CHAPTER 10

Causal inference using more advanced models

Chapter 9 discussed situations in which it is dangerous to use a standard linear regression of outcome on predictors and an indicator variable for estimating causal effects: when there is imbalance or lack of complete overlap or when ignorability is in doubt. This chapter discusses these issues in more detail and provides potential solutions for each.

10.1 Imbalance and lack of complete overlap

In a study comparing two treatments (which we typically label "treatment" and "control"), causal inferences are cleanest if the units receiving the treatment are comparable to those receiving the control. Until Section 10.5, we shall restrict ourselves to ignorable models, which means that we only need to consider observed pre-treatment predictors when considering comparability.

For ignorable models, we consider two sorts of departures from comparability—*imbalance* and *lack of complete overlap*. Imbalance occurs if the distributions of relevant pre-treatment variables differ for the treatment and control groups. Lack of complete overlap occurs if there are regions in the space of relevant pre-treatment variables where there are treated units but no controls, or controls but no treated units.

Imbalance and lack of complete overlap are issues for causal inference largely because they force us to rely more heavily on model specification and less on direct support from the data.

When treatment and control groups are *unbalanced*, the simple comparison of group averages, $\bar{y}_1 - \bar{y}_0$, is not, in general, a good estimate of the average treatment effect. Instead, some analysis must be performed to adjust for pre-treatment differences between the groups.

When treatment and control groups do not completely *overlap*, the data are inherently limited in what they can tell us about treatment effects in the regions of nonoverlap. No amount of adjustment can create direct treatment/control comparisons, and one must either restrict inferences to the region of overlap, or rely on a model to extrapolate outside this region.

Thus, lack of complete overlap is a more serious problem than imbalance. But similar statistical methods are used in both scenarios, so we discuss these problems together here.

Imbalance and model sensitivity

When attempting to make causal inferences by comparing two samples that differ in terms of the "treatment" or causing variable of interest (participation in a program, taking a drug, engaging in some activity) but that also differ in terms of confounding covariates (predictors related both to the treatment and outcome), we can be misled

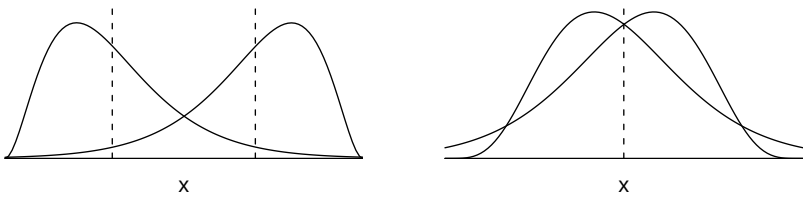

Figure 10.1 *Imbalance in distributions across treatment and control groups. (a) In the left panel, the groups differ in their averages (dotted vertical lines) but cover the same range of x. (b) The right panel shows a more subtle form of imbalance, in which the groups have the same average but differ in their distributions.*

if we do not appropriately control for those confounders. The examples regarding the effect of a treatment on health outcomes in Section 9.1 illustrated this point in a simple setting.

Even when all the confounding covariates are measured (hence ignorability is satisfied), however, it can be difficult to properly control for them if the distributions of the predictors are not similar across groups. Broadly speaking, any differences across groups can be referred to as lack of *balance* across groups. The terms "imbalance" and "lack of balance" are commonly used as a shorthand for differences in averages, but more broadly they can refer to more general differences in distributions across groups. Figure 10.1 provides two examples of imbalance. In the first case the groups have different means (dotted vertical lines) and different skews. In the second case groups have the same mean but different skews. In both examples the standard deviations are the same across groups though differences in standard deviation might be another manifestation of imbalance.

