CHAPTER 9

Causal inference using regression on the treatment variable

9.1 Causal inference and predictive comparisons

So far, we have been interpreting regressions *predictively*: given the values of several inputs, the fitted model allows us to predict y, considering the n data points as a simple random sample from a hypothetical infinite "superpopulation" or probability distribution. Then we can make comparisons across different combinations of values for these inputs.

This chapter and the next consider *causal inference*, which concerns what *would happen* to an outcome y as a result of a hypothesized "treatment" or intervention. In a regression framework, the treatment can be written as a variable T:[1]

$$T_i = \begin{cases} 1 & \text{if unit } i \text{ receives the "treatment"} \\ 0 & \text{if unit } i \text{ receives the "control,"} \end{cases}$$

or, for a continuous treatment,

$$T_i = \text{level of the "treatment" assigned to unit } i.$$

In the usual regression context, predictive inference relates to comparisons *between* units, whereas causal inference addresses comparisons of different treatments if applied to the *same* units. More generally, causal inference can be viewed as a special case of prediction in which the goal is to predict what *would have happened* under different treatment options. We shall discuss this theoretical framework more thoroughly in Section 9.2. Causal interpretations of regression coefficients can only be justified by relying on much stricter assumptions than are needed for predictive inference.

To motivate the detailed study of regression models for causal effects, we present two simple examples in which predictive comparisons do not yield appropriate causal inferences.

Hypothetical example of zero causal effect but positive predictive comparison

Consider a hypothetical medical experiment in which 100 patients receive the treatment and 100 receive the control condition. In this scenario, the causal effect represents a comparison between what would have happened to a given patient had he or she received the treatment compared to what would have happened under control. We first suppose that the treatment would have no effect on the health status of any given patient, compared with what would have happened under the control. That is, the *causal effect* of the treatment is zero.

However, let us further suppose that treated and control groups systematically differ, with healthier patients receiving the treatment and sicker patients receiving

[1] We use a capital letter for the vector T (violating our usual rule of reserving capitals for matrices) in order to emphasize the treatment as a key variable in causal analyses, and also to avoid potential confusion with t, which we sometimes use for "time."

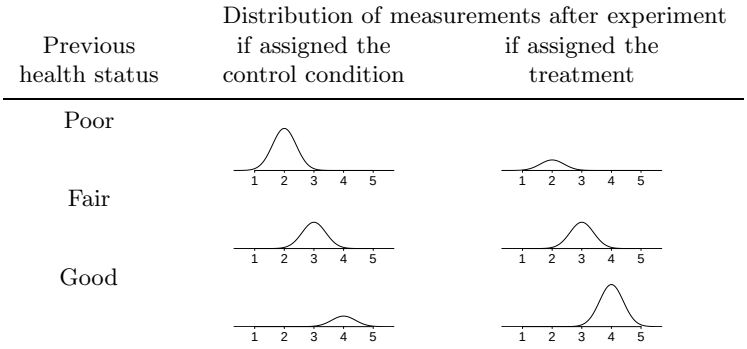

Figure 9.1 *Hypothetical scenario of* zero causal effect *of treatment: for any value of previous health status, the distributions of potential outcomes are* identical *under control and treatment. However, the* predictive comparison *between treatment and control could be positive, if healthier patients receive the treatment and sicker patients receive the control condition.*

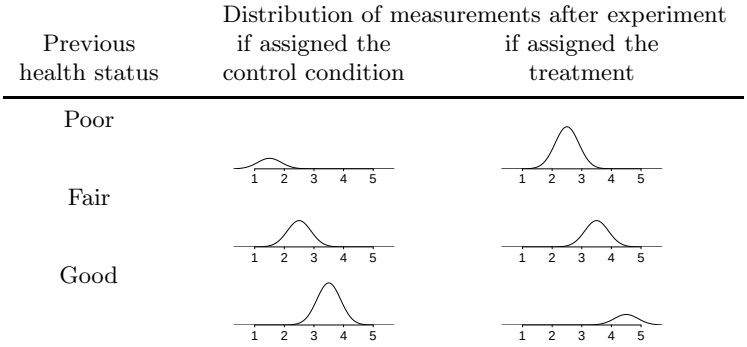

Figure 9.2 *Hypothetical scenario of* positive causal effect *of treatment: for any value of previous health status, the distributions of potential outcomes are centered at* higher *values for treatment than for control. However, the* predictive comparison *between treatment and control could be zero, if sicker patients receive the treatment and healthier patients receive the control condition. Compare to Figure 9.1.*

the control. This scenario is illustrated in Figure 9.1, where the distribution of outcome health status measurements is centered at the same place for the treatment and control conditions within each previous health status category (reflecting the lack of causal effect) but the heights of each distribution reflect the differential proportions of the sample that fell in each condition. This scenario leads to a positive *predictive comparison* between the treatment and control groups, even though the causal effect is zero. This sort of discrepancy between the predictive comparison and the causal effect is sometimes called self-selection bias, or simply selection bias, because participants are selecting themselves into different treatments.

Hypothetical example of positive causal effect but zero positive predictive comparison

Conversely, it is possible for a truly nonzero treatment effect to not show up in the predictive comparison. Figure 9.2 illustrates. In this scenario, the treatment has a positive effect for all patients, whatever their previous health status, as displayed

by outcome distributions that for the treatment group are centered one point to the right of the corresponding (same previous health status) distributions in the control group. So, for any given unit, we would expect the outcome to be better under treatment than control. However, suppose that this time, sicker patients are given the treatment and healthier patients are assigned to the control condition, as illustrated by the different heights of these distributions. It is then possible to see equal average outcomes of patients in the two groups, with sick patients who received the treatment canceling out healthy patients who received the control.

Previous health status plays an important role in both these scenarios because it is related both to treatment assignment and future health status. If a causal estimate is desired, simple comparisons of average outcomes across groups that ignore this variable will be misleading because the effect of the treatment will be "confounded" with the effect of previous health status. For this reason, such predictors are sometimes called *confounding covariates*.

Adding regression predictors; "omitted" or "lurking" variables

The preceding theoretical examples illustrate how a simple predictive comparison is not necessarily an appropriate estimate of a causal effect. In these simple examples, however, there is a simple solution, which is to compare treated and control units conditional on previous health status. Intuitively, the simplest way to do this is to compare the averages of the current health status measurements across treatment groups only within each previous health status category; we discuss this kind of subclassification strategy in Section 10.2.

Another way to estimate the causal effect in this scenario is to regress the outcome on two inputs: the treatment indicator and previous health status. If health status is the only confounding covariate—that is, the only variable that predicts both the treatment and the outcome—and if the regression model is properly specified, then the coefficient of the treatment indicator corresponds to the average causal effect in the sample. In this example a simple way to avoid possible misspecification would be to discretize health status using indicator variables rather than including it as a single continuous predictor.

In general, then, causal effects can be estimated using regression if the model includes all confounding covariates (predictors that can affect treatment assignment or the outcome) and if the model is correct. If the confounding covariates are all observed (as in this example), then accurate estimation comes down to proper modeling and the extent to which the model is forced to extrapolate beyond the support of the data. If the confounding covariates are not observed (for example, if we suspect that healthier patients received the treatment, but no accurate measure of previous health status is included in the model), then they are "omitted" or "lurking" variables that complicate the quest to estimate causal effects.

We consider these issues in more detail in the rest of this chapter and the next, but first we will provide some intuition in the form of an algebraic formula.

Formula for omitted variable bias

We can quantify the bias incurred by excluding a confounding covariate in the context where a simple linear regression model is appropriate and there is only one confounding covariate. First define the "correct" specification as

$$y_i = \beta_0 + \beta_1 T_i + \beta_2 x_i + \epsilon_i \tag{9.1}$$

where T_i is the treatment and x_i is the covariate for unit i.

If instead the confounding covariate, x_i, is ignored, one can fit the model

$$y_i = \beta_0^* + \beta_1^* T_i + \epsilon_i^*$$

What is the relation between these models? To understand, it helps to define a third regression,

$$x_i = \gamma_0 + \gamma_1 T_i + \nu_i$$

If we substitute this representation of x into the original, correct, equation, and rearrange terms, we get

$$y_i = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)T_i + \epsilon_i + \beta_2\nu_i \tag{9.2}$$

Equating the coefficients of T in (9.1) and (9.2) yields

$$\beta_1^* = \beta_1 + \beta_2^*\gamma_1$$

This correspondence helps demonstrate the definition of a confounding covariate. If there is no association between the treatment and the purported confounder (that is, $\gamma_1 = 0$) or if there is no association between the outcome and the confounder (that is, $\beta_2 = 0$) then the variable is not a confounder because there will be no bias ($\beta_2^*\gamma_1 = 0$).

This formula is commonly presented in regression texts as a way of describing the bias that can be incurred if a model is specified incorrectly. However, this term has little meaning outside of a context in which one is attempting to make causal inferences.

9.2 The fundamental problem of causal inference

We begin by considering the problem of estimating the causal effect of a treatment compared to a control, for example in a medical experiment. Formally, the *causal effect* of a treatment T on an outcome y for an observational or experimental unit i can be defined by comparisons between the outcomes that would have occurred under each of the different treatment possibilities. With a binary treatment T taking on the value 0 (control) or 1 (treatment), we can define *potential outcomes*, y_i^0 and y_i^1 for unit i as the outcomes that would be observed under control and treatment conditions, respectively.[2](These ideas can also be directly generalized to the case of a treatment variable with multiple levels.)

The problem

For someone assigned to the treatment condition (that is, $T_i = 1$), y_i^1 is observed and y_i^0 is the unobserved *counterfactual* outcome—it represents what *would have* happened to the individual if assigned to control. Conversely, for control units, y_i^0 is observed and y_i^1 is counterfactual. In either case, a simple treatment effect for unit i can be defined as

$$\text{treatment effect for unit } i = y_i^1 - y_i^0$$

Figure 9.3 displays hypothetical data for an experiment with 100 units (and thus 200 potential outcomes). The top panel displays the data we would like to be able to see in order to determine causal effects for each person in the dataset—that is, it includes both potential outcomes for each person.

[2] The word "counterfactual" is sometimes used here, but we follow Rubin (1990) and use the term "potential outcome" because some of these potential data are actually observed.

