Chapter 17

Regression in the context of randomized experiments

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. Causal interpretations of regression coefficients can only be justified by relying on much stricter assumptions than are needed for predictive inference. Fortunately data arising from randomized experiments are an ideal setting for using regression to estimate a treatment effect because the design of the experiment guarantees that treatment assignment is independent of the potential outcomes. We illustrate use of regression in the setting of randomized experiments in this chapter.

17.1 The data: pre-treatment covariates, treatments, and potential outcomes

In this chapter we consider a scenario in which there is a sample of items or *study units*, each of which has been randomly assigned to receive either Treatment 1 or Treatment 0. Often these are labeled as the *treatment group* and the *control group* although alternately these can represent two distinct treatments. As shown in Figure 17.1, there is typically (but not necessarily) some background information on each observation, then there is the treatment assignment, and then, after some period of time (allowing for the treatment to have its effect) an outcome is measured. We thus can have at least three sorts of measurements on each item i:

- The pre-treatment measurements, also called *covariates*, x_i (which, as noted above, are not strictly required for causal inference),
- The treatment z_i, which equals 1 for treated units and 0 for controls,
- The outcome measurement, y_i, which we label as y_i^1 when units have been exposed to the treatment and y_i^0 when units have not been exposed to the treatment. As discussed in detail in Section 16.1, the notation y_i^0, y_i^1 refers to *potential outcomes* under different possible treatment assignments, and y_i refers to which of these is actually observed.

Here are some examples in this framework:

- Units are high school students, x is college admissions test score on the first try, $T = 1$ for students who are assigned to get coaching or 0 otherwise, y is test score on the second try, six months later.
- Units are male or female musicians auditioning for jobs at leading orchestras, x is their sex (for example, $x = 1$ for men and 0 for women), $z = 1$ if the audition is performed behind a screen or 0 if the evaluators can see the musician audition, and y is the outcome (defined as 1 if the applicant gets the job offer and 0 otherwise).

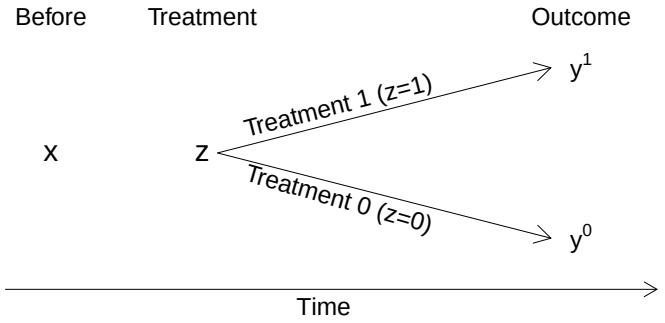

Figure 17.1 *Basic framework for causal inference. The treatment effect is* $y^1 - y^0$*, but we can never observe both* potential outcomes y^1 *and* y^0 *on the same item. For any given unit, the unobserved outcome is called the* counterfactual.

More generally, there can be multiple pre-treatment measurements, multiple treatment levels, and multiple outcome measurements.

17.2 Example: the effect of showing children an educational television show

We illustrate the use of regression to estimate a causal effect in the context of an educational experiment that was undertaken around 1970 on a set of 192 elementary school classes.[1] The goal was to measure the effect of a new educational television program, "The Electric Company," on children's reading ability. We discussed this study briefly on page 6 in Section 1.2. Selected classes of children in grades 1–4 were randomized into treated and control groups. At the beginning and 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; that is, we treat the classes as the observational units in this study.

Displaying the data two different ways

Figure 17.2 displays the distribution of the outcome, average post-treatment test scores, in the control and treatment group for each grade. Recall that 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. Rather than try to cleverly put all the data on a single plot, we arrange them on a 4×2 grid, using a common scale for all the graphs to facilitate comparisons among grades and between treatment and control. We also extend the axis all the way to zero, which is not strictly necessary, in the interest of clarity of presentation. In this example, as in many others, we are not concerned with the exact counts in the histogram; thus, we simplify the display by eliminating y-axes, and we similarly clarify the x-axis by removing tick marks and using minimal labeling.

