal?

As it turns out this design is particularly *poorly* suited to a difference-in-differences approach because, by design, a simple application of difference-in-differences is likely to suffer from bias due to regression to the mean effects. For example, consider our hypothetical regression discontinuity example with grade retention imposed based on student test scores falling below a given threshold. The regression discontinuity design is capitalizing on the fact that, for students with scores close to the threshold, falling above or below the cutoff is for all intents and purposes randomly assigned. For instance you might imagine that students

with test scores in that range all have scores that come from a probability distribution centered at the threshold value.

If we take this seriously then we realize that if nothing changed for those students over time, that is, *if there were no treatment effect*, then the next time we administered a similar test then there would be a high probability that the students who scored in the lower part of this distribution (so below the cutoff) the first time would be likely to score higher the next time and that the students who score relatively high (so above the cutoff) the first time would be likely to score lower the next time. Therefore if there were truly no treatment effect then we would still (repeatedly across samples) estimate a *negative* treatment effect due to this regression to the mean effect.

The regression discontinuity analysis avoids this problem by conditioning on the assignment variable. This sets up direct comparisons between units with the same value of the assignment variable (confounder) and highlights the role of specification of the models for $\mathrm{E}(y(1) \mid x)$ and $\mathrm{E}(y(0) \mid x)$.

19.5 Forward and reverse causal inference

Statistical methods for causal and policy analysis are more focused on "effects of causes" than on "causes of effects." That is, in the standard approach it is natural to study the effect of a treatment, but it is not in general possible to define the causes of any particular outcome. This has led some researchers to dismiss the search for causes as "cocktail party chatter" that is outside the realm of science. We argue here that the search for causes can be understood within traditional statistical frameworks as a part of model checking and hypothesis generation. We argue that it can make sense to ask *questions* about the causes of effects, but the answers to these questions will be in terms of effects of causes.

Example: cancer clusters. A map shows an unexpected number of cancers in some small location, and this raises the questions: What is going on? Why are there so many cancers in this place? These questions might be resolved via the identification of some carcinogenic agent (for example, a local environmental exposure) or, less interestingly, some latent variable (some systematic difference between people living in and not living in this area), or perhaps some statistical explanation (for example, a multiple-comparisons argument demonstrating that an apparent anomalous count in a local region is no anomaly at all, after accounting for all the possible clusters that could have been found). The question, Why so many cancers in that location?, until it is explained in some way, refers to a data pattern that does not fit one's pre-existing models. Researchers are motivated to look for an explanation because this might lead to a deeper understanding and ultimately, through the identification of some carcinogenic agent, to recommendations for policy that may lead to a reduction in cancer rates.

The cancer-cluster problem is typical of a lot of scientific reasoning. Some anomaly is observed and it needs to be explained. To borrow concepts from the philosophy of science, the resolution of the anomaly may be an entirely new paradigm or a revision of an existing research programme. In the cancer example, potential causes might be addressed via a hierarchical regression model as discussed in our next book

The purpose of the present discussion is to place "Why" questions within a statistical framework of causal inference. We argue that a question such as "Why are there so many cancers in this place?" can be viewed not directly as a question of causal inference, but rather indirectly as an identification of a problem with an existing statistical model, motivating the development of more sophisticated statistical models that can directly address causation in terms of counterfactuals and potential outcomes.

Forward and reverse causal questions

We distinguish two broad classes of causal queries:

1. *Forward causal questions*, or the estimation of "effects of causes." What might happen if we do z? What is the effect of some manipulation, for example the effect of smoking on health, the effect of schooling on earnings, the effect of campaigns on election outcomes, and so forth?
2. *Reverse causal inference*, or the search for "causes of effects." What causes y? Why do more attractive people earn more money? Why does *per capita* income vary some much by country? Why do many poor people vote for Republicans and rich people vote for Democrats? Why did the economy collapse?

When statistical and econometric methodologists write about causal inference, they generally focus on forward causal questions. We are taught to answer questions of the type "What if?", rather than "Why?" As we have discussed in the preceding chapters, causal questions are typically framed in terms of manipulations: if x were changed by one unit, how much would y be expected to change? But reverse causal questions are important too. They are a natural way to think, and in many ways, it is the reverse causal questions that motivate the research, including experiments and observational studies, that we use to answer the forward questions.