(Hypothetical) complete data:

Unit, i	Pre-treatment inputs X_i			Treatment indicator T_i	Potential outcomes y_i^0	y_i^1	Treatment effect $y_i^1 - y_i^0$
1	2	1	50	0	**69**	*75*	*6*
2	3	1	98	0	**111**	*108*	*−3*
3	2	2	80	1	*92*	**102**	*10*
4	3	1	98	1	*112*	**111**	*−1*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	4	1	104	1	*111*	**114**	*3*

Observed data:

Unit, i	Pre-treatment inputs X_i			Treatment indicator T_i	Potential outcomes y_i^0	y_i^1	Treatment effect $y_i^1 - y_i^0$
1	2	1	50	0	69	?	?
2	3	1	98	0	111	?	?
3	2	2	80	1	?	102	?
4	3	1	98	1	?	111	?
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
100	4	1	104	1	?	114	?

Figure 9.3 *Illustration of the fundamental problem of causal inference. For each unit, we have observed some pre-treatment inputs, and then the treatment* ($T_i = 1$) *or control* ($T_i = 0$) *is applied. We can then observe only one of the potential outcomes,* (y_i^0, y_i^1)*. As a result, we cannot observe the treatment effect,* $y_i^1 - y_i^0$*, for any of the units.*
The top table shows what the complete data might look like, if it were possible to observe both potential outcomes on each unit. For each pair, the observed outcome is displayed in boldface. The bottom table shows what would actually be observed.

The so-called *fundamental problem of causal inference* is that at most one of these two potential outcomes, y_i^0 and y_i^1, can be observed for each unit i. The bottom panel of Figure 9.3 displays the data that can actually be observed. The y_i^1 values are "missing" for those in the control group and the y_i^0 values are "missing" for those in the treatment group.

Ways of getting around the problem

We cannot observe *both* what happens to an individual after taking the treatment (at a particular point in time) *and* what happens to that same individual after not taking the treatment (at the same point in time). Thus we can never measure a causal effect directly. In essence, then, we can think of causal inference as a prediction of what would happen to unit i if $T_i = 0$ or $T_i = 1$. It is thus predictive inference in the potential-outcome framework. Viewed this way, estimating causal effects requires one or some combination of the following: close substitutes for the potential outcomes, randomization, or statistical adjustment. We discuss the basic strategies here and go into more detail in the remainder of this chapter and the next.

Close substitutes. One might object to the formulation of the fundamental problem of causal inference by noting situations where it appears one can actually measure both y_i^0 and y_i^1 on the same unit. Consider, for example drinking tea one evening and milk another evening, and then measuring the amount of sleep each time. A careful consideration of this example reveals the implicit assumption that there are no systematic differences between days that could also affect sleep. An additional assumption is that applying the treatment on one day has no effect on the outcome on another day.

More pristine examples can generally be found in the natural and physical sciences. For instance, imagine dividing a piece of plastic into two parts and then exposing each piece to a corrosive chemical. In this case, the hidden assumption is that pieces are identical in how they would respond with and without treatment, that is, $y_1^0 = y_2^0$ and $y_1^1 = y_2^1$.

As a third example, suppose you want to measure the effect of a new diet by comparing your weight before the diet and your weight after. The hidden assumption here is that the pre-treatment measure can act as a substitute for the potential outcome under control, that is, $y_i^0 = x_i$.

It is not unusual to see studies that attempt to make causal inferences by substituting values in this way. It is important to keep in mind the strong assumptions often implicit in such strategies.

Randomization and experimentation. A different approach to causal inference is the "statistical" idea of using the outcomes observed on a sample of units to learn about the distribution of outcomes in the population.

The basic idea is that since we cannot compare treatment and control outcomes for the same units, we try to compare them on similar units. Similarity can be attained by using randomization to decide which units are assigned to the treatment group and which units are assigned to the control group. We will discuss this strategy in depth in the next section.

Statistical adjustment. For a variety of reasons, it is not always possible to achieve close similarity between the treated and control groups in a causal study. In observational studies, units often end up treated or not based on characteristics that are predictive of the outcome of interest (for example, men enter a job training program because they have low earnings and future earnings is the outcome of interest). Randomized experiments, however, can be impractical or unethical, and even in this context imbalance can arise from small-sample variation or from unwillingness or inability of subjects to follow the assigned treatment.

When treatment and control groups are not similar, modeling or other forms of statistical adjustment can be used to fill in the gap. For instance, by fitting a regression (or more complicated model), we may be able to estimate what would have happened to the treated units had they received the control, and vice versa. Alternately, one can attempt to divide the sample into subsets within which the treatment/control allocation mimics an experimental allocation of subjects. We discuss regression approaches in this chapter. We discuss imbalance and related issues more thoroughly in Chapter 10 along with a description of ways to help observational studies mimic randomized experiments.

9.3 Randomized experiments

We begin with the cleanest scenario, an experiment with units randomly assigned to receive treatment and control, and with the units in the study considered as a random sample from a population of interest. The random sampling and random

treatment assignment allow us to estimate the average causal effect of the treatment in the population, and regression modeling can be used to refine this estimate.

Average causal effects and randomized experiments

Although we cannot estimate individual-level causal effects (without making strong assumptions, as discussed previously), we can design studies to estimate the population average treatment effect:

$$\text{average treatment effect} = \text{avg}\,(y_i^1 - y_i^0),$$

for the units i in a larger population. The cleanest way to estimate the population average is through a randomized experiment in which each unit has a positive chance of receiving each of the possible treatments.[3] If this is set up correctly, with treatment assignment either entirely random or depending only on recorded data that are appropriately modeled, the coefficient for T in a regression corresponds to the causal effect of the treatment, among the population represented by the n units in the study.

Considered more broadly, we can think of the control group as a group of units that could just as well have ended up in the treatment group, they just happened not to get the treatment. Therefore, on average, their outcomes represent what would have happened to the treated units had they not been treated; similarly, the treatment group outcomes represent what might have happened to the control group had they been treated. Therefore the control group plays an essential role in a causal analysis.

For example, if n_0 units are selected at random from the population and given the control, and n_1 other units are randomly selected and given the treatment, then the observed sample averages of y for the treated and control units can be used to estimate the corresponding population quantities, $\text{avg}(y^0)$ and $\text{avg}(y^1)$, with their difference estimating the average treatment effect (and with standard error $\sqrt{s_0^2/n_0 + s_1^2/n_1}$; see Section 2.3). This works because the y_i^0's for the control group are a random sample of the values of y_i^0 in the entire population. Similarly, the y_i^1's for the treatment group are a random sample of the y_i^1's in the population.

Equivalently, if we select $n_0 + n_1$ units at random from the population, and then randomly assign n_0 of them to the control and n_1 to the treatment, we can think of each of the sample groups as representing the corresponding population of control or treated units. Therefore the control group mean can act as a counterfactual for the treatment group (and vice versa).

What if the $n_0 + n_1$ units are selected nonrandomly from the population but then the treatment is assigned at random within this sample? This is common practice, for example, in experiments involving human subjects. Experiments in medicine, for instance, are conducted on volunteers with specified medical conditions who are willing to participate in such a study, and experiments in psychology are often conducted on university students taking introductory psychology courses. In this case, causal inferences are still justified, but inferences no longer generalize to the entire population. It is usual instead to consider the inference to be appropriate to a hypothetical superpopulation from which the experimental subjects were drawn. Further modeling is needed to generalize to any other population. A study

[3] Ideally, each unit should have a nonzero probability of receiving each of the treatments, because otherwise the appropriate counterfactual (potential) outcome cannot be estimated for units in the corresponding subset of the population. In practice, if the probabilities are highly unequal, the estimated population treatment effect will have a high standard error due to the difficulty of reliably estimating such a rare event.

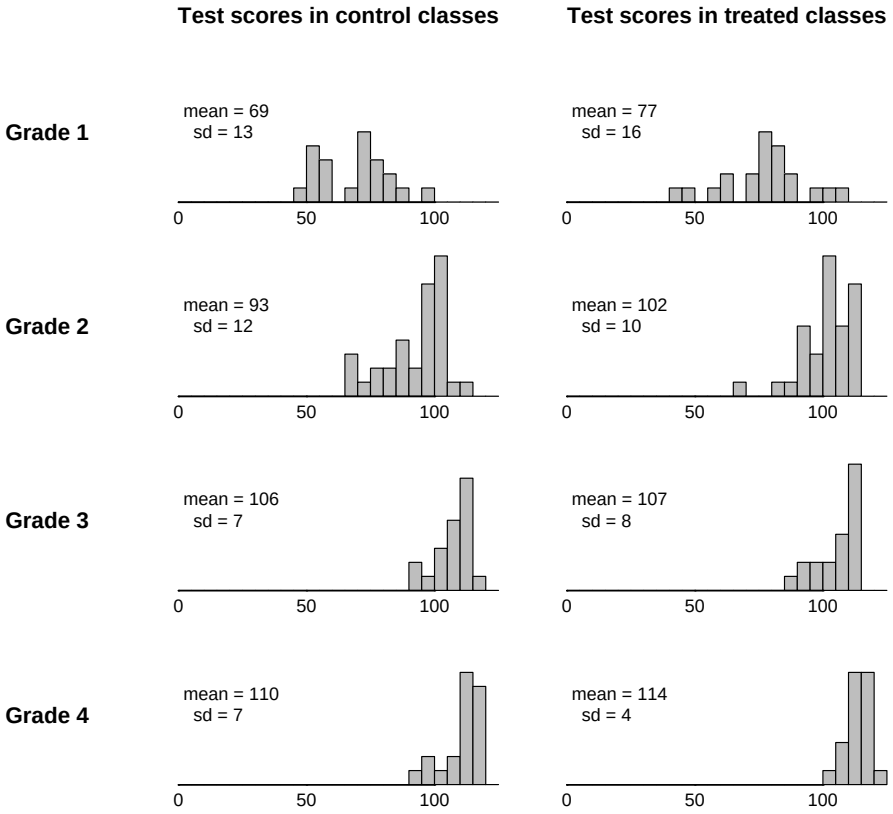

Figure 9.4 *Post-treatment test scores from an experiment measuring the effect of an educational television program, The Electric Company, on children's reading abilities. The experiment was applied on a total of 192 classrooms in four grades. At the end of the experiment, the average reading test score in each classroom was recorded.*

in which causal inferences are merited for a specific sample or population is said to have *internal validity*, and when those inferences can be generalized to a broader population of interest the study is said to have *external validity*.

We illustrate with a simple binary treatment (that is, two treatment levels, or a comparison of treatment to control) in an educational experiment. We then briefly discuss more general categorical, continuous, and multivariate treatments.

Example: showing children an educational television show

Figure 9.4 summarizes data from an educational experiment performed around 1970 on a set of elementary school classes. The treatment in this experiment was exposure to a new educational television show called The Electric Company. In each of four grades, the classes were randomized into treated and control groups. At the end of the school year, students in all the classes were given a reading test, and the average test score within each class was recorded. Unfortunately, we do not have data on individual students, and so our entire analysis will be at the classroom level.