As discussed in Section 2.3, all graphs can be considered as comparisons. Figure 17.2 most directly allows comparisons between grades. This comparison is useful—if for no other reason than to ground ourselves and confirm that scores are higher in the higher grades—but we are more interested in the comparison of treatment to control within each grade.

Thus, it might be more helpful to arrange the histograms as shown in Figure 17.3, with treatment and control aligned for each grade. With four histograms arranged horizontally on a page, we need to save some space and so we restrict the x-axes to the combined range

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

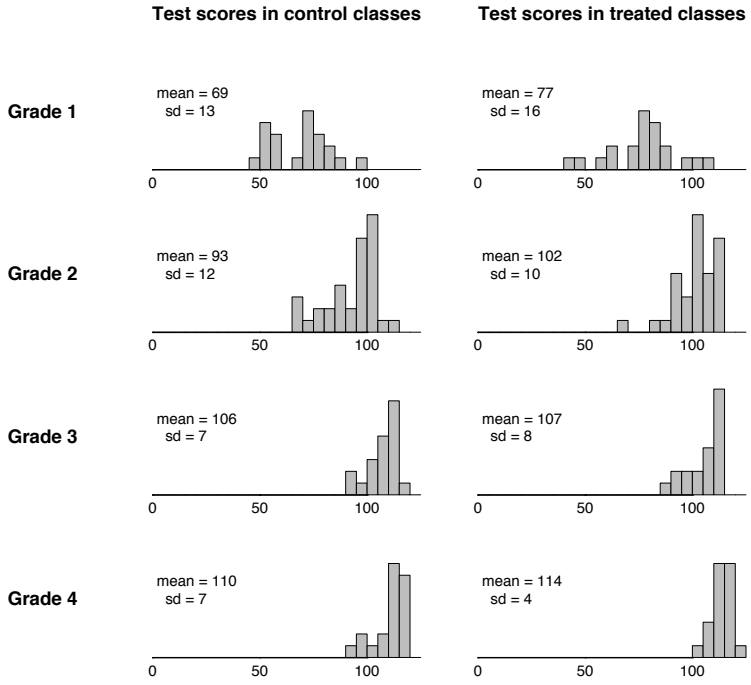

Figure 17.2 *Post-treatment test scores from an experiment measuring the effect of an educational television program, the Electric Company, on children's reading abilities. The experiment was applied on a total of 192 classrooms in four grades. At the end of the experiment, the average reading test score in each classroom was recorded.*

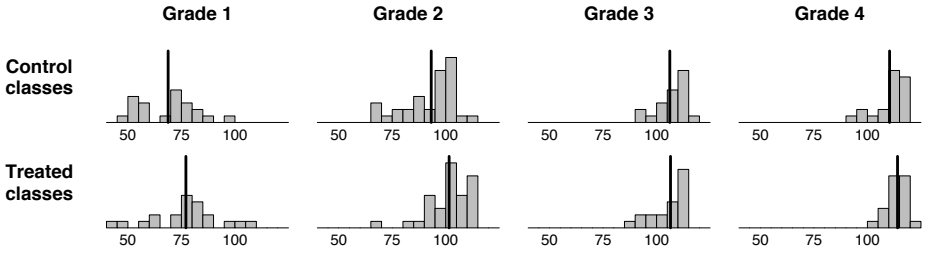

Figure 17.3 *Data from the Electric Company experiment, from Figure 17.2 on page 333, displayed in a different orientation to allow easier comparison between treated and control groups in each grade. For each histogram, the average is indicated by a vertical line.*

of the data. We also indicate the average value in each group with a vertical line to allow easier comparisons of control to treatment in each grade.

Visual comparisons across treatment groups within grades suggest that watching the Electric Company may have led to small increases in average test scores, particularly for the lower grades. This plot also reveals an apparent ceiling with respect to the reading assessment used. There is not much room for improvement for those classes that scored well in the later grades. This feature of the assessment makes it difficult to know whether the relatively small estimated effects at the higher grades might have been larger if the

assessments had included more difficult items to allow for more room for improvement among the more advanced classes.