Imbalance creates problems primarily because it forces us to rely more on the correctness of our model than we would have to if the samples were balanced. To see this, consider what happens when we try to make inferences about the effect of a treatment variable, for instance a new reading program, on test score, y, while controlling for a crucial confounding covariate, pre-test score, x. Suppose that the true treatment effect is θ and the relations between the response variable, y, and the sole confounding covariate, x, is quadratic, as indicated by the following regressions, written out separately for the members of each treatment group:

$$\begin{aligned} \text{treated:}\ y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \theta + \text{error}_i \\ \text{controls:}\ y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \text{error}_i \end{aligned}$$

Averaging over each treatment group separately, solving the second equation for β_0, plugging back into the first, and solving for θ yields the estimate,

$$\hat{\theta} = \bar{y}_1 - \bar{y}_0 - \beta_1(\bar{x}_1 - \bar{x}_0) - \beta_2(\overline{x_1^2} - \overline{x_0^2}), \tag{10.1}$$

where $\bar{y}_1$ and $\bar{y}_0$ denote the average of the outcome test scores in the treatment and control groups respectively, $\bar{x}_1$ and $\bar{x}_0$ represent average pre-test scores for treatment and control groups respectively, and $\overline{x_1^2}$ and $\overline{x_0^2}$ represent these averages for squared pre-test scores. Ignoring x (that is, simply using the raw treatment/control comparison $\bar{y}_1 - \bar{y}_0$) is a poor estimate of the treatment effect: it will be off by the amount $\beta_1(\bar{x}_1 - \bar{x}_0) + \beta_2(\overline{x_1^2} - \overline{x_0^2})$, which corresponds to systematic pre-treatment differences between groups 0 and 1. The magnitude of this bias depends on how different the distribution of x is across treatment and control groups (specifically with regard to variance in this case) and how large β_1 and β_2 are. The closer the

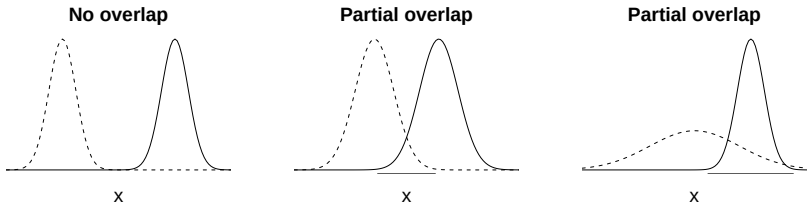

Figure 10.2 *Lack of complete overlap in distributions across treatment and control groups. Dashed lines indicate distributions for the control group; solid lines indicate distributions for the treatment group. (a) Two distributions with no overlap; (b) two distributions with partial overlap; (c) a scenario in which the* range *of one distribution is a subset of the range of the other.*

distributions of pre-test scores across treatment and control groups, the smaller this bias will be.

Moreover, a linear model regression using x as a predictor would also yield the wrong answer; it will be off by the amount $\beta_2(\overline{x_1^2} - \overline{x_0^2})$. The closer the distributions of pre-test scores across treatment and control groups, however, the smaller $(\overline{x_1^2} - \overline{x_0^2})$ will be, and the less worried we need to be about correctly specifying this model as quadratic rather than linear.

Lack of complete overlap and model extrapolation

Overlap describes the extent to which the range of the data is the same across treatment groups. There is *complete overlap* if this range is the same in the two groups. Figure 10.1 illustrated treatment and control confounder distributions with complete overlap.

As discussed briefly in the previous chapter, lack of complete overlap creates problems because it means that there are treatment observations for which we have no counterfactuals (that is, control observations with the same covariate distribution) and vice versa. A model fitted to data such as these is forced to extrapolate beyond the support of the data. The illustrations in Figure 10.2 display several scenarios that exhibit lack of complete overlap.

If these are distributions for an important confounding covariate, then areas where there is no overlap represent observations about which we may not want to make causal inferences. Observations in these areas have no empirical counterfactuals. Thus, any inferences regarding these observations would have to rely on modeling assumptions in place of direct support from the data. Adhering to this structure would imply that in the setting of Figure 10.2a, it would be impossible to make data-based causal inferences about any of the observations. Figure 10.2b shows a scenario in which data-based inferences are only possible for the region of overlap, which is underscored on the plot. In Figure 10.2c, causal inferences are possible for the full treatment group but only for a subset of the control group (again indicated by the underscored region).