Figure 9.4 displays the distribution of average post-treatment test scores in the control and treatment group for each grade. (The experimental treatment was applied to classes, not to schools, and so we treat the average test score in each class as

a single measurement.) We break up the data by grade for convenience and because it is reasonable to suppose that the effects of this show could vary by grade.

Analysis as a completely randomized experiment. The experiment was performed in two cities (Fresno and Youngstown). For each city and grade, the experimenters selected a small number of schools (10–20) and, within each school, they selected the two poorest reading classes of that grade. For each pair, one of these classes was randomly assigned to continue with its regular reading course and the other was assigned to view the TV program.

This is called a *paired comparisons* design (which in turn is a special case of a *randomized block* design, with exactly two units within each block). For simplicity, however, we shall analyze the data here as if the treatment assignment had been completely randomized within each grade. In a *completely randomized experiment* on n units (in this case, classrooms), one can imagine the units mixed together in a bag, completely mixed, and then separated into two groups. For example, the units could be labeled from 1 to n, and then permuted at random, with the first n_1 units receiving the treatment and the others receiving the control. Each unit has the same probability of being in the treatment group and these probabilities are independent of each other.

Again, for the rest of this chapter we pretend that the Electric Company experiment was completely randomized within each grade. In Section 23.1 we return to the example and present an analysis appropriate to the paired design that was actually used.

Basic analysis of a completely randomized experiment

When treatments are assigned completely at random, we can think of the different treatment groups (or the treatment and control groups) as a set of random samples from a common population. The population average under each treatment, $\text{avg}(y^0)$ and $\text{avg}(y^1)$, can then be estimated by the sample average, and the population average difference between treatment and control, $\text{avg}(y^1) - \text{avg}(y^0)$—that is, the average causal effect—can be estimated by the difference in sample averages, $\bar{y}_1 - \bar{y}_0$.

Equivalently, the average causal effect of the treatment corresponds to the coefficient θ in the regression, $y_i = \alpha + \theta T_i + \text{error}_i$. We can easily fit the four regressions (one for each grade) in R:

R code
```
for (k in 1:4) {
  display (lm (post.test ~ treatment, subset=(grade==k)))
}
```

The estimates and uncertainty intervals for the Electric Company experiment are graphed in the left panel of Figure 9.5. The treatment appears to be generally effective, perhaps more so in the low grades, but it is hard to be sure given the large standard errors of estimation.

Controlling for pre-treatment predictors

In this study, a pre-test was given in each class at the beginning of the school year (before the treatment was applied). In this case, the treatment effect can also be estimated using a regression model: $y_i = \alpha + \theta T_i + \beta x_i + \text{error}_i$ on the pre-treatment predictor x.[4] Figure 9.6 illustrates for the Electric Company experiment. For each

[4] We avoid the term *confounding covariates* when describing adjustment in the context of a randomized experiment. Predictors are included in this context to increase precision. We expect

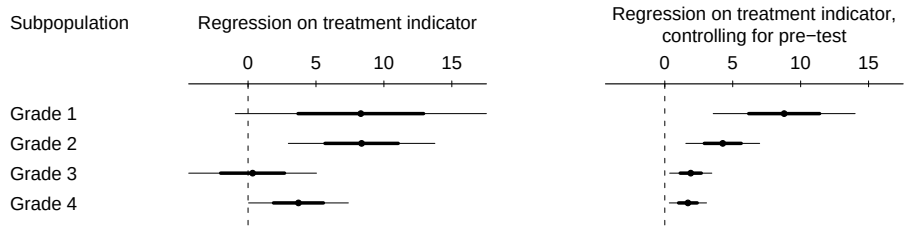

Figure 9.5 *Estimates, 50%, and 95% intervals for the effect of the Electric Company television show (see data in Figures 9.4 and 9.6) as estimated in two ways: first, from a regression on treatment alone, and second, also controlling for pre-test data. In both cases, the coefficient for treatment is the estimated causal effect. Including pre-test data as a predictor increases the precision of the estimates.*
Displaying these coefficients and intervals as a graph facilitates comparisons across grades and across estimation strategies (controlling for pre-test or not). For instance, the plot highlights how controlling for pre-test scores increases precision and reveals decreasing effects of the program for the higher grades, a pattern that would be more difficult to see in a table of numbers.
Sample sizes are approximately the same in each of the grades. The estimates for higher grades have lower standard errors because the residual standard deviations of the regressions are lower in these grades; see Figure 9.6.

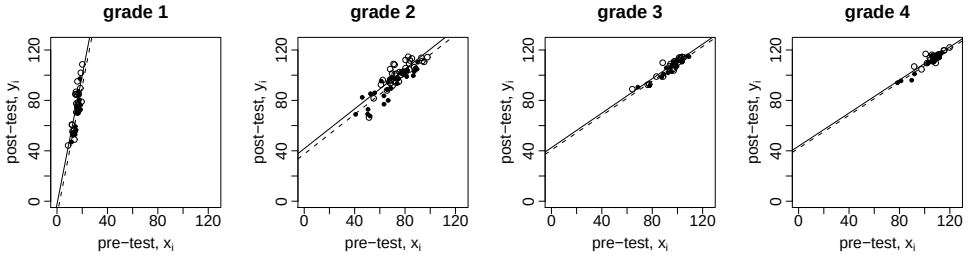

Figure 9.6 *Pre-test/post-test data for the Electric Company experiment. Treated and control classes are indicated by circles and dots, respectively, and the solid and dotted lines represent parallel regression lines fit to the treatment and control groups, respectively. The solid lines are slightly higher than the dotted lines, indicating slightly positive estimated treatment effects. Compare to Figure 9.4, which displays only the post-test data.*

grade, the difference between the regression lines for the two groups represents the treatment effect as a function of pre-test score. Since we have not included any interaction in the model, this treatment effect is assumed constant over all levels of the pre-test score.

For grades 2–4, the pre-test was the same as the post-test, and so it is no surprise that all the classes improved whether treated or not (as can be seen from the plots). For grade 1, the pre-test was a subset of the longer test, which explains why the pre-test scores for grade 1 are so low. We can also see that the distribution of post-test scores for each grade is similar to the next grade's pre-test scores, which makes sense.

In any case, for estimating causal effects (as defined in Section 9.2) we are interested in the difference between treatment and control conditions, not in the simple improvement from pre-test to post-test. The pre-post improvement is not a

them to be related to the outcome but not to the treatment assignment due to the randomization. Therefore they are not confounding covariates.

causal effect (except under the assumption, unreasonable in this case, that under the control there would be no change from pre-post change).

In the regression

$$y_i = \alpha + \theta T_i + \beta x_i + \text{error}_i \tag{9.3}$$

the coefficient for the treatment indicator still represents the average treatment effect, but controlling for pre-test can improve the efficiency of the estimate. (More generally, the regression can control for multiple pre-treatment predictors, in which case the model has the form $y_i = \alpha + \theta T_i + X_i\beta + \text{error}_i$, or alternatively α can be removed from the equation and considered as a constant term in the linear predictor $X\beta$.)

The estimates for the Electric Company study appear in the right panel of Figure 9.5. It is now clear that the treatment is effective, and it appears to be more effective in the lower grades. A glance at Figure 9.6 suggests that in the higher grades there is less room for improvement; hence this particular test might not be the most effective for measuring the benefits of The Electric Company in grades 3 and 4.

It is only appropriate to control for pre-treatment predictors, or, more generally, predictors that would not be affected by the treatment (such as race or age). This point will be illustrated more concretely in Section 9.7.

Gain scores

An alternative way to specify a model that controls for pre-test measures is to use these measures to transform the response variable. A simple approach is to subtract the pre-test score, x_i, from the outcome score, y_i, thereby creating a "gain score," g_i. Then this score can be regressed on the treatment indicator (and other predictors if desired), $g_i = \alpha + \theta T_i + \text{error}_i$. (In the simple case with no other predictors, the regression estimate is simply $\hat{\theta} = \bar{g}^T - \bar{g}^C$, the average difference of gain scores in the treatment and control groups.)

In some cases the gain score can be more easily interpreted than the original outcome variable y. Using gain scores is most effective if the pre-treatment score is comparable to the post-treatment measure. For instance, in our Electric Company example it would not make sense to create gain scores for the classes in grade 1 since their pre-test measure was based on only a subset of the full test.

One perspective on this model is that it makes an unnecessary assumption, namely, that $\beta = 1$ in model (9.3). On the other hand, if this assumption is close to being true then θ may be estimated more precisely. One way to resolve this concern about misspecification would simply be to include the pre-test score as a predictor as well, $g_i = \alpha + \theta T_i + \gamma x_i + \text{error}_i$. However, in this case, $\hat{\theta}$, the estimate of the coefficient for T, is equivalent to the estimated coefficient from the original model, $y_i = \alpha + \theta T_i + \beta x_i + \text{error}_i$ (see Exercise 9.7).

More than two treatment levels, continuous treatments, and multiple treatment factors

Going beyond a simple treatment-and-control setting, multiple treatment effects can be defined relative to a baseline level. With random assignment, this simply follows general principles of regression modeling.

If treatment levels are numerical, the treatment level can be considered as a continuous input variable. To conceptualize randomization with a continuous treatment variable, think of choosing a random number that falls anywhere in the continuous range. As with regression inputs in general, it can make sense to fit more compli-

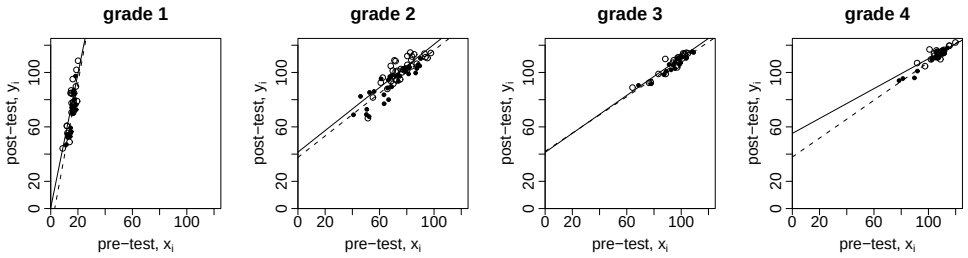

Figure 9.7 *Pre-test/post-test data for the Electric Company experiment. Treated and control classes are indicated by circles and dots, respectively, and the solid and dotted lines represent separate regression lines fit to the treatment and control groups, respectively. For each grade, the difference between the solid and dotted lines represents the estimated treatment effect as a function of pre-test score.*

cated models if suggested by theory or supported by data. A linear model—which estimates the average effect on y for each additional unit of T—is a natural starting point, though it may need to be refined.