The question discussed in the current section is: How can we incorporate reverse causal questions into a statistical framework that is centered around forward causal inference? Even methods such as path analysis or structural modeling, which in some settings can be used to determine the direction of causality from data, are still ultimately answering forward casual questions of the sort, What happens to y when we change x? Our resolution is as follows: Forward causal inference is about estimation; reverse causal questions are about model checking and hypothesis generation. To put it another way, we ask reverse causal *questions* all the time, but we do not perform reverse causal *inference*.

We do not try to answer "Why" questions; rather, "Why" questions motivate "What if" questions that can be studied using standard statistical tools such as experiments, observational studies, and structural equation models.

Example: political campaigns. Consider the forward causal question, What is the effect of money on elections? To be answered, the question needs to be made more precise by defining the treatments and outcome. For example: supposing a challenger in a given congressional election race is given a $100 donation, how much will this change his or her expected vote share? It is not so easy to get an accurate answer to this question, but the causal quantity is well defined (and can be specified even more precisely, for example with further details about the contribution, as necessary).

Now a reverse causal question: Why do incumbents running for reelection to Congress get so much more funding than challengers? Many possible answers have been suggested, including the idea that people like to support a winner, that incumbents have higher name recognition, that certain people give money in exchange for political favors, and that incumbents are generally of higher political "quality" than challengers and get more support of all types. Various studies could be performed to evaluate these different hypotheses, all of which could be true to different extent (and in some interacting ways).

Now the notation. It is generally accepted (and we agree) that forward causal inferences can be handled in a potential-outcome or graphical-modeling framework involving a treatment variable T, an outcome y, and pre-treatment variables, x, so that the causal effect is defined (in the simple case of binary treatment) as $y(T=1, x) - y(T=0, x)$. The actual estimation will likely involve some modeling (for example, some curve of the effect of money on votes that is linear at the low end, so that a $20 contribution has twice the expected effect as

$10), but there is little problem in defining the treatment effect. The challenge arises from estimating these effects from observational data.

Reverse causal reasoning is different; it involves asking questions and searching for new variables that might not yet even be in our model. We can frame reverse causal questions as model checking. It goes like this: what we see is some pattern in the world that needs an explanation. What does it mean to "need an explanation"? It means that existing explanations—the existing model of the phenomenon—does not do the job. This model might be implicit. For example, if we ask, Why do incumbents get more contributions than challengers, we are comparing to an implicit model that all candidates get the same. If we gather some numbers on dollar funding, compare incumbents to challengers, and find the difference is large and precisely estimated, then we are comparing to the implicit model that there is variation but not related to incumbency status. If we get some measure for candidate quality (for example, previous elective office and other qualifications) and still see a big difference between the funds given to incumbents and challengers, then it seems we need more explanation. And so forth.

Here is an example told to us by a statistician working in a business environment:

> A lot of real world problems are of the reverse causality type, and it's an embarrassment for us to ignore them. . . . Most business problems are reverse causal. Take for example P&G who spend a huge amount of money on marketing and advertising activities. The money is spread out over many vehicles, such as television, radio, newspaper, supermarket coupons, events, emails, display ads, search engine, etc. The key performance metric is sales amount.
>
> If sales amount suddenly drops, then the executives will want to know what caused the drop. This is the classic reverse causality question. Many possible hypotheses could be generated . . . TV ad was not liked, coupons weren't distributed on time, emails suffered a deliverability problem, etc. By a process of elimination, one can drill down to a small set of plausible causes. This is all complex work that gives approximate answers.
>
> The same question can be posed as a forward causal problem. We now start with a list of treatments. We will divide the country into regions, and vary the so-called marketing mix, that is, the distribution of spend across the many vehicles. This generates variations in the spend patterns by vehicle, which allows us to estimate the effect of each of the constituent vehicles on the overall sales performance.

This is the pattern: a quantitative analyst notices an anomaly, a pattern that cannot be explained by current models. The reverse-causal question is: Why did this happen? And this leads to improved modeling. Often these models are implicit. For example, consider the much-asked question, Why did violent crime in American cities decline so much in recent decades? This is a question asked in reference to summaries of crime statistics, not with reference to any particular model. But the Why question points to factors that should be considered. Economists have had similar discussions along the lines of, Why was there a financial crisis in 2008, and why did it happen then rather than sooner or later?