Example: evaluating the effectiveness of high-quality child care

We illustrate with data collected regarding the development of nearly 4500 children born in the 1980s. A subset of 290 of these children who were premature and with low birth weight (between 1500 and 2500 grams) received special services in the first few years of life, including high-quality child care (five full days a week) in the

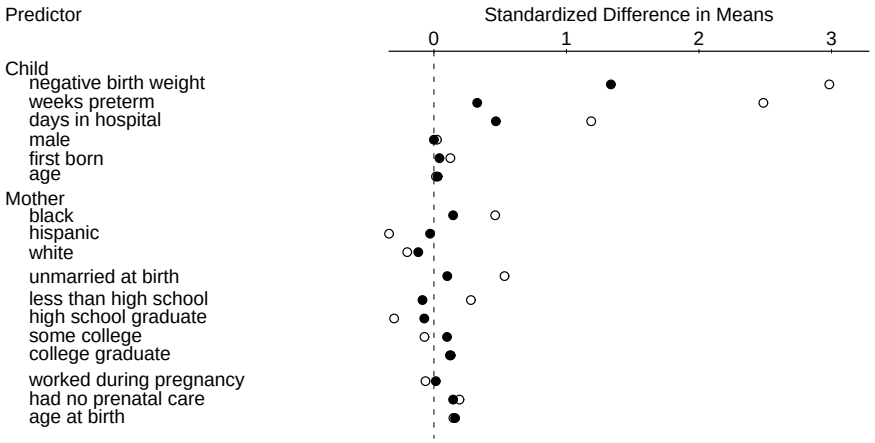

Figure 10.3 *Imbalance in averages of confounding covariates across treatment groups. Open circles represent differences in averages for the unmatched groups standardized by the pooled within-group standard deviations for unmatched groups. Solid circles represent differences in averages for matched groups standardized by the pooled within-group standard deviation for unmatched groups to facilitate comparisons. Negative birth weight is defined as 2500 grams minus the child's weight at birth.*

second and third years of life as part of a formal intervention (the Infant Health and Development Program). We want to evaluate the impact of this intervention on the children's subsequent cognitive outcomes by comparing the outcomes for children in the intervention group to the outcomes in a comparison group of 4091 children who did not participate in the program. The outcome of interest is test score at age 3; this test is similar to an IQ measure so we simplistically refer to these scores as IQ scores from now on.

Missing data. Incomplete data arise in virtually all observational studies. For this sample dataset, we imputed missing data once, using a model-based random imputation (see Chapter 25 for a general discussion of this approach). We excluded the most severely low-birth-weight children (those at or below 1500 grams) from the sample because they are so different from the comparison sample. For these reasons, results presented here do not exactly match the complete published analysis, which multiply imputed the missing values.

Examining imbalance for several covariates

To illustrate the ways in which the treated and comparison groups differ, the open circles in Figure 10.3 display the standardized differences in mean values (differences in averages divided by the pooled within-group standard deviations for the treatment and control groups) for a set of confounding covariates that we think predict both program participation and subsequent test scores. Many of these differences are large given that they are shown in standard-deviation units.

Setting up the plot to reveal systematic patterns of imbalance. In Figure 10.3, the characteristics of this sample are organized by whether they pertain to the child or to the mother. Additionally, continuous and binary predictors have been coded when possible such that the larger values are typically associated with lower test scores for children. For instance, "negative birth weight" is defined as the child's birth weight subtracted from 2500 grams, the cutoff for the official designation of

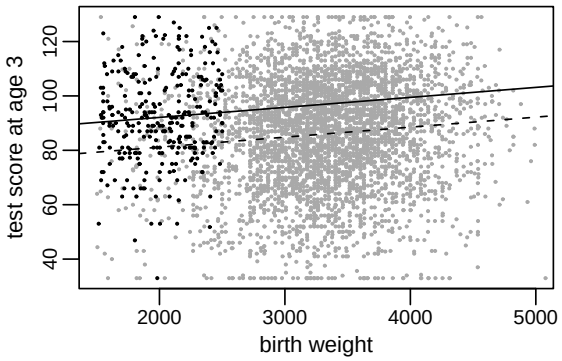

Figure 10.4 *Data from an intervention targeting low birth weight, premature children (black dots), and data from a comparison group of children (gray dots). Test scores at age 3 are plotted against birth weight. The solid line and dotted lines are regressions fit to the black and gray points, respectively.*

low birth weight. Therefore, high values of this predictor reflect children whom we would expect to have lower test scores than children with lower values for negative birth weight. Categorical variables have been broken out into indicators for each category and organized so that the category associated with lowest test scores comes first.