With several discrete treatments that are unordered (such as in a comparison of three different sorts of psychotherapy), we can move to multilevel modeling, with the group index indicating the treatment assigned to each unit, and a second-level model on the group coefficients, or treatment effects. We shall illustrate such modeling in Section 13.5 with an experiment from psychology. We shall focus more on multilevel modeling as a tool for fitting data, but since the treatments in that example are randomly assigned, their coefficients can be interpreted as causal effects.

Additionally, different combinations of multiple treatments can be administered randomly. For instance, depressed individuals could be randomly assigned to receive nothing, drugs, counseling sessions, or a combination of drugs and counseling sessions. These combinations could be modeled as two treatments and their interaction or as four distinct treatments.

The assumption of no interference between units

Our discussion so far regarding estimation of causal effects using experiments is contingent upon another, often overlooked, assumption. We must assume also that the treatment assignment for one individual (unit) in the experiment does not affect the outcome for another. This has been incorporated into the "stable unit treatment value assumption" (SUTVA). Otherwise, we would need to define a different potential outcome for the i^{th} unit not just for each treatment received by that unit but for each combination of treatment assignments received by every other unit in the experiment. This would enormously complicate even the definition, let alone the estimation, of individual causal effects. In settings such as agricultural experiments where interference between units is to be expected, it can be modeled directly, typically using spatial interactions.

9.4 Treatment interactions and poststratification

Interactions of treatment effect with pre-treatment inputs

Once we include pre-test in the model, it is natural to allow it to interact with treatment effect. The treatment is then allowed to affect both the intercept and the slope of the pre-test/post-test regression. Figure 9.7 shows the Electric Company

data with separate regression lines estimated for the treatment and control groups. As with Figure 9.6, for each grade the difference between the regression lines is the estimated treatment effect as a function of pre-test score.

We illustrate in detail for grade 4. First, we fit the simple model including only the treatment indicator:

```
lm(formula = post.test ~ treatment, subset=(grade==4))                R output
            coef.est coef.se
(Intercept)    110.4     1.3
treatment        3.7     1.8
  n = 42, k = 2
  residual sd = 6.0, R-Squared = 0.09
```

The estimated treatment effect is 3.7 with a standard error of 1.8. We can improve the efficiency of the estimator by controlling for the pre-test score:

```
lm(formula = post.test ~ treatment + pre.test, subset=(grade==4))     R output
            coef.est coef.se
(Intercept)     42.0     4.3
treatment        1.7     0.7
pre.test         0.7     0.0
  n = 42, k = 3
  residual sd = 2.2, R-Squared = 0.88
```

The new estimated treatment effect is 1.7 with a standard error of 0.7. In this case, controlling for the pre-test reduced the estimated effect. Under a clean randomization, controlling for pre-treatment predictors in this way should reduce the standard errors of the estimates.[5] (Figure 9.5 shows the estimates for the Electric Company experiment in all four grades.)

Complicated arise when we include the interaction of treatment with pre-test:

```
lm(formula = post.test ~ treatment + pre.test + treatment:pre.test,   R output
    subset=(grade==4))
                   coef.est coef.se
(Intercept)        37.84     4.90
treatment          17.37     9.60
pre.test            0.70     0.05
treatment:pre.test -0.15     0.09
  n = 42, k = 4
  residual sd = 2.1, R-Squared = 0.89
```

The estimated treatment effect is now $17 - 0.15x$, which is difficult to interpret without knowing the range of x. From Figure 9.7 we see that pre-test scores range from approximately 80 to 120; in this range, the estimated treatment effect varies from $17-0.15\cdot 80 = 5$ for classes with pre-test scores of 80 to $17-0.15\cdot 120 = -1$ for classes with pre-test scores of 120. This range represents the *variation* in estimated treatment effects as a function of pre-test score, *not* uncertainty in the estimated treatment effect.

To get a sense of the uncertainty, we can plot the estimated treatment effect as a function of x, overlaying random simulation draws to represent uncertainty:

[5] Under a clean randomization, controlling for pre-treatment predictors in this way does not change what we are estimating. If the randomization was less than pristine, however, the addition of predictors to the equation may help us control for unbalanced characteristics across groups. Thus, this strategy has the potential to move us from estimating a noncausal estimand (due to lack of randomization) to estimating a causal estimand by in essence "cleaning" the randomization.

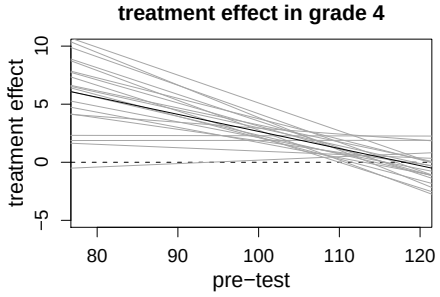

Figure 9.8 *Estimate and uncertainty for the effect of viewing The Electric Company (compared to the control treatment) for fourth-graders. Compare to the data in the rightmost plot in Figure 9.7. The dark line here—the estimated treatment effect as a function of pre-test score—is the difference between the two regression lines in the grade 4 plot in Figure 9.7. The gray lines represent 20 random draws from the uncertainty distribution of the treatment effect.*

R code

```
lm.4 <- lm (post.test ~ treatment + pre.test + treatment:pre.test,
  subset=(grade==4))
lm.4.sim <- sim (lm.4)
plot (0, 0, xlim=range (pre.test[grade==4]), ylim=c(-5,10),
  xlab="pre-test", ylab="treatment effect",
  main="treatment effect in grade 4")
abline (0, 0, lwd=.5, lty=2)
for (i in 1:20){
  curve (lm.4.sim$beta[i,2] + lm.4.sim$beta[i,4]*x, lwd=.5, col="gray",
    add=TRUE)}
curve (coef(lm.4)[2] + coef(lm.4)[4]*x, lwd=.5, add=TRUE)
```

This produces the graph shown in Figure 9.8.

Finally, we can estimate a mean treatment effect by averaging over the values of x in the data. If we write the regression model as $y_i = \alpha + \theta_1 T_i + \beta x_i + \theta_2 T_i x_i + \text{error}_i$, then the treatment effect is $\theta_1 + \theta_2 x$, and the summary treatment effect in the sample is $\frac{1}{n}\sum_{i=1}^{n}(\theta_1 + \theta_2 x_i)$, averaging over the n fourth-grade classrooms in the data. We can compute the average treatment effect as follows:

R code

```
n.sims <- nrow(lm.4.sim$beta)
effect <- array (NA, c(n.sims, sum(grade==4)))
for (i in 1:n.sims){
  effect[i,] <- lm.4.sim$beta[i,2] + lm.4.sim$beta[i,4]*pre.test[grade==4]
}
avg.effect <- rowMeans (effect)
```

The `rowMeans()` function averages over the grade 4 classrooms, and the result of this computation, `avg.effect`, is a vector of length `n.sims` representing the uncertainty in the average treatment effect. We can summarize with the mean and standard error:

R code

```
print (c (mean(avg.effect), sd(avg.effect)))
```

The result is 1.8 with a standard deviation of 0.7—quite similar to the result from the model controlling for pre-test but with no interactions. In general, for a linear regression model, the estimate obtained by including the interaction, and then averaging over the data, reduces to the estimate with no interaction. The motivation for including the interaction is thus to get a better idea of how the treatment effect varies with pre-treatment predictors, not simply to estimate an average effect.

Poststratification

We have discussed how treatment effects interact with pre-treatment predictors (that is, regression inputs). To estimate an average treatment effect, we can poststratify—that is, average over the population.[6]

For example, suppose we have treatment variable T and pre-treatment control variables x_1, x_2, and our regression predictors are x_1, x_2, T, and the interactions x_1T and x_2T, so that the linear model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 T + \beta_4 x_1 T + \beta_5 x_2 T +$ error. The estimated treatment effect is then $\beta_3 + \beta_4 x_1 + \beta_5 x_2$, and its average, in a linear regression, is simply $\beta_3 + \beta_4 \mu_1 + \beta_5 \mu_2$, where μ_1 and μ_2 are the averages of x_1 and x_2 in the population. These population averages might be available from another source, or else they can be estimated using the averages of x_1 and x_2 in the data at hand. Standard errors for summaries such as $\beta_3 + \beta_4 \mu_1 + \beta_5 \mu_2$ can be determined analytically, but it is easier to simply compute them using simulations.

Modeling interactions is important when we care about differences in the treatment effect for different groups, and poststratification then arises naturally if a population average estimate is of interest.

9.5 Observational studies

In theory, the simplest solution to the fundamental problem of causal inference is, as we have described, to randomly sample a different set of units for each treatment group assignment from a common population, and then apply the appropriate treatments to each group. An equivalent approach is to randomly assign the treatment conditions among a selected set of units. Either of these approaches ensures that, on average, the different treatment groups are *balanced* or, to put it another way, that the $\bar{y}^0$ and $\bar{y}^1$ from the sample are estimating the average outcomes under control and treatment for the same population.

In practice, however, we often work with *observational data* because, compared to experiments, observational studies can be more practical to conduct and can have more realism with regard to how the program or treatment is likely to be "administered" in practice. As we have discussed, however, in observational studies treatments are observed rather than assigned (for example, comparisons of smokers to nonsmokers), and it is not at all reasonable to consider the observed data under different treatments as random samples from a common population. In an observational study, there can be systematic differences between groups of units that receive different treatments—differences that are outside the control of the experimenter—and they can affect the outcome, y. In this case we need to rely on more data than just treatments and outcomes and implement a more complicated analysis strategy that will rely upon stronger assumptions. The strategy discussed in this chapter, however, is relatively simple and relies on controlling for confounding covariates through linear regression. Some alternative approaches are described in Chapter 10.

[6] In survey sampling, *stratification* refers to the procedure of dividing the population into disjoint subsets (strata), sampling separately within each stratum, and then combining the stratum samples to get a population estimate. Poststratification is the analysis of an unstratified sample, breaking the data into strata and reweighting as would have been done had the survey actually been stratified. Stratification can adjust for potential differences between sample and population using the survey design; poststratification makes such adjustments in the data analysis.

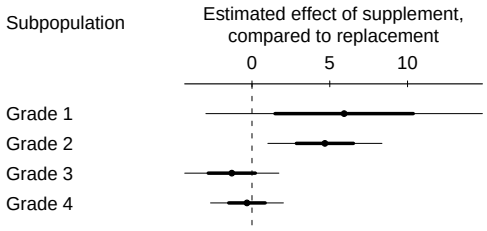

Figure 9.9 *Estimates, 50%, and 95% intervals for the effect of The Electric Company as a supplement rather than a replacement, as estimated by a regression on the supplement/replacement indicator also controlling for pre-test data. For each grade, the regression is performed only on the treated classes; this is an observational study embedded in an experiment.*

Electric Company example

Here we illustrate an observational study for which a simple regression analysis, controlling for pre-treatment information, may yield reasonable causal inferences.