Paired comparisons design

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 an example of a *matched pairs* design (which in turn is a special case of a *randomized block* design, with exactly two units within each block). Recall that the idea behind this is that there are characteristics of the school—both observable and, potentially, unobservable—that are predictive of future student outcomes that we would like to adjust for explicitly by forcing balance through our design.

For simplicity we shall analyze this experiment *as if the treatment assignment had been completely randomized within each grade*. In the companion volume we return to the example and show how to use a multilevel model to account for the pairing in the design.

Simple difference estimate, appropriate for a completely randomized experiment with no pre-treatment variables

We start by estimating a single treatment effect using the simplest possible estimate, a regression of post-test on treatment indicator, which would be the appropriate analysis had the data come from a completely randomized experiment with no available pre-treatment information.

When treatments are assigned completely at random, we can think of the treatment and control groups as two 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, $y_i = \alpha + \theta z_i + \text{error}_i$. In R:

R code
```
display(lm(post_test ~ treatment, data=electric))
```

Applied to the Electric Company data, this yields an estimate of 5.7 with a standard error of 2.5. This estimate is a starting point, but we should be able to do better, as it makes sense that the effects of the television show could vary by grade, and indeed this appears to show up from visual comparisons of the histograms in Figure 17.2.

Separate analysis within each grade

Given the large variation in test scores between grade to grade, it makes sense to take the next step and perform a separate regression analysis on each grade's data. This is equivalent to fitting a model in which treatment effects vary by grade—that is, an interaction between treatment and grade indicators—and where the residual variance can be different from grade to grade as well.

In R we fit this as four separate models:

R code
```
for (k in 1:4) {
  display(lm(post_test ~ treatment, data=electric, subset=(grade==k))
}
```

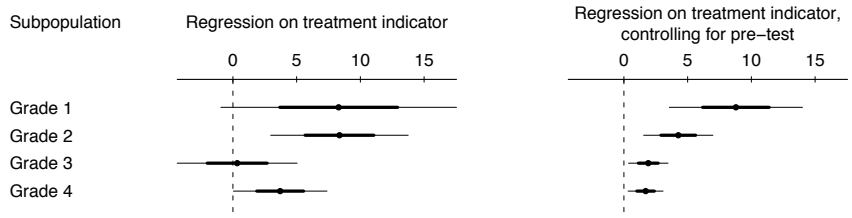

Figure 17.4 *Estimates, 50%, and 95% intervals for the effect of the Electric Company television show (see data in Figures 17.2 and 17.5) as estimated in two ways: first, from a regression on treatment alone, and second, also adjusting for pre-test data. In both cases, the coefficient for treatment is 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 (adjusting for pre-test or not). For instance, the plot highlights how adjusting 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 17.5.

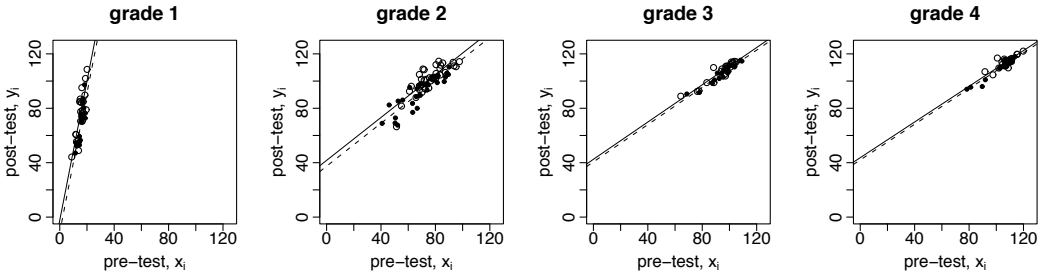

Figure 17.5 *Pre-test/post-test data for the Electric Company experiment. Treated and control classes are indicated by circles and dots, respectively, and the solid and dotted lines represent parallel regression lines fit to the treatment and control groups, respectively. The solid lines are slightly higher than the dotted lines, indicating slightly positive estimated treatment effects. Compare to Figure 17.2, which displays only the post-test data.*

The resulting estimates and uncertainty intervals for the Electric Company experiment are graphed in the left panel of Figure 17.4. 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.