Again, the point of this section is to discuss how such Why questions fit into statistics and policy analysis, not as inferential questions to be answered with estimates or confidence intervals, but as the identification of statistical anomalies that motivate improved models.

Relation to formal models of causal inference

Now let us put some mathematical structure on the problem using the potential outcome framework. Let y_i denote the outcome of interest for unit i, say an indicator whether individual i developed cancer. We may also observe characteristics of units, individuals in this case, that are known to, or expected to, be related to the outcome, in this case, related to the probability of developing cancer. Denote those by w_i. Finally, there is a characteristic of the units, denoted by u_i, that the researcher feels should not affect of the outcome, and

so one would expect that in populations homogeneous in w_i, there should be no association between the outcome and u_i:

$$y_i \perp u_i \,|\, w_i.$$

This attribute u_i may be the location of individuals. It may be a binary characteristic, say female versus male, or an indicator for a subpopulation. The key is that the researcher interprets this variable as an attribute that should not be correlated with the outcome in homogeneous subpopulations.

However, suppose the data reject this hypothesis, and show a substantial association between u_i and the outcome, even after adjusting for differences in the other attributes. In a graphical representation of the model there would be evidence for an arrow between u_i and y_i, in addition to the arrow from w_i and y_i, and possibly a general association between u_i and w_i.

Such a finding is consistent with two alternative models. First, it may be that there is a cause, its value for unit i denoted by x_i, such that the potential outcome given the cause is not associated with u_i, possibly after conditioning on w_i. Let $y_i(x)$ denote the potential outcomes in the potential outcome framework, and $y_i(x_i)$ the realized outcome. Formally we would hypothesize that, for all x,

$$y_i(x) \perp u_i \,|\, w_i.$$

Thus, the observed association between y_i and u_i is the result of a causal effect of x_i on u_i, and an association between the cause x_i and the attribute u_i. In the cancer example there may be a carcinogenic agent that is more common in the area with high cancer rates.

The second possible explanation that is still consistent with no causal link between u_i and y_i is that the researcher omitted an important attribute, say v_i. Given this attribute and the original attributes w_i, the association between u_i and y_i would disappear:

$$y_i \perp u_i \,|\, v_i, w_i.$$

For example, it may be that individuals in this area have a different genetic background that makes them more susceptible to the particular cancer.

Both these alternative models could in principle provide a satisfactory explanation for the anomaly of the strong association between u_i and the outcome. Whether these models do so in practice, and which of these models do, and in fact whether the original association is even viewed as an anomaly, depends on the context and the researcher. Consider the finding that taller individuals command higher wages in the labor market. Standard economic models suggest that wages should be related to productivity, rather than height, and such an association might therefore be viewed as an anomaly. It may be that childhood nutrition affects both adult height and components of productivity. One could view that as an explanation of the first type, with childhood nutrition as the cause that could be manipulated to affect the outcome. One could also view health as a characteristic of adult individuals that is a natural predictor of productivity.

As stressed before, there are generally not unique answers to these questions. In the case of the association between height and earnings, one researcher might find that health is the omitted attribute, and another researcher might find that childhood nutrition is the omitted cause. Both could be right, and which answer is more useful will depend on the context. The point is that the finding that height itself is correlated with earnings is the starting point for an analysis that explores causes of earnings, that is, alternative models for earnings determination, that would reproduce the correlation between height and earnings without the implication that intervening in height would change earnings.

What does this mean for statistical practice?

A key theme in this discussion is the distinction between causal *statements* and causal *questions*. A reverse causal question does not in general have a well-defined answer, even in a setting where all possible data are made available. But this does not mean that such questions are valueless or that they fall outside the realm of statistics. A reverse question places a focus on an anomaly—an aspect of the data unlikely to be reproducible by the current (possibly implicit) model—and points toward possible directions of model improvement.

It has been (correctly) said that one of the main virtues of the potential outcome framework is that it motivates us to be explicit about interventions and outcomes in forward causal inference. Similarly, one of the main virtues of reverse causal thinking is that it motivates us to be explicit about what we consider to be problems with our model.