The educational experiment described in Section 9.3 actually had an embedded observational study. Once the treatments had been assigned, the teacher for each class assigned to the Electric Company treatment chose to either *replace* or *supplement* the regular reading program with the Electric Company television show. That is, all the classes in the treatment group watched the show, but some watched it instead of the regular reading program and others got it in addition.[7]

The simplest starting point to analyzing these observational data (now limited to the randomized treatment group) is to consider the choice between the two treatment options—"replace" or "supplement"—to be randomly assigned conditional on pre-test scores. This is a strong assumption but we use it simply as a starting point. We can then estimate the treatment effect by regression, as with an actual experiment. In the R code, we create a variable called `supp` that equals 0 for the replacement form of the treatment, 1 for the supplement, and NA for the controls. We then estimate the effect of the supplement, as compared to the replacement, for each grade:

R code

```
for (k in 1:4) {
  ok <- (grade==k) & (!is.na(supp))
  lm.supp <- lm (post.test ~ supp + pre.test, subset=ok)
}
```

The estimates are graphed in Figure 9.9. The uncertainties are high enough that the comparison is inconclusive except in grade 2, but on the whole the pattern is consistent with the reasonable hypothesis that supplementing is more effective than replacing in the lower grades.

Assumption of ignorable treatment assignment

As opposed to making the same assumption as the completely randomized experiment, the key assumption underlying the estimate is that, *conditional* on the confounding covariates used in the analysis (here as inputs in the regression analysis), the distribution of units across treatment conditions is, in essence, "random"

[7] This procedural detail reveals that the treatment effect for the randomized experiment is actually more complicated than described earlier. As implemented, the experiment estimated the effect of making the program available, either as a supplement or replacement for the current curriculum.

(in this case, pre-test score) with respect to the potential outcomes. To help with the intuition here, one could envision units being randomly assigned to treatment conditions conditional on the confounding covariates; however, of course, no actual randomized assigment need take place.

Ignorability is often formalized by the conditional independence statement,

$$y^0, y^1 \perp T \,|\, X.$$

This says that the distribution of the potential outcomes, (y^0, y^1), is the same across levels of the treatment variable, T, once we condition on confounding covariates X.

This assumption is referred to as *ignorability* of the treatment assignment in the statistics literature and *selection on observables* in econometrics. Said another way, we would not necessarily expect any two classes to have had the same probability of receiving the supplemental version of the treatment. However, we expect any two classes at the same levels of the confounding covariates (that is, pre-treatment variables; in our example, average pre-test score) to have had the same probability of receiving the supplemental version of the treatment. A third way to think about the ignorability assumption is that it requires that we control for all confounding covariates, the pre-treatment variables that are associated with both the treatment and the outcome.

If ignorability holds, then causal inferences can be made without modeling the treatment assignment process—that is, we can *ignore* this aspect of the model as long as analyses regarding the causal effects condition on the predictors needed to satisfy ignorability. Randomized experiments represent a simple case of ignorability. Completely randomized experiments need not condition on any pre-treatment variables—this is why we can use a simple difference in means to estimate causal effects. Randomized experiments that block or match satisfy ignorability conditional on the design variables used to block or match, and therefore these variables need to be included when estimating causal effects.

In the Electric Company supplement/replacement example, an example of a *nonignorable assignment mechanism* would be if the teacher of each class chose the treatment that he or she believed would be more effective for that particular class based on unmeasured characteristics of the class that were related to their subsequent test scores. Another nonignorable assignment mechanism would be if, for example, supplementing was more likely to be chosen by more "motivated" teachers, with teacher motivation also associated with the students' future test scores.

For ignorability to hold, it is not necessary that the two treatments be equally likely to be picked, but rather that the probability that a given treatment is picked should be equal, conditional on our confounding covariates.[8] In an experiment, one can control this at the design stage by using a random assignment mechanism. In an observational study, the "treatment assignment" is not under the control of the statistician, but one can aim for ignorability by conditioning in the analysis stage on as much pre-treatment information in the regression model as possible. For example, if teachers' motivation might affect treatment assignment, it would be advisable to have a pre-treatment measure of teacher motivation and include this as an input in the regression model. This would increase the plausibility of the ignorability assumption. Realistically, this may be a difficult characteristic to

[8] As further clarification, consider two participants of a study for which ignorability holds. If we define the probability of treatment participation as $\Pr(T = 1|X)$, then this probability must be equal for these two individuals. However, suppose there exists another variable, w, that is associated with treatment participation (conditional on X) but not with the outcome (conditional on X). We do not require that $\Pr(T = 1 \,|\, X, W)$ be the same for these two participants.

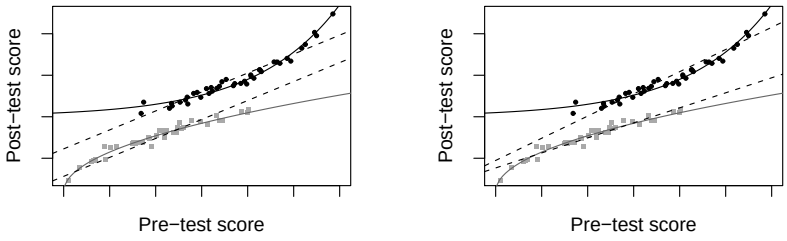

Figure 9.10 *Hypothetical before/after data demonstrating the potential problems in using linear regression for causal inference. The dark dots and line correspond to the children who received the educational supplement; the lighter dots and line correspond to the children who did not receive the supplement. The dashed lines are regression lines fit to the observed data. The model shown in the right panel allows for an interaction between receiving the supplement and pre-test scores.*

measure, but other teacher characteristics such as years of experience and schooling might act as partial proxies.

In general, one can never prove that the treatment assignment process in an observational study is ignorable—it is always possible that the choice of treatment depends on relevant information that has not been recorded. In an educational study this information could be characteristics of the teacher or school that are related both to treatment assignment and to post-treatment test scores. Thus, if we interpret the estimates in Figure 9.9 as causal effects, we do so with the understanding that we would prefer to have further pre-treatment information, especially on the teachers, in order to be more confident in ignorability.

If we believe that treatment assignments depend on information not included in the model, then we should choose a different analysis strategy. We discuss some options at the end of the next chapter.

Judging the reasonableness of regression as a modeling approach, assuming ignorability

Even if the ignorability assumption appears to be justified, this does not mean that simple regression of our outcomes on confounding covariates and a treatment indicator is necessarily the best modeling approach for estimating treatment effects. There are two primary concerns related to the distributions of the confounding covariates across the treatment groups: lack of complete overlap and lack of balance. For instance, consider our initial hypothetical example of a medical treatment that is supposed to affect subsequent health measures. What if there were no treatment observations among the group of people whose pre-treatment health status was highest? Arguably, we could not make any causal inferences about the effect of the treatment on these people because we would have no empirical evidence regarding the counterfactual state. Lack of overlap and balance forces stronger reliance on our modeling than if covariate distributions were the same across treatment groups. We provide a brief illustration in this chapter and discuss in greater depth in Chapter 10.

Suppose we are interested in the effect of a supplementary educational activity (such as viewing The Electric Company) that was not randomly assigned. Suppose, however, that only one predictor, pre-test score, is necessary to satisfy ignorability—that is, there is only one confounding covariate. Suppose further, though, that those individuals who participate in the supplementary activity tend to have higher pre-

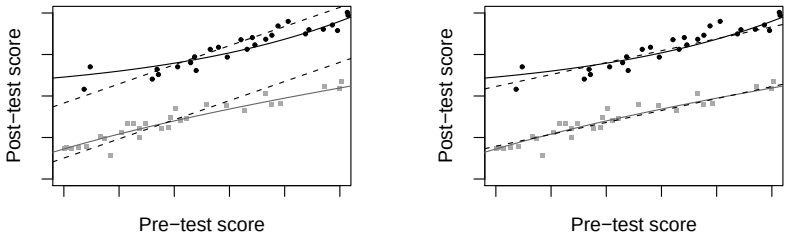

Figure 9.11 *Hypothetical before/after data demonstrating the potential problems in using linear regression for causal inference. The dark dots and line correspond to the children who received the educational supplement; the lighter dots and line correspond to the children who did not receive the supplement. The dashed lines are regression lines fit to the observed data. Plots are restricted to observations in the region where there is overlap in terms of the pre-treatment test score across treatment and control groups. The left panel shows only the portion of the plot in Figure 9.10 where there is overlap. The right panel shows regression lines fit only using observations in this overlapping region.*

test scores, on average, than those who do not participate. One realization of this hypothetical scenario is illustrated in Figure 9.10. The dark line represents the true relation between pre-test scores (x-axis) and post-test scores (y-axis) for those who receive the supplement. The lighter line represents the true relation between pre-test scores and post-test scores for those who do not receive the supplement. Estimated linear regression lines are superimposed for these data. The linear model has problems fitting the true nonlinear regression relation—a problem that is compounded by the lack of overlap of the two groups in the data. Because there are no "control" children with high test scores and virtually no "treatment" children with low test scores, these linear models, to create counterfactual predictions, are forced to extrapolate over portions of the space where there are no data to support them. These two problems combine to create, in this case, a substantial underestimate of the true average treatment effect. Allowing for an interaction, as illustrated in the right panel, does not solve the problem.

In the region of pre-test scores where there are observations from both treatment groups, however, even the incorrectly specified linear regression lines do not provide such a bad fit to the data. And no model extrapolation is required, so diagnosing this lack of fit would be possible. This is demonstrated in the left panel of Figure 9.11 by restricting the plot from the left panel of Figure 9.10 to the area of overlap. Furthermore, if the regression lines are fit only using this restricted sample they fit quite well in this region, as is illustrated in the right panel of Figure 9.11. Some of the strategies discussed in the next chapter use this idea of limiting analyses to observations with the region of complete overlap.