Sample sizes are approximately the same in each of the grades, but the estimates for higher grades have lower standard errors because the residual standard deviations of the regressions are lower in these grades (see Figure 17.5).

17.3 Including pre-treatment predictors

Adjusting for pre-test to get more precise estimates

In the Electric Company experiment, a pre-test was given in each class at the beginning of the school year, before the treatment was applied, and we can use this information to improve our treatment effect estimates, using a regression model such as, $y_i = \alpha + \theta z_i + \beta x_i + \text{error}_i$, where z_i indicates the treatment (1 if Electric Company and 0 if control) and x_i denotes the average pre-test scores of the students in classroom i. As before, we perform this estimate separately for each grade:

R code
```
for (k in 1:4) {
  display(lm(post_test ~ treatment + pre_test, data=electric, subset=(grade==k))
}
```

Figure 17.5 shows the before-after data for the Electric Company experiment. For each grade, the difference between the regression lines for the two groups represents the estimated treatment effect as a function of pre-test score. Since we have not included any interaction in the model, the treatment effect within each grade 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 in Figure 17.2). 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 the regression,

$$y_i = \alpha + \theta z_i + \beta x_i + \text{error}_i, \tag{17.1}$$

the coefficient θ still represents the average treatment effect in the grade, but adjusting for pre-test, x_i, can improve the efficiency of the estimate. More generally, the regression can adjust for multiple pre-treatment predictors, in which case the model has the form $y_i = \alpha + \tau z_i + x_i\beta + \text{error}_i$, where x_i and β are now vectors of predictors and coefficients, respectively; alternatively α could 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 17.4. It now appears that the treatment is effective on average for each of the grades, although, consistent with our earlier visual inspection of the data, the effects seem larger in the lower grades.

Crucially, it is only appropriate to adjust for *pre*-treatment predictors, or, more generally, predictors that would not be affected by the treatment (such as race or age). We discuss this point in more detail in Section 17.5.

Problems with simple before-after comparisons

Given that we have pre-test and post-test measurements, why not simply summarize the treatment effect by their difference? Why bother with a controlled experiment at all? The problem with the simple before-after estimate is that, when estimating causal effects we are interested in the difference between treatment and control conditions, not in the simple improvement from pre-test to post-test. The improvement is not a causal effect (except under the assumption, unreasonable in this case, that under the control there would be no change in reading ability during the school year).

There are real-world settings in which there is no control, and then strong assumptions do need to be made to draw any causal inferences. But when a control group is available, it should be used. In a regression context, the treatment effect is the coefficient on the treatment indicator, and if there is a treatment group but no control group, then the treatment variable $z_i = 1$ for all units, and there is no way to compute the regression coefficient of z without strong modeling assumptions that cannot be checked with the data at hand.

Gain scores

More generally, 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), $g_i = \alpha + \tau z_i + \text{error}_i$. In the simple case with no other predictors, the regression estimate is simply $\hat{\tau} = \bar{g}^T - \bar{g}^C$, the average difference of gain scores in the treatment and control groups.

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. Also, as with other ways of adjusting for predictors, it is never appropriate to create a gain score by subtracting a variable that was measured after the treatment assignment.

One perspective on the analysis of gain scores is that it implicitly makes an unnecessary assumption, namely, that $\beta = 1$ in model (17.1). To see this, note the algebraic equivalence between $y_i = \alpha + \tau z_i + x_i + \text{error}_i$ and $y_i - x_i = \alpha + \tau z_i + \text{error}_i$. On the other hand, if this assumption is close to being true then τ 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, $g_i = \alpha + \tau z_i + \gamma x_i + \text{error}_i$. However, in this case, $\hat{\tau}$, the estimate of the coefficient for z, is equivalent to the estimated coefficient from the original model, $y_i = \alpha + \tau z_i + \beta x_i + \text{error}_i$ (see Exercise 17.3).