In terms of graphical models, the anomalies also suggest that the current model is inadequate. In combination with the data, the model suggests the presence of arrows that do not agree with our prior understanding. The implication is that one needs to build a more general model involving additional variables that would eliminate the arrow between the attribute and the outcome.

By formalizing reverse casual reasoning within the process of data analysis, we hope to make a step toward connecting our statistical reasoning to the ways that we naturally think and talk about causality, to be able to talk about reverse causal questions in a way that is complementary to, rather than outside of, the mainstream formalisms of statistics and econometrics.

Just as we view graphical exploratory data analysis as a form of checking models (which maybe implicit), we similarly hold that reverse causal questions arise from anomalies—aspects of our data that are not readily explained—and that the search for causal explanations is, in statistical terms, an attempt to improve our models so as to reproduce the patterns we see in the world.

19.6 Bibliographic note

Instrumental variables formulations date back to work in the economics literature by Tinbergen (1930) and Haavelmo (1943). Angrist and Krueger (2001) present an upbeat applied review of instrumental variables. Imbens (2004) provides a review of statistical methods for causal inference that is a little less enthusiastic about instrumental variables. Woolridge (2001, chapter 5) provides a crisp overview of instrumental variables from a classical econometric perspective; Lancaster (2004, chapter 8) uses a Bayesian framework. The "always taker," "complier," and "never taker" categorizations here are adapted from Angrist, Imbens, and Rubin (1996), who reframe the classic econometric presentation of instrumental variables in statistical language and clarify the assumptions and the implications when the assumptions are not satisfied. For a discussion of all of the methods discussed in this chapter from an econometric standpoint, see Angrist and Krueger (1999).

The Vietnam draft lottery example comes from several papers including Angrist (1990). The weather and fish price example comes from Angrist, Graddy, and Imbens (2000). The sex of child example comes from Angrist and Evans (1998).

Glickman and Normand (2000) derive an instrumental variables estimate using a latent-data model; see also Carroll et al. (2004).

Imbens and Rubin (1997) discuss a Bayesian approach to instrumental variables in the context of a randomized experiment with noncompliance. Hirano et al. (2000) extend this framework to include covariates. Barnard et al. (2003) describe further extensions that additionally accommodate missing outcome and covariate data. For discussions of prior distributions for instrumental variable models, see Dreze (1976), Maddala (1976), Kleibergen and Zivot (2003), and Hoogerheide, Kleibergen and van Dijk (2006).

For a discussion of use of instrumental variables to estimate bounds for the average treatment effect (as opposed to the local average treatment effect), see Robins (1989), Manski (1990), and Balke and Pearl (1997). Robins (1994) discusses estimation issues.

Sarsons (2015) discusses problems with unexamined assumptions in instrumental variables analysis, in the context of a previously-fitted model of income and political conflict that used rainfall as an instrument even though the exclusion restriction was violated in this case.

Our discussion of weak instruments may look different from the bulk of the econometrics literature on weak instruments. That literature focuses on the fact that in an overidentified instrumental variables model (that is, one with more instruments than causal variables), if the instruments are not sufficiently predictive (as measured by a test of the hypothesis that the coefficients in the first stage model are all equal to zero) then the resulting effect estimates may be biased. Given the difficulty of meeting the assumptions required of even one instrument, our advice is to stick to situations in which there is one instrument and one treatment variable (or at least the same number of instruments as treatment variables).

For more on the Sesame Street encouragement study, see Bogatz and Ball (1971) and Murphy (1991).

Regression discontinuity analysis is described by Thistlethwaite and Campbell (1960). More recent work in econometrics includes Hahn, Todd, and van der Klaauw (2001) and Linden (2006). Gelman and Zelizer (2015) and Gelman and Imbens (2017) explore some potential pitfalls when regression discontinuity analysis is applied in an automatic manner where it is not appropriate. The example regarding children's promotion in school was drawn from work by Jacob and Lefgren (2004).