Examining overlap in the Electric Company embedded observational study

For the Electric Company data we can use plots such as in Figure 9.10–9.11 to assess the appropriateness of the modeling assumptions and the extent to which we are relying on unsupported model extrapolations. For the most part, Figure 9.12 reveals a reasonable amount of overlap in pre-test scores across treatment groups within each grade. Grade 3, however, has some classrooms with average pre-test scores that are lower than the bulk of the sample, all of which received the supplement. It might be appropriate to decide that no counterfactual classrooms exist in our data for these classrooms and thus the data cannot support causal inferences for these

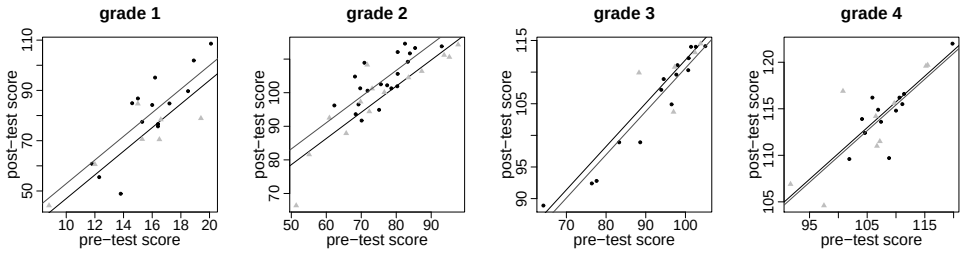

Figure 9.12 *Pre-test/post-test data examining the overlap in pre-test scores across treatment groups as well as the extent to which models are being extrapolated to regions where there is no support in the data. Classrooms that watched The Electric Company as a supplement are represented by the dark points and regression line; classrooms that watched The Electric Company as a replacement are represented by the lighter points and regression line. No interactions were included when estimating the regression lines.*

classrooms. The sample sizes for each grade make it difficult to come to any firm conclusions one way or another, however.

Therefore, we must feel confident in the (probably relatively minor) degree of model extrapolation relied upon by these estimates in order to trust a causal interpretation.

9.6 Understanding causal inference in observational studies

Sometimes the term "observational study" refers to a situation in which a specific intervention was offered nonrandomly to a population or in which a population was exposed nonrandomly to a well-defined treatment. The primary characteristic that distinguishes causal inference in these settings from causal inference in randomized experiments is the inability to identify causal effects without making assumptions such as ignorability. (Other sorts of assumptions will be discussed in the next chapter.)

Often, however, observational studies refer more broadly to survey data settings where no intervention has been performed. In these settings, there are other aspects of the research design that need to be carefully considered as well. The first is the mapping between the "treatment" variable in the data and a policy or intervention. The second considers whether it is possible to separately identify the effects of multiple treatment factors. When attempting causal inference using observational data, it is helpful to formalize exactly what the experiment might have been that would have generated the data, as we discuss next.

Defining a "treatment" variable

A causal effect needs to be defined with respect to a cause, or an intervention, on a particular set of experimental units. We need to be able to conceive of each unit as being able to experience each level of the treatment variable for which causal effects will be defined for that unit. Thus, the "effect" of height on earnings is ill-defined without reference to a treatment that could change one's height. Otherwise what does it mean to define a potential outcome for a person that would occur *if* he or she had been shorter or taller?

More subtly, consider the effect of single-motherhood on children's outcomes. We might be able to envision several different kinds of interventions that could change

a mother's marital status either before or after birth: changes in tax laws, participation in a marriage encouragement program for unwed parents, new child support enforcement policies, divorce laws, and so on. These potential "treatments" vary in the timing of marriage relative to birth and even the strength of the marriages that might result, and consequently might be expected to have different effects on the children involved. Therefore, this conceptual mapping to a hypothetical intervention can be important for choice of study design, analysis, and interpretation of results.

Consider, for instance, a study that examines Korean children who were randomly assigned to American families for adoption. This "natural experiment" allows for fair comparisons across conditions such as being raised in one-parent versus two-parent households. However, this is a different kind of treatment altogether than considering whether a couple should get married. There is no attempt to compare *parents* who are similar to each other; instead, it is the *children* who are similar on average at the outset. The treatment in question then has to do with the child's placement in a family. This addresses an interesting although perhaps less policy-relevant question (at least in terms of policies that affect incentives for marriage formation or dissolution).

Multiple treatment factors

It is difficult to directly interpret more than one input variable causally in an observational study. Suppose we have two variables, A and B, whose effects we would like to estimate from a single observational study. To estimate causal effects, we must consider implicit treatments—and to estimate both effects at once, we would have to imagine a treatment that affects A while leaving B unchanged, and a treatment that affects B while leaving A unchanged. In examples we have seen, it is generally difficult to envision both these interventions: if A comes before B in time or logical sequence, then we can estimate the effect of B controlling for A but not the reverse (because of the problem with controlling for post-treatment variables, which we discuss in greater detail in the next section).

More broadly, for many years a common practice when studying a social problem (for example, poverty) was to compare people with different outcomes, throwing many inputs into a regression to see which was the strongest predictor. As opposed to the way we have tried to frame causal questions thus far in this chapter, as the effect of causes, this is a strategy that searches for the causes of an effect. This is an ill-defined notion that we will avoid for exactly the kind of reasons discussed in this chapter.[9]

Thought experiment: what would be an ideal randomized experiment?

If you find yourself confused about what can be estimated and how the various aspects of your study should be defined, a simple strategy is to try to formalize the randomized experiment you would have liked to have done to answer your causal question. A perfect mapping rarely exists between this experimental ideal and your data so often you will be forced instead to figure out, given the data you have, what randomized experiment could be thought to have generated such data.

[9] Also, philosophically, looking for the most important cause of an outcome is a confusing framing for a research question because one can always find an earlier cause that affected the "cause" you determine to be the strongest from your data. This phenomenon is sometimes called the "infinite regress of causation."

For instance, if you were interested in the effect of breastfeeding on children's cognitive outcomes, what randomized experiment would you want to perform assuming no practical, legal, or moral barriers existed? We could imagine randomizing mothers to either breastfeed their children exclusively or bottle-feed them formula exclusively. We would have to consider how to handle those who do not adhere to their treatment assignment, such as mothers and children who are not able to breastfeed, and children who are allergic to standard formula. Moreover, what if we want to separately estimate the physiological effects of the breast milk from the potential psychological implications (to both mother and child) of nursing at the breast and the more extended physical contact that is often associated with breastfeeding? In essence, then, we think that perhaps breastfeeding represents several concurrent treatments. Perhaps we would want to create a third treatment group of mothers who feed their babies with bottles of expressed breast milk. This exercise of considering the randomized experiment helps to clarify what the true nature of the intervention is that we are using our treatment variable to represent.

Just as in a randomized experiment, all causal inference requires a comparison of at least two treatments (counting "control" as a treatment). For example, consider a study of the effect on weight loss of a new diet. The treatment (following the diet) may be clear but the control is not. Is it to try a different diet? To continue eating "normally"? To exercise more? Different control conditions imply different counterfactual states and thus induce different causal effects.

Finally, thinking about hypothetical randomized experiments can help with problems of trying to establish a causal link between two variables when neither has temporal priority and when they may have been simultaneously determined. For instance, consider a regression of crime rates in each of 50 states using a cross section of data, where the goal is to determine the "effect" of the number of police officers while controlling for the social, demographic, and economic features of each state as well as characteristics of the state (such as the crime rate) that might affect decisions to increase the size of the police force. The problem is that it may be difficult (if not impossible) to disentangle the "effect" of the size of the police force on crime from the "effect" of the crime rate on the size of the police force.

If one is interested in figuring out policies that can affect crime rates, it might be more helpful to conceptualize both "number of police officers" and "crime rate" as outcome variables. Then one could imagine different treatments (policies) that could affect these outcomes. For example, the number of police officers could be affected by a bond issue to raise money earmarked for hiring new police, or a change in the retirement age, or a reallocation of resources within local and state government law enforcement agencies. These different treatments could have different effects on the crime rate.

9.7 Do not control for post-treatment variables

As illustrated in the examples of this chapter, we recommend controlling for pretreatment covariates when estimating causal effects in experiments and observational studies. However, it is generally not a good idea to control for variables measured *after* the treatment. In this section and the next we explain why controlling for a post-treatment variable messes up the estimate of total treatment effect, and also the difficulty of using regression on "mediators" or "intermediate outcomes" (variables measured post-treatment but generally prior to the primary outcome of interest) to estimate so-called mediating effects.

Consider a hypothetical study of a treatment that incorporates a variety of social

unit, i	treatment, T_i	observed intermediate outcome, z_i	potential intermediate outcomes, z_i^0	z_i^1	final outcome, y_i
1	0	0.5	**0.5**	*0.7*	y_1
2	1	0.5	*0.3*	**0.5**	y_2
⋮	⋮	⋮	⋮	⋮	⋮

Figure 9.13 *Hypothetical example illustrating the problems with regressions that control on a continuous intermediate outcome. If we control for z when regressing y on T, we will be essentially making comparisons between units such as 1 and 2 above, which differ in T but are identical in z. The trouble is that such units are not, in fact, comparable, as can be seen by looking at the potential outcomes, z^0 and z^1 (which can never both be observed, but which we can imagine for the purposes of understanding this comparison). Unit 1, which received the control, has higher potential outcomes than unit 2, which received the treatment. Matching on the observed z inherently leads to misleading comparisons as measured by the potential outcomes, which are the more fundamental quantity.*
The coefficient θ in regression (9.6) thus in general represents an inappropriate comparison of units that fundamentally differ. See Figure 9.14 for a similar example with a discrete intermediate outcome.

services including high-quality child care and home visits by trained professionals. We label y as the child's IQ score, z as the parenting quality, T as the *randomly assigned* binary treatment, and x as a pre-treatment background variable (which could in general be a vector). The goal here is to measure the effect of T on y, and we shall explain why it is not a good idea to control for the intermediate outcome, z, in making this estimate.

To keep things clean, we shall assume a linear regression for the intermediate outcome:

$$z = 0.3 + 0.2T + \gamma x + \text{error}, \tag{9.4}$$

with independent errors.[10] We further suppose that the pre-treatment variable x has been standardized to have mean 0. Then, on average, we would see parenting quality at 0.3 for the controls and 0.5 for the treated parents. Thus the causal effect of the treatment on parenting quality is 0.2. An interaction of T and x could be easily added and interpreted as well if it is desired to estimate systematic variation of treatment effects.

Similarly, a model for y given T and x—excluding z—is straightforward, with the coefficient of T representing the total effect of the treatment on the child's cognitive outcome:

$$\text{regression estimating the treatment effect:} \quad y = \theta T + \beta x + \epsilon. \tag{9.5}$$

The difficulty comes if z is added to this model. Adding z as a predictor could improve the model fit, explaining much of the variation in y:

$$\text{regression including intermediate outcome:} \quad y = \theta^* T + \beta^* x + \delta^* z + \epsilon^*. \tag{9.6}$$

We add the asterisks here because adding a new predictor changes the interpretation of each of the parameters. Unfortunately, the new coefficient θ^* does *not*, in general, estimate the effect of T.

Figure 9.13 illustrates the problem with controlling for an intermediate outcome.

[10] We use the notation γ for the coefficient of x because we are saving β for the regression of y; see model (9.5).