Displaying the confounders in this way and plotting standardized averages—rather than displaying a table of numbers—facilitate comparisons across predictors and methods (the dark points, to be described later, correspond to results obtained from another strategy) and allow us to more clearly identify trends when they exist. For instance, compared to the control group, the at-risk treatment group generally has characteristics associated with lower test scores—such as low birth weight for the child (coded as high "negative birth weight"), mother unmarried at birth, and mother not a high school graduate.

Figure 10.4, which shows a scatterplot and regression lines of test scores on birth weight, illustrates that, not only do the average birth weights differ in the two groups (lack of balance), but there are many control observations (gray dots) who have birth weights far out of the range of birth weights experienced in the treatment population (black dots). This is an example of *lack of complete overlap* in this predictor across groups. If birth weight is a confounding covariate that we need to control for to achieve ignorability, Figure 10.4 demonstrates that if we want to make inferences about the effect of the program on children with birth weights above 2500 grams, we will have to rely on model extrapolations that may be inappropriate.

Imbalance is not the same as lack of overlap

Figure 10.5 illustrates the distinction between balance and overlap. Imbalance does not necessarily imply lack of complete overlap; conversely, lack of complete overlap does not necessarily necessarily result in imbalance in the sense of different average values in the two groups. Ultimately, lack of overlap is a more serious problem, corresponding to a lack of data that limits the causal conclusions that can be made without uncheckable modeling assumptions.

Figure 10.5a demonstrates complete overlap across groups in terms of mother's education. Each category includes observations in each treatment group. However,

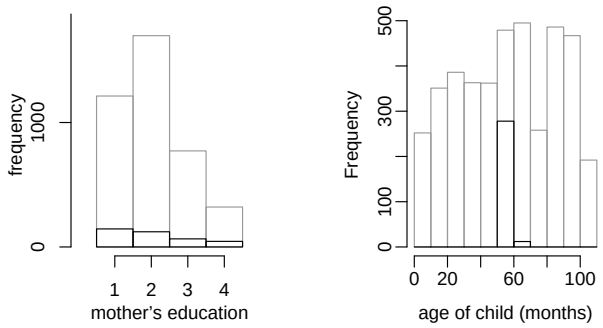

Figure 10.5 *Comparisons of the treatment (black histogram bars) and control (gray histogram bars) groups for the child-intervention study, with respect to two of the pretreatment variables. There is lack of complete overlap for child age, but the averages are similar across groups. In contrast, mother's education shows complete overlap, but imbalance exists in that the distributions differ for the two groups.*

the percentages falling in each category (and the overall average, were we to code these categories as 1–4) differ when comparing treatment and control groups—thus there is clearly imbalance.

Figure 10.5b shows balance in mean values but without complete overlap. As the histograms show, the averages of children's ages differ little across treatment groups, but the vast majority of control children have ages that are not represented in the treatment group. Thus there is a lack of complete overlap across groups for this variable. More specifically, there is complete overlap in terms of the treatment observations, but not in terms of the control observations. If we believe age to be a crucial confounding covariate, we probably would not want to make inferences about the full set of controls in this sample.

10.2 Subclassification: effects and estimates for different subpopulations

Assuming we are willing to trust the ignorability assumption, how can we assess whether we are relying too strongly on modeling assumptions? And if we are uncertain of our assumptions, how can we proceed cautiously? Section 9.5 illustrated a check for overlap in one continuous predictor across treatment groups. In this section we demonstrate a check that accommodates many predictors and discuss options for more flexible modeling.