Sometimes, however, gain score models are motivated by concern that the pre-treatment score may have been measured with error. In this case we would prefer to adjust for the latent "true score" rather than the observed pre-treatment measurement. Setting up a regression adjusting for an unmeasured variable is more complicated, and we discuss this sort of measurement-error model in our companion volume on multilevel models. For here, we merely note that there are settings where performing a simple regression using the gain score can be a useful approximation to that latent-variable regression. In a randomized experiment such as the Electric Company study this is less of a concern because the treatment assignment is independent of all pre-treatment variables that have not been already accounted for in the design.

Another motivation for use of gain scores is the desire to interpret effects on changes in the outcome rather than the effect on the outcome on its own. Compare this interpretation to the interpretation of a treatment effect estimate from a model that controls for the pre-test; in this case we could interpret an effect on the outcome for those with the same value of the pre-test. The difference between these interpretations is subtle.

17.4 Varying treatment effects and interactions

Once we include pre-test in the model, it is natural to interact it with the treatment effect. The treatment effect is then allowed to vary with the level of the pre-test. Figure 17.6 shows the Electric Company data with separate regression lines estimated for the treatment and control groups. As with Figure 17.5, 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:

R output

```
lm(formula = post_test ~ treatment, data=electric, subset=(grade==4))
            coef.est coef.se
(Intercept)    110.4     1.3
treatment        3.7     1.8
  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 estimate by adjusting for the pre-test score:

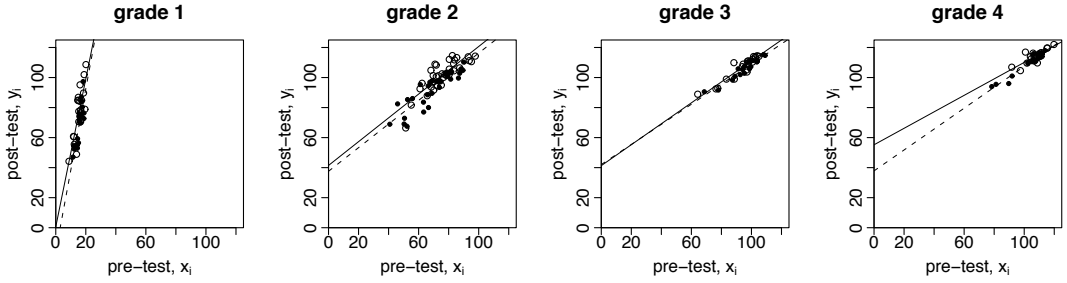

Figure 17.6 *Pre-test/post-test data for the Electric Company experiment. Treated and control classes are indicated by circles and dots, respectively, and the solid and dotted lines represent separate regression lines fit to the treatment and control groups, respectively—that is, the model interacts treatment with pre-test score. For each grade, the difference between the solid and dotted lines represents the estimated treatment effect as a function of pre-test score. Compare to Figure 17.5, which displays the same data but with parallel regression lines in each graph. The non-parallel lines in the current figure represent interactions in the fitted model.*

R output
```
lm(formula = post_test ~ treatment + pre_test, data=electric, subset=(grade==4))
            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, adjusting for the pre-test reduced the estimated effect. Under a clean randomization, adjusting for pre-treatment predictors in this way should reduce the standard errors of the estimates. As discussed earlier, under a clean randomization, adjusting for pre-treatment predictors in this way does not change what we are estimating. However, if the predictor has a strong association with the outcome it can help to bring each estimate closer (on average) to the truth, and if the randomization was less than pristine, the addition of predictors to the equation may help us adjust for *systematically* unbalanced characteristics across groups. Thus, this strategy has the potential to adjust for problems with the randomization.

Figure 17.4 shows the estimates for the Electric Company experiment in all four grades. Complications arise when we include the interaction of treatment with pre-test, as we show in this analysis of the fourth grade data:

R output
```
lm(formula = post_test ~ treatment + pre_test + treatment:pre_test,
    data=electric, 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 17.6 we see that pre-test scores range from approximately 80 to 120; in this range, the estimated treatment effect varies from $17 - 0.15 * 80 = 5$ for classes with pre-test scores of 80 to $17 - 0.15 * 120 = -1$ for classes with pre-test scores of 120. This range represents the *variation* in estimated treatment effects as a function of pre-test

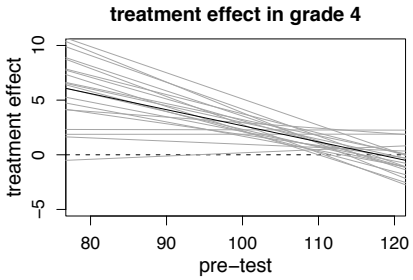

Figure 17.7 *Estimate and uncertainty for the effect of viewing the Electric Company (compared to the control treatment) for fourth-graders. Compare to the data in the rightmost plot in Figure 17.6. The dark line here—the estimated treatment effect as a function of pre-test score—is the difference between the two regression lines in the grade 4 plot in Figure 17.6. The gray lines represent 20 random draws from the uncertainty distribution of the treatment effect.*

score, *not* uncertainty in the estimated treatment effect. Centering x before including it in the model allows the treatment coefficient to represent that treatment effect for classes with the mean pre-test score for the sample.

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:

R code
```
fit_4 <- stan_glm(post_test ~ treatment + pre_test + treatment:pre_test,
  data=electric, subset=(grade==4))
sims_4 <- as.matrix(fit_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=0.5, lty=2)
for (i in 1:20){
  curve(sims_4[i,2] + sims_4[i,4]*x, lwd=0.5, col="gray", add=TRUE)}
est_4 <- apply(sims_4, 2, median)
curve(est_4[2] + est_4[4]*x, lwd=0.5, add=TRUE)
```

This produces the graph shown in Figure 17.7.

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 $y_i = \alpha + \tau_1 z_i + \beta x_i + \tau_2 z_i x_i + \text{error}_i$, then the treatment effect is $\tau_1 + \tau_2 x$, and the summary treatment effect in the sample is $\frac{1}{n}\sum_{i=1}^{n}(\tau_1 + \tau_2 x_i)$, averaging over the n fourth-grade classrooms in the data. We can compute the estimated average treatment effect as follows:

R code
```
n_sims <- nrow(sims_4)
effect <- array(NA, c(n_sims, sum(grade==4)))
for (i in 1:n_sims){
  effect[i,] <- sims_4[i,2] + sims_4[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—similar to the result from the model adjusting 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 to simply estimate an average effect.

Identification of treatment interactions is also important when we want to generalize experimental results to a broader population. If treatment effects vary with pre-treatment characteristics and the distribution of these characteristics varies between the experimental sample and the population of interest, the average treatment effects will typically be different. If these characteristics are observable, however, we should be able to extrapolate the one estimate to the other.

Using poststratification to combine conditional treatment effects to obtain an average treatment effect

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.

We discussed this in Section 15.1 in the general regression context where the challenge is to line up sample and population. Here we apply the idea to causal inference, where the challenge is to line up treatment and control groups.

For example, suppose we have treatment variable z and pre-treatment control variables x_1, x_2, and our regression predictors are x_1, x_2, z, and the interactions $x_1 z$ and $x_2 z$, so that the linear model is: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 z + \beta_4 x_1 z + \beta_5 x_2 z + \text{error}$. 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 obtained using `stan_glm()` as discussed in earlier chapters.

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.

17.5 Do not control for post-treatment variables

As illustrated in the examples of this chapter, we recommend adjusting for pre-treatment 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. By "control for" a variable, we specifically are referring to including it as a regression input. As we discuss in Section 19.1, information on post-treatment variables can be included in the more complicated framework of instrumental variables or in more general mediator strategies.

In this section we explain why naively adjusting for a post-treatment variable can bias the estimate of the treatment effect, *even when the treatment has been randomly assigned to study participants*. In the next section we describe the related 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 services including high-quality child care and home visits by trained professionals. We label y as the child's IQ score measured 2 years after the treatment regime has been completed,

q as a continuous parenting quality measure (ranging from 0 to 1) measured one year after treatment completion, z as the *randomly assigned* binary treatment, and x as the pre-treatment measure reflecting whether both parents have a high school education (in general this could be a vector of pre-treatment predictors). The goal here is to measure the effect of z on y, and we shall explain why it is not a good idea to control for the intermediate outcome, q, when estimating this effect.