The coefficient of T in regression (9.6) corresponds to a comparison of units that are identical in x and z but differ in T. The trouble is, they will then automatically differ in their *potential outcomes*, z^0 and z^1. For example, consider two families, one with $z = 0.5$ but one with $T = 0$ and one with $T = 1$. Under the (simplifying) assumption that the effect of T is to increase z by exactly 0.2 (recall the assumed model (9.4)), the first family has potential outcomes $z^0 = 0.5, z^1 = 0.7$, and the second family has potential outcomes $z^0 = 0.3, z^1 = 0.5$. Thus, given two families with the same intermediate outcome z, the one that received the treatment has lower underlying parenting skills. Thus, in the regression of y on (x, T, z), the coefficient of T represents a comparison of families that differ in their underlying characteristics. This is an inevitable consequence of controlling for an intermediate outcome.

This reasoning suggests a strategy of estimating treatment effects conditional on the potential outcomes—in this example, including both z^0 and z^1, along with T and x, in the regression. The practical difficulty here (as usual) is that we observe at most one potential outcome for each observation, and thus such a regression would require imputation of z^0 or z^1 for each case (perhaps, informally, by using pre-treatment variables as proxies for z^0 and z^1), and correspondingly strong assumptions.

9.8 Intermediate outcomes and causal paths

Randomized experimentation is often described as a "black box" approach to causal inference. We see what goes into the box (treatments) and we see what comes out (outcomes), and we can make inferences about the relation between these inputs and outputs, without the ability to see what happens *inside* the box. This section discusses what happens when we use standard techniques to try to ascertain the role of post-treatment, or *mediating* variables, in the causal path between treatment and outcomes. We present this material at the end of this chapter because the discussion relies on concepts from the analysis of both randomized experiments and observational studies.

Hypothetical example of a binary intermediate outcome

Continuing the hypothetical experiment on child care, suppose that the randomly assigned treatment increases children's IQ points after three years by an average of 10 points (compared to the outcome under usual care). We would additionally like to know to what extent these positive results were the result of improved parenting practices. This question is sometimes phrased as: "What is the 'direct' effect of the treatment, net the effect of parenting?" Does the experiment allow us to evaluate this question? The short answer is no. At least not without making further assumptions.

Yet it would not be unusual to see such a question addressed by simply running a regression of the outcome on the randomized treatment variable along with a predictor representing (post-treatment) "parenting" added to the equation; recall that this is often called a *mediating* variable or mediator. Implicitly, the coefficient on the treatment variable then creates a comparison between those randomly assigned to treatment and control, within subgroups defined by post-treatment parenting practices. Let us consider what is estimated by such a regression.

For simplicity, assume these parenting practices are measured by a simple categorization as "good" or "poor." The simple comparison of the two groups can mislead, because parents who demonstrate good practices after the treatment is applied are likely to be different, on average, from the parents who would have been classified

Parenting potential	Parenting quality after assigned to control	Parenting quality after assigned to treat	Child's IQ score after assigned to control	Child's IQ score after assigned to treat	Proportion of sample
Poor parenting either way	Poor	Poor	60	70	0.1
Good parenting if treated	Poor	Good	65	80	0.7
Good parenting either way	Good	Good	90	100	0.2

Figure 9.14 *Hypothetical example illustrating the problems with regressions that control on intermediate outcomes. The table shows, for three categories of parents, their potential parenting behaviors and the potential outcomes for their children under the control and treatment conditions. The proportion of the sample falling into each category is also provided. In actual data, we would not know which category was appropriate for each individual parent—it is the fundamental problem of causal inference that we can observe at most one treatment condition for each person—but this theoretical setup is helpful for understanding the properties of statistical estimates. See Figure 9.13 for a similar example with a continuous intermediate outcome.*

as having good parenting practices even in the absence of the treatment. Therefore such comparisons, in essence, lose the advantages originally imparted by the randomization and it becomes unclear what such estimates represent.

Regression controlling for intermediate outcomes cannot, in general, estimate "mediating" effects

Some researchers who perform these analyses will claim that these models are still useful because, if the estimate of the coefficient on the treatment variable goes to zero after including the mediating variable, then we have learned that the entire effect of the treatment acts through the mediating variable. Similarly, if the treatment effect is cut in half, they might claim that half of the effect of the treatment acts through better parenting practices or, equivalently, that the effect of treatment net the effect of parenting is half the total value. This sort of conclusion is *not* generally appropriate, however, as we illustrate with a hypothetical example.

Hypothetical scenario with direct and indirect effects. Figure 9.14 displays potential outcomes of the children of the three different kinds of parents in our sample: those who will demonstrate poor parenting practices with or without the intervention, those whose parenting will get better if they receive the intervention, and those who will exhibit good parenting practices with or without the intervention. We can think of these categories as reflecting parenting *potential*. For simplicity, we have defined the model deterministically, with no individual variation within the three categories of family.

Here the effect of the intervention is 10 IQ points on children whose parents' parenting practices were unaffected by the treatment. For those parents who would improve their parenting due to the intervention, the children get a 15-point improvement. In some sense, philosophically, it is difficult (some would say impossible) to even define questions such as "what percentage of the treatment effect can be attributed to improved parenting practices" since treatment effects (and fractions attributable to various causes) can differ across people. How can we ever say for those families that have good parenting, if treated, what portion of their treatment effect can be attributed to differences in parenting practices as compared to the effects experienced by the families whose parenting practices would not change based on their treatment assignment? If we assume, however, that the effect on children

due to sources other than parenting practices stays constant over different types of people (10 points), then we might say that, at least for those with the potential to have their parenting improved by the intervention, this improved parenting accounts for about $(15 - 10)/15 = 1/3$ of the effect.

A regression controlling for the intermediate outcome does not generally work. However, if one were to try to estimate this effect using a regression of the outcome on the randomized treatment variable and observed parenting behavior, the coefficient on the treatment indicator will be -1.5, falsely implying that the treatment has some sort of negative "direct effect" on IQ scores!

To see what is happening here, recall that this coefficient is based on comparisons of treated and control groups *within* groups defined by *observed* parenting behavior. Consider, for instance, the comparison between treated and control groups within those observed to have poor parenting behavior. The group of parents who did not receive the treatment and are observed to have poor parenting behavior is a mixture of those who would have exhibited poor parenting either way and those who exhibited poor parenting simply because they did not get the treatment. Those in the treatment group who exhibited poor parenting are all those who would have exhibited poor parenting either way. Those whose poor parenting is not changed by the intervention have children with lower test scores on average—under either treatment condition—than those whose parenting would have been affected by the intervention.

The regression controlling for the intermediate outcome thus implicitly compares unlike groups of people and underestimates the treatment effect, because the treatment group in this comparison is made up of lower-performing children, on average. A similar phenomenon occurs when we make comparisons across treatment groups among those who exhibit good parenting. Those in the treatment group who demonstrate good parenting are a mixture of two groups (good parenting if treated and good parenting either way) whereas the control group is simply made up of the parents with the highest-performing children (good parenting either way). This estimate does not reflect the effect of the intervention net the effect of parenting. It does not estimate any causal effect. It is simply a mixture of some nonexperimental comparisons.

This example is an oversimplification, but the basic principles hold in more complicated settings. In short, randomization allows us to calculate causal effects of the variable randomized, but not other variables unless a whole new set of assumptions is made. Moreover, the benefits of the randomization for treatment effect estimation are generally destroyed by including post-treatment variables. These assumptions and the strategies that allow us to estimate the effects conditional on intermediate outcomes in certain situations will be discussed at the end of Chapter 10.

What can be estimated: principal stratification

We noted earlier that questions such as "What proportion of the treatment effect works through variable A?" are in some sense, inherently unanswerable. What can we learn about the role of intermediate outcomes or mediating variables? As we discussed in the context of Figure 9.14, treatment effects can vary depending on the extent to which the mediating variable (in this example, parenting practices) is affected by the treatment. The key theoretical step here is to divide the population into categories based on their potential outcomes for the mediating variable—what would happen under each of the two treatment conditions. In statistical parlance, these categorizations are sometimes called *principal strata*. The problem is that

the principal stratum labels are generally unobserved. It is theoretically possible to statistically infer principal-stratum categories based on covariates, especially if the treatment was randomized—because then at least we know that the distribution of principal strata is the same across the randomized groups. In practice, however, this reduces to making the same kinds of assumptions as are made in typical observational studies when ignorability is assumed.

Principal strata are important because they can define, even if only theoretically, the categories of people for whom the treatment effect can be estimated from available data. For example, if treatment effects were nonzero only for the study participants whose parenting practices had been changed, and if we could reasonably exclude other causal pathways, even stronger conclusions could be drawn regarding the role of this mediating variable. We discuss this scenario of *instrumental variables* in greater detail in Section 10.5.

Intermediate outcomes in the context of observational studies

If trying to control directly for mediating variables is problematic in the context of randomized experiments, it should come as no surprise that it generally is also problematic for observational studies. The concern is nonignorability—systematic differences between groups defined conditional on the post-treatment intermediate outcome. In the example above if we could control for the true parenting potential designations, the regression would yield the correct estimate for the treatment effect if we are willing to assume constant effects across groups (or willing to posit a model for how effects change across groups). One conceivably can obtain the same result by controlling sufficiently for covariates that adequately proxy this information.

In observational studies, researchers often already know to control for many predictors. So it is possible that these predictors will mitigate some of the problems we have discussed. On the other hand, studying intermediate outcomes in an observational study involves two ignorability problems to deal with rather than just one, making it all the more challenging to obtain trustworthy results.

Well-switching example. As an example where the issues discussed in this and the previous section come into play, consider one of the logistic regressions from Chapter 5:

$$\Pr(\text{switch}) = \text{logit}^{-1}(-0.21 - 0.90 \cdot \text{dist100} + 0.47 \cdot \text{arsenic} + 0.17 \cdot \text{educ4}),$$

predicting the probability that a household switches drinking-water wells as a function of distance to the nearest safe well, arsenic level of the current well, and education of head of household.

This model can simply be considered as data description, but it is natural to try to interpret it causally: being further from a safe well makes one less likely to switch, having a higher arsenic level makes switching more likely, and having more education makes one more likely to switch. Each of these coefficients is interpreted with the other two inputs held constant—and this is what we want to do, in isolating the "effects" (as crudely interpreted) of each variable. For example, households that are farther from safe wells turn out to be more likely to have high arsenic levels, and in studying the "effect" of distance, we would indeed like to compare households that are otherwise similar, including in their arsenic level. This fits with a psychological or decision-theoretic model in which these variables affect the perceived costs and benefits of the switching decision (as outlined in Section 6.8).