Subclassification

We saw in Chapter 3 that mother's educational attainment is an important predictor of her child's test scores. Education level also traditionally is associated with participation in interventions such as this program for children with low birth weights. Let us make the (unreasonable) assumption for the moment that this is the only confounding covariate (that is, the only predictor associated with both participation in this program and test scores). How would we want to estimate causal effects? In this case a simple solution would be to estimate the difference in mean test scores within each subclass defined by mother's education. These averages as well as the associated standard error and sample size in each subclass are displayed in Figure 10.6. These point to positive effects for all participants, though not all

Mother's education	Treatment effect estimate ± s.e.	Sample size treated	controls
Not a high school grad	9.3 ± 1.3	126	1358
High school graduate	4.0 ± 1.8	82	1820
Some college	7.9 ± 2.3	48	837
College graduate	4.6 ± 2.1	34	366

Figure 10.6 *Estimates ± standard errors of the effect on children's test scores of a child care intervention, for each of four subclasses formed by mother's educational attainment. The study was of premature infants with low birth weight, most of whom were born to mothers with low levels of education.*

effects are statistically significant, with by far the largest effects for the children whose mothers had not graduated from high school.

Recall that there is overlap on this variable across the treatment and control groups as is evidenced by the sample sizes for treated and control observations within each subclass in Figure 10.6. If there were a subclass with observations only from one group, we would not be able to make inferences for this type of person. Also, if there were a subclass with only a small number of observations in either the treatment group or the control group, we would probably be wary of making inferences for these children as well.

To get an estimate of the overall effect for those who participated in the program, the subclass-specific estimates could be combined using a weighted average where the weights are defined by the number of children in each subclass who participated in the program:

$$\text{Est. effect on the treated} = \frac{9.3 \cdot 126 + 4.0 \cdot 82 + 7.9 \cdot 48 + 4.6 \cdot 34}{126 + 82 + 48 + 34} = 7.0, \quad (10.2)$$

with a standard error of $\sqrt{\frac{1.3^2 \cdot 126^2 + 1.8^2 \cdot 82^2 + 2.3^2 \cdot 5.3^2 + 2.1^2 \cdot 34^2}{(126+82+48+34)^2}} = 0.9$.

This analysis is similar to a regression with interactions between the treatment and mother's educational attainment. To calculate the average treatment effect for program participants, we would have to poststratify—that is, estimate the treatment effect separately for each category of mother's education, and then average these effects based on the distribution of mother's education in the population.

This strategy has the advantage of imposing overlap and, moreover, forcing the control sample to have roughly the same covariate distribution as the treated sample. This reduces reliance on the type of model extrapolations discussed previously. Moreover, one can choose to avoid modeling altogether after subclassifying, and simply can take a difference in averages across treatment and control groups to perform inferences, therefore completely avoiding making assumptions about the parametric relation between the response and the confounding covariates.

One drawback of subclassifying, however, is that when controlling for a continuous variable, some information may be lost when discretizing the variable. A more substantial drawback is that it is difficult to control for many variables at once.

Average treatment effects: whom do we average over?

Figure 10.6 demonstrated how treatment effects can vary over different subpopulations. Why did we weight these subclass-specific estimates by the number of treated children in each subclass rather than the total number of children in each subclass?

For this application, we are interested in the effect of the intervention *for the sort of children who would have participated in it*. Weighting using the number of treatment children in each subclass forces the estimate implicitly to be representative of the treatment children we observe. The effect we are trying to estimate is sometimes called the *effect of the treatment on the treated*.

If we had weighted instead by the number of control children in each subclass we could estimate the effect of the treatment on the controls. However, this particular intervention was designed for the special needs of low-birth-weight, premature children—not for typical children—and there is little interest in its effect on comparison children who would not have participated.

The effect of the intervention might vary, for instance, for children with different initial birth weights, and since we know that the mix of children's birth weights differs in treatment and comparison groups, the average effects across these groups could also differ. Moreover, we saw in Figure 10.4 that there are so many control observations with no counterfactual observations in the treatment group with regard to birth weight that these data are likely inappropriate for drawing inferences about the control group either directly (the effect of the treatment on the controls) or as part of an average effect across the entire sample.

Again, this is related to poststratification. We can think of the estimate of the effect of the treatment on the treated as a poststratified version of the estimate of the average causal effect. As the methods we discuss in this section rely on more and more covariates, however, it can be more attractive to apply methods that more directly estimate the effect of the treatment on the treated, as we discuss next.