Ashenfelter, Zimmerman, and Levine (2003) discuss "fixed effects" and difference-in-differences methods for causal inference. The twins and birth weight example was based on a paper by Almond, Chay, and Lee (2005). Another interesting twins example examining the returns from education on earnings can be found in Ashenfelter and Krueger (1994). Aaronson (1998) and Chay and Greenstone (2003) provide further examples of the application of these approaches. The busing and housing prices example is from Bogart and Cromwell (2000). Card and Krueger (1994) discuss a classic example of a difference-in-differences model that uses panel data.

Wainer, Palmer, and Bradlow (1998) provide a friendly introduction to selection bias. Heckman (1979) and Diggle and Kenward (1994) are influential works on selection models in econometrics and biostatistics, respectively. Rosenbaum and Rubin (1983b), Rosenbaum (2002a), and Greenland (2005) consider sensitivity of inferences to ignorability assumptions.

Sobel (1990, 1998) discusses the assumptions needed for structural equation modeling more generally.

Section 19.5 is taken from Gelman and Imbens (2013), and the quote on page 413 is from Kaiser Fung.

19.7 Exercises

1. *Regression discontinuity analysis:* Suppose you are trying to evaluate the effect of a new procedure for coronary bypass surgery that is supposed to help with the postoperative healing process. The new procedure is risky, however, and is rarely performed in patients who are over 80 years old. Data from this (hypothetical) example are displayed in Figure 19.3.
   (a) Does this seem like an appropriate setting in which to implement a regression discontinuity analysis?
   (b) The folder `Bypass` contains data for this example: `stay` is the length of hospital stay after surgery, `age` is the age of the patient, and `new` is the indicator for whether the new surgical procedure was used. Preoperative disease severity (`severity`) was unobserved

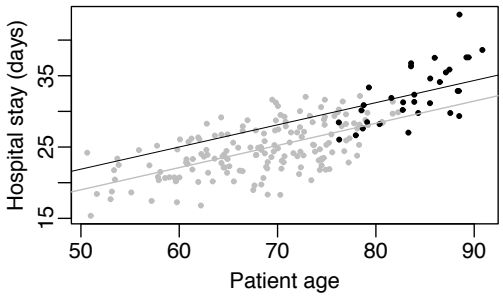

Figure 19.3 *Hypothetical data of length of hospital stay and age of patients, with separate points and regression lines plotted for each treatment condition: the new procedure in gray and the old procedure in black.*

by the researchers, but we have access to it for illustrative purposes. Can you find any evidence using these data that the regression discontinuity design is inappropriate?

(c) Estimate the treatment effect using a regression discontinuity estimate (ignoring) severity. Estimate the treatment effect in any way you like, taking advantage of the information in severity. Explain the discrepancy between these estimates.

2. *Instrumental variables:* Come up with a hypothetical example in which it would be appropriate to estimate treatment effects using an instrumental variables strategy. For simplicity, stick to an example with a binary instrument and binary treatment variable.

(a) Simulate data for this imaginary example if all the assumptions are met. Estimate the local average treatment effect for the data by dividing the intent-to-treat effect by the percentage of compliers. Show that two-stage least squares yields the same point estimate.

(b) Now simulate data in which the exclusion restriction is not met (so, for instance, those whose treatment level is left unaffected by the instrument have a treatment effect of half the magnitude of the compliers) but the instrument is strong (say, 80% of the population are compliers), and see how far off your estimate is.

(c) Finally, simulate data in which the exclusion restriction is violated in the same way, but where the instrument is weak (only 20% of the population are compliers), and see how far off your estimate is.

3. *Intermediate outcomes:* In Exercise 17.9, you estimated the effect of incumbency on votes for Congress. Now consider an additional variable: money raised by the congressional candidates. Assume this variable has been coded in some reasonable way to be positive in districts where the Democrat has raised more money and negative in districts where the Republican has raised more.

(a) Explain why it is inappropriate to include money as an additional input variable to "improve" the estimate of incumbency advantage in the regression in Exercise 17.9.

(b) Suppose you are interested in estimating the effect of money on the election outcome. Set this up as a causal inference problem (that is, define the treatments and potential outcomes).

(c) Explain why it is inappropriate to simply estimate the effect of money using instrumental variables, with incumbency as the instrument. Which of the instrumental variables assumptions would be reasonable in this example and which would be implausible?

(d) How could you estimate the effect of money on congressional election outcomes?

See Campbell (2002) and Gerber (2004) for more on this topic.