Due to the randomization we know that unconditional ignorability holds, that is, $y^0, y^1 \perp z$. Thus a model for y given z alone would serve to unbiasedly estimate the treatment effect:

$$\text{regression estimating the treatment effect:} \quad y = \tau z + \epsilon.$$

The coefficient on z in this equation is equivalent to the difference in mean outcomes across treatment groups. The model has no intercept because the treatment effect is defined relative to the control group so by definition it is zero when $z = 0$.

As discussed above, we also know that when z is randomized, ignorability also holds conditional on x, that is, $y^0, y^1 \perp z \,|\, x$. Therefore a model for y given z and x—excluding q—is straightforward, with the coefficient on z representing the total effect of the treatment on the child's cognitive outcome:

$$\text{regression estimating the treatment effect:} \quad y = \tau z + \beta x + \epsilon.$$

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

$$\text{regression including intermediate outcome:} \quad y = \tau^* z + \beta^* x + \delta^* q + \epsilon^*. \tag{17.2}$$

We add the asterisks here because adding a new predictor changes the meaning of each of the parameters. Unfortunately, the new coefficient τ^* does *not*, in general, estimate the effect of z. Formally this is because it is not in general true that $y^0, y^1 \perp z \,|\, x, q$ if q was measured post-treatment, and in particular if q was affected by the treatment.

For instance, suppose that the true model for q given z and x is represented by a linear regression:

$$q = 0.3 + 0.2z + \gamma x + \text{error}, \tag{17.3}$$

with independent errors. By saying this is the true model, we are implying that $\mathrm{E}(q(0)|x) = 0.3 + \gamma x$ and $\mathrm{E}(q(1)|x) = 0.3 + \gamma x + 0.2$. We further suppose that the pre-treatment variable x has been standardized to have mean 0. We can see then that parenting quality has an average of 0.3 for the controls and 0.5 for the treated parents. Thus z increases parenting quality by 0.2 points on this 0 to 1 scale.

Figure 17.8 illustrates the problem with controlling for an intermediate variable that has been affected by the treatment. The coefficient of z in regression (17.2) corresponds to a comparison of units that are identical in x and q but differ in z. The trouble is, they will then automatically differ in their *potential outcomes*, q^0 and q^1. For example, consider two families, both with $q = 0.5$ and $x = 0$, but one with $z = 0$ and one with $z = 1$. Under the (simplifying) assumption that the effect of z is to increase q by exactly 0.2 (as in the assumed model (17.3)), the first family has potential outcomes $q^0 = 0.5, q^1 = 0.7$, and the second family has potential outcomes $q^0 = 0.3, q^1 = 0.5$. Thus, given two families with the same observed intermediate outcome q, the one that received the treatment has lower underlying parenting skills. Thus, in the regression of y on (x, z, q), the coefficient of z represents a comparison of families that differ fundamentally in their underlying characteristics. This is an inevitable consequence of controlling for an intermediate outcome that has been affected by the treatment.

This reasoning suggests a strategy of estimating treatment effects conditional on the potential outcomes—in this example, including both q^0 and q^1, along with z and x, in

unit, i	pre-treatment covariate, x_i	treatment, z_i	observed intermediate outcome, q_i	potential intermediate outcomes, q_i^0	q_i^1	final outcome, y_i
1	0	0	0.5	**0.5**	*0.7*	y_1
2	0	1	0.5	*0.3*	**0.5**	y_2
⋮	⋮	⋮	⋮	⋮	⋮	

Figure 17.8 *Hypothetical example illustrating the problems with regressions that control on a continuous intermediate outcome. If we control for q when regressing y on z, we will be essentially making comparisons between units such as 1 and 2 above, which differ in z but are identical in q. The trouble is that such units are not, in fact, comparable, as can be seen by looking at the parenting quality potential outcomes, q^0 and q^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 parenting quality potential outcomes than unit 2, which received the treatment. Matching on the observed q thus inherently leads to misleading comparisons as reflected in the parenting quality potential outcomes.*
The coefficient τ in regression (17.2) thus in general represents an inappropriate comparison of units that fundamentally differ. See Figure 17.9 for a similar example with a discrete intermediate outcome.