10.3 Matching: subsetting the data to get overlapping and balanced treatment and control groups

Matching refers to a variety of procedures that restrict and reorganize the original sample in preparation for a statistical analysis. In the simplest form of matching, one-to-one matching, the data points are divided into pairs—each containing one treated and one control unit—with the two units matched into a pair being as similar as possible on relevant pre-treatment variables. The number of units in the two groups will not in general be equal—typically there are more controls than treated units, as in Figure 10.5, for example—and so there will be some leftover units unmatched. In settings with poor overlap, there can be unmatched units from both groups, so that the matched pairs represent the region of data space where the treatment and control groups overlap.

Once the matched units have been selected out of the larger dataset, they can be analyzed by estimating a simple difference in average outcomes across treatment groups or by using regression methods to estimate the effect of the treatment in the area of overlap.

Matching and subclassification

Matching on one variable is similar to subclassification except that it handles continuous variables more precisely. For instance, a treatment observation might be matched to control observations that had the closest age to their own as opposed to being grouped into subclasses based on broader age categories. Thus, matching has the same advantages of stratification in terms of creating balance and forcing overlap, and may even be able to create slightly better balance. However many

matching methods discard observations even when they are within the range of overlap, which is likely inefficient.

Matching has some advantages over subclassification when controlling for many variables at once. Exact matching is difficult with many confounders, but "nearest-neighbor" matching is often still possible. This strategy matches treatment units to control units that are "similar" in terms of their confounders where the metric for similarity can be defined in any variety of ways, one of the most popular being the *Mahalanobis distance*, which is defined in matrix notation as $d(x^{(1)}, x^{(2)}) = (x^{(1)} - x^{(2)})^t \Sigma^{-1} (x^{(1)} - x^{(2)})$, where $x^{(1)}$ and $x^{(2)}$ represent the vectors of predictors for points 1 and 2, and Σ is the covariance of the predictors in the dataset. Recently, other algorithms have been introduced to accomplish this same task—finding similar treatment and control observations—that rely on algorithms originally created for genetic or data mining applications. Another matching approach, which we describe next, compares the input variables for treatment and control cases in order to find an effective scale on which to match.

Propensity score matching

One way to simplify the issue of matching or subclassifying on many confounding covariates at once is to create a one-number summary of all the covariates and then use this to match or subclassify. We illustrate using a popular summary, the propensity score, with our example of the intervention for children with low birth weights. It seems implausible that mother's education, for example, is the only predictor we need to satisfy the ignorability assumption in our example. We would like to control for as many predictors as possible to allow for the possibility that any of them is a confounding covariate. We also want to maintain the beneficial properties of matching. How can we match on many predictors at once?

Propensity score matching provides a solution to this problem. The *propensity score* for the i^{th} individual is defined as the probability that he or she receives the treatment given everything we observe before the treatment (that is, all the confounding covariates for which we want to control). Propensity scores can be estimated using standard models such as logistic regression, where the outcome is the treatment indicator and the predictors are all the confounding covariates. Then matches are found by choosing for each treatment observation the control observation with the closest propensity score.

In our example we randomly ordered the treatment observations, and then each time a control observation was chosen as a match for a given treatment observation it could not be used again. More generally, methods have been developed for matching multiple control units to a single treated unit, and vice versa; these ideas can be effective, especially when there is overlap but poor balance (so that, for example, some regions of predictor space contain many controls and few treated units, or the reverse). From this perspective, matching can be thought of as a way of discarding observations so that the remaining data show good balance and overlap.

The goal of propensity score matching is not to ensure that each *pair of matched observations* is similar in terms of all their covariate values, but rather that the matched groups are similar *on average* across all their covariate values. Thus, the adequacy of the model used to estimate the propensity score can be evaluated by examining the balance that results on average across the matched groups.