However, in the well-switching example as in many regression problems, additional assumptions beyond the data are required to justify the convenient interpre-

tation of multiple regression coefficients as causal effects—what would happen to y if a particular input were changed, with all others held constant—and it is rarely appropriate to give more than one coefficient such an interpretation, and then only after careful consideration of ignorability. Similarly, we cannot learn about causal pathways from observational data without strong assumptions.

For example, a careful estimate of the effect of a potential intervention (for example, digging new, safe wells in close proximity to existing high-arsenic households) should include, if not an actual experiment, a model of what would happen in the particular households being affected, which returns us to the principles of observational studies discussed earlier in this chapter.

9.9 Bibliographic note

The fundamental problem of causal inference and the potential outcome notation were introduced by Rubin (1974, 1978). Related earlier work includes Neyman (1923) and Cox (1958). For other approaches to causal inference, see Pearl (2000) along with many of the references in Section 10.8.

The stable unit treatment value assumption was defined by Rubin (1978); see also Sobel (2006) for a more recent discussion in the context of a public policy intervention and evaluation. Ainsley, Dyke, and Jenkyn (1995) and Besag and Higdon (1999) discuss spatial models for interference between units in agricultural experiments. Gelman (2004d) discusses treatment interactions in before/after studies.

Campbell and Stanley (1963) is an early presentation of causal inference in experiments and observational studies from a social science perspective; see also Achen (1986) and Shadish, Cook, and Campbell (2002). Rosenbaum (2002b) and Imbens (2004) present overviews of inference for observational studies. Dawid (2000) offers another perspective on the potential-outcome framework. Leamer (1978, 1983) explores the challenges of relying on regression models for answering causal questions.

Modeling strategies also exist that rely on ignorability but loosen the relatively strict functional form imposed by linear regression. Examples include Hahn (1998), Heckman, Ichimura and Todd (1998), Hirano, Imbens, and Ridder (2003), and Hill and McCulloch (2006).

The example regarding the Korean babies up for adoption was inspired by Sacerdote (2004). The Electric Company experiment is described by Ball and Bogatz (1972) and Ball et al. (1972).

Rosenbaum (1984) provides a good discussion of the dangers outlined in Section 9.8 involved in trying to control for post-treatment outcomes. Raudenbush and Sampson (1999), Rubin (2000), and Rubin (2004) discuss direct and indirect effects for multilevel designs. We do not attempt here to review the vast literature on structural equation modeling; Kenny, Kashy, and Bolger (1998) is a good place to start.

The term "principal stratification" was introduced by Frangakis and Rubin (2002); examples of its application include Frangakis et al. (2003) and Barnard et al. (2003). Similar ideas appear in Robins (1989, 1994).

9.10 Exercises

1. Suppose you are interested in the effect of the presence of vending machines in schools on childhood obesity. What randomized experiment would you want to do (in a perfect world) to evaluate this question?
2. Suppose you are interested in the effect of smoking on lung cancer. What ran-

domized experiment could you plausibly perform (in the real world) to evaluate this effect?

3. Suppose you are a consultant for a researcher who is interested in investigating the effects of teacher quality on student test scores. Use the strategy of mapping this question to a randomized experiment to help define the question more clearly. Write a memo to the researcher asking for needed clarifications to this study proposal.
4. The table below describes a hypothetical experiment on 2400 persons. Each row of the table specifies a category of person, as defined by his or her pre-treatment predictor x, treatment indicator T, and potential outcomes y^0, y^1. (For simplicity, we assume unrealistically that all the people in this experiment fit into these eight categories.)

Category	# persons in category	x	T	y^0	y^1
1	300	0	0	4	6
2	300	1	0	4	6
3	500	0	1	4	6
4	500	1	1	4	6
5	200	0	0	10	12
6	200	1	0	10	12
7	200	0	1	10	12
8	200	1	1	10	12

   In making the table we are assuming omniscience, so that we know both y^0 and y^1 for all observations. But the (nonomniscient) investigator would only observe x, T, and y^T for each unit. (For example, a person in category 1 would have $x=0, T=0, y=4$, and a person in category 3 would have $x=0, T=1, y=6$.)

   (a) What is the average treatment effect in this population of 2400 persons?

   (b) Is it plausible to believe that these data came from a randomized experiment? Defend your answer.

   (c) Another population quantity is the mean of y for those who received the treatment minus the mean of y for those who did not. What is the relation between this quantity and the average treatment effect?

   (d) For these data, is it plausible to believe that treatment assignment is ignorable given sex? Defend your answer.

5. For the hypothetical study in the previous exercise, figure out the estimate and the standard error of the coefficient of T in a regression of y on T and x.
6. You are consulting for a researcher who has performed a randomized trial where the treatment was a series of 26 weekly therapy sessions, the control was no therapy, and the outcome was self-report of emotional state one year later. However, most people in the treatment group did not attend every therapy session. In fact there was a good deal of variation in the number of therapy sessions actually attended. The researcher is concerned that her results represent "watered down" estimates because of this variation and suggests adding in another predictor to the model: number of therapy sessions attended. What would you advise her?
7. Gain-score models: in the discussion of gain-score models in Section 9.3, we noted that if we include the pre-treatment measure of the outcome in a gain score model, the coefficient on the treatment indicator will be the same as if we had just run a standard regression of the outcome on the treatment indicator and the pre-treatment measure. Show why this is true.

8. Assume that linear regression is appropriate for the regression of an outcome, y, on treatment indicator, T, and a single confounding covariate, x. Sketch hypothetical data (plotting y versus x, with treated and control units indicated by circles and dots, respectively) and regression lines (for treatment and control group) that represent each of the following situations:

   (a) No treatment effect,
   (b) Constant treatment effect,
   (c) Treatment effect increasing with x.

9. Consider a study with an outcome, y, a treatment indicator, T, and a single confounding covariate, x. Draw a scatterplot of treatment and control observations that demonstrates each of the following:

   (a) A scenario where the difference in means estimate would not capture the true treatment effect but a regression of y on x and T would yield the correct estimate.
   (b) A scenario where a linear regression would yield the wrong estimate but a nonlinear regression would yield the correct estimate.

10. The folder `sesame` contains data from an experiment in which a randomly selected group of children was encouraged to watch the television program Sesame Street and the randomly selected control group was not.

    (a) The goal of the experiment was to estimate the effect on child cognitive development of watching more Sesame Street. In the experiment, encouragement but not actual watching was randomized. Briefly explain why you think this was done. (Hint: think of practical as well as statistical reasons.)
    (b) Suppose that the investigators instead had decided to test the effectiveness of the program simply by examining how test scores changed from before the intervention to after. What assumption would be required for this to be an appropriate causal inference? Use data on just the control group from this study to examine how realistic this assumption would have been.

11. Return to the Sesame Street example from the previous exercise.

    (a) Did encouragement (the variable `viewenc` in the dataset) lead to an increase in post-test scores for letters (`postlet`) and numbers (`postnumb`)? Fit an appropriate model to answer this question.
    (b) We are actually more interested in the effect of watching Sesame Street regularly (`regular`) than in the effect of being encouraged to watch Sesame Street. Fit an appropriate model to answer this question.
    (c) Comment on which of the two previous estimates can plausibly be interpreted causally.

12. Messy randomization: the folder `cows` contains data from an agricultural experiment that was conducted on 50 cows to estimate the effect of a feed additive on six outcomes related to the amount of milk fat produced by each cow.

    Four diets (treatments) were considered, corresponding to different levels of the additive, and three variables were recorded before treatment assignment: lactation number (seasons of lactation), age, and initial weight of cow.

    Cows were initially assigned to treatments completely at random, and then the distributions of the three covariates were checked for balance across the treatment groups; several randomizations were tried, and the one that produced the

"best" balance with respect to the three covariates was chosen. The treatment assignment is ignorable (because it depends only on fully observed covariates and not on unrecorded variables such as the physical appearances of the cows or the times at which the cows entered the study) but unknown (because the decisions whether to rerandomize are not explained).

We shall consider different estimates of the effect of additive on the mean daily milk fat produced.

(a) Consider the simple regression of mean daily milk fat on the level of additive. Compute the estimated treatment effect and standard error, and explain why this is not a completely appropriate analysis given the randomization used.

(b) Add more predictors to the model. Explain your choice of which variables to include. Compare your estimated treatment effect to the result from (a).

(c) Repeat (b), this time considering additive level as a categorical predictor with four letters. Make a plot showing the estimate (and standard error) of the treatment effect at each level, and also showing the inference the model fit in part (b).

13. The folder `congress` has election outcomes and incumbency for U.S. congressional election races in the 1900s.

(a) Take data from a particular year, t, and estimate the effect of incumbency by fitting a regression of $v_{i,t}$, the Democratic share of the two-party vote in district i, on $v_{i,t-2}$ (the outcome in the previous election, two years earlier), I_{it} (the incumbency status in district i in election t, coded as 1 for Democratic incumbents, 0 for open seats, -1 for Republican incumbents), and P_{it} (the incumbent *party*, coded as 1 if the sitting congressmember is a Democrat and -1 if he or she is a Republican). In your analysis, include only the districts where the congressional election was contested in both years, and do not pick a year ending in "2." (District lines in the United States are redrawn every ten years, and district election outcomes v_{it} and $v_{i,t-2}$ are not comparable across redistrictings, for example, from 1970 to 1972.)

(b) Plot the fitted model and the data, and discuss the political interpretation of the estimated coefficients.

(c) What assumptions are needed for this regression to give a valid estimate of the causal effect of incumbency? In answering this question, define clearly what is meant by incumbency as a "treatment variable."

See Erikson (1971), Gelman and King (1990), Cox and Katz (1996), Levitt and Wolfram (1997), Ansolabehere, Snyder, and Stewart (2000), Ansolabehere and Snyder (2002), and Gelman and Huang (2006) for further work and references on this topic.

14. Causal inference based on data from individual choices: our lives involve trade-offs between monetary cost and physical risk, in decisions ranging from how large a car to drive, to choices of health care, to purchases of safety equipment. Economists have estimated people's implicit balancing of dollars and danger by comparing different jobs that are comparable but with different risks, fitting regression models predicting salary given the probability of death on the job. The idea is that a riskier job should be compensated with a higher salary, with the slope of the regression line corresponding to the "value of a statistical life."

(a) Set up this problem as an individual choice model, as in Section 6.8. What are an individual's options, value function, and parameters?

(b) Discuss the assumptions involved in assigning a causal interpretation to these regression models.

See Dorman and Hagstrom (1998), Costa and Kahn (2002), and Viscusi and Aldy (2002) for different perspectives of economists on assessing the value of a life, and Lin et al. (1999) for a discussion in the context of the risks from radon exposure.