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 (prediction) of q^0 or q^1 for each case (perhaps, informally, by using pre-treatment variables as proxies for q^0 and q^1), and correspondingly strong assumptions.

17.6 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 (naive) techniques to try to ascertain the role of post-treatment, or *mediating* variables, in the causal path between treatment and outcomes as part of a well-intentioned attempt to peer inside the black box.

Hypothetical example of a binary intermediate outcome

Continuing the hypothetical experiment on child care, suppose that the randomly assigned treatment increases children's IQ after three years by an average of 10 points (compared to the outcome under usual care). In comparison to the previous section, now we would like to understand 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 of the effect of parenting?" Does the experiment allow us to easily evaluate this question? The short answer is no. At least not without making further assumptions.

Yet it is not unusual to see such a question addressed by simply running a regression of the outcome on the randomized treatment variable along with the (post-treatment) mediating variable "parenting" added to the equation. In this model, the coefficient on the treatment variable 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 that parenting quality is now measured by a simple categorization: "good" or "poor." The simple comparison of the two groups can be misleading, because

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 17.9 *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 17.8 for a similar example with a continuous intermediate outcome.*

parents who demonstrate good parenting practices after the treatment is applied are likely to be different, on average, from the parents who would have been classified 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 they represent. Said another way, in general this approach will lead to biased estimates of the average treatment effect.

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 is (statistically indistinguishable from) 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 generally *not* appropriate, however, as we illustrate with a hypothetical example.

Hypothetical scenario with direct and indirect effects. Figure 17.9 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. We have also ruled out the existence of parents whose parenting quality is adversely affected by the intervention.

Here the effect of the intervention is 10 IQ points for children whose parents' parenting practices were unaffected by the treatment. For those parents who would improve their parenting due to the intervention, the children would benefit from 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 would be -1.5, falsely implying that the treatment has some sort of negative "direct effect" on IQ scores!

To understand what is happening here, recall that this coefficient is based on comparisons of treated and control group 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. Estimation strategies that allow us to estimate the effects conditional on intermediate outcomes and the corresponding assumptions will be discussed at the end of Chapter 18.

17.7 Bibliographic note

The fundamental problem of causal inference and the potential outcome notation were formally introduced by Rubin (1974, 1978) who also coined the term ignorability. Related earlier work includes Neyman (1923) and Cox (1958).

The Electric Company experiment is described by Ball and Bogatz (1972) and Ball et al. (1972). 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.

Rosenbaum (1984) provides a helpful discussion of the dangers outlined in Section 17.6 involved in trying to control for post-treatment outcomes; see also Montgomery, Nyhan, and Torres (2017). 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).

17.8 Exercises

1. *External validity:* Comment on the external validity of the Electric Company example.
2. *Compliance:* 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?
3. *Gain scores:* In the discussion of gain-score models in Section 17.3, 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.
4. *Pre-test and post-test:* 100 students are given a pre-test, then a treatment or control is randomly assigned to each, then they get a post-test. Given the following regression model:

$$\text{post\_test} = a + b * \text{pre\_test} + \theta * z + \text{error},$$

   where $z = 1$ for treated units and 0 for controls. Further suppose that `pre_test` has mean 40 and standard deviation 15. Suppose $b = 0.7$ and $\theta = 10$ and the mean for `post_test` is 50 for the students in the control group. Further suppose that the residual standard deviation of the regression is 10.

   (a) Determine a.

   (b) What is the standard deviation of the post-test scores for the students in the control group?

   (c) What are the mean and standard deviation of the post-test scores in the treatment group?
5. *Sketching the regression model for causal inference:* Assume that linear regression is appropriate for the regression of an outcome, y, on treatment indicator, z, 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.
6. *Linearity assumptions and causal inference:* Consider a study with an outcome, y, a treatment indicator, z, 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 z 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.

7. *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 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, because the decisions of whether to re-randomize 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 from the model fit in part (b).
8. *Causal inference based on data from individual choices:* our lives involve tradeoffs 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 13.7. 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 (2004), 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 exposure to radon gas.
9. *Estimating causal effects:* 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.