Computation of propensity score matches

The first step in creating matches is to fit a model to predict who got the intervention based on the set of predictors we think are necessary to achieve ignorability (confounding covariates). A natural starting point would be a logistic regression, something like,

R code
```
ps.fit.1 <- glm (treat ~ as.factor(educ) + as.factor(ethnic) + b.marr  +
  work.dur + prenatal + mom.age + sex + first + preterm + age +
  dayskidh + bw + unemp.rt, data=cc2, family=binomial(link="logit"))
```

In our example, we evaluated several different model fits before settling on one that provided balance that seemed adequate. In each case we evaluated the adequacy of the model by evaluating the balance that resulted from matching on the estimated propensity scores from that model. Model variations tried excluding variables and including interactions and quadratic terms. We finally settled on,

R code
```
ps.fit.2 <- glm (treat ~ bwg + as.factor(educ) + bwg:as.factor(educ) +
  as.factor(ethnic) + b.marr + as.factor(ethnic):b.marr +
  work.dur + prenatal + preterm + age + mom.age + sex + first,
  data=cc2, family=binomial(link="logit"))
```

We then create predicted values:[1]

R code
```
pscores <- predict (ps.fit.2, type="link")
```

The regression model is messy, but we are not concerned with all its coefficients; we are only using it as a tool to construct a balanced comparison between treatment and control groups. We used the estimated propensity scores to create matches, using a little R function called `matching` that finds for each treatment unit in turn the control unit (not previously chosen) with the closest propensity score:[2]

R code
```
matches <- matching (z=cc2$treat, score=pscores)
matched <- cc2[matches$matched,]
```

Then the full dataset was reduced to only the treated observations and only those control observations that were chosen as matches.

The differences between treated and control averages, for the matched subset, are displayed by the solid dots in Figure 10.3. The imbalance has decreased noticeably compared to the unmatched sample. Certain variables (birth weight and the number of days the children were in the hospital after being born) still show imbalance, but none of our models succeeded in balancing those variables. We hope the other variables are more important in predicting future test scores (which appears to be reasonable from the previous literature on this topic).

The process of fitting, assessing, and selecting a model for the propensity scores has completely ignored the outcome variable. We have judged the model solely by the balance that results from subsequent matches on the associated propensity scores. This helps the researcher to be "honest" when fitting the propensity score model because a treatment effect estimate is not automatically produced each time a new model is fit.

[1] We use the `type="link"` option to get predictions on the scale of the linear predictor, that is, $\tilde{X}\beta$. If we wanted predictions on the probability scale, we would set `type="response"`. In this example, similar results would arise from using either approach.

[2] Here we have performed the matching mostly "manually" in the sense of setting up a regression on the treatment variable and then using the predicted probabilities to select a subset of matched units for the analysis. Various more automatic methods for propensity score estimation, matching, and balancing have be implemented in R and other software packages; see the end of this chapter for references.

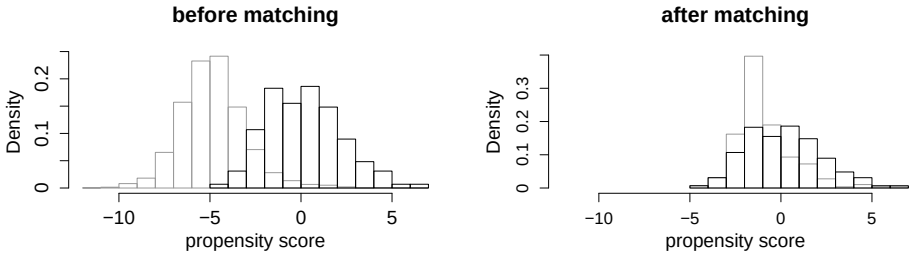

Figure 10.7 *(a) Distribution of logit propensity scores for treated (dark lines) and control groups (gray lines)* before *matching. (b) Distributions of logit propensity scores for treated (dark lines) and control groups (gray lines)* after matching.

Having created and checked appropriateness of the matches by examining balance, we fit a regression model just on the matched data including all the predictors considered so far, along with an indicator to estimate the treatment effect:

R code

```
reg.ps <- lm (ppvtr.36 ~ treat + hispanic + black + b.marr + lths +
  hs + ltcoll + work.dur + prenatal + mom.age + sex + first +
  preterm + age + dayskidh + bw, data=matched)
```

Given the balance and overlap that the matching procedure has achieved, we are less concerned than in the standard regression context about issues such as deviations from linearity and model extrapolation. Our estimated treatment effect from the matched dataset is 10.2 (with a standard error of 1.6), which can be compared to the standard regression estimate of 11.7 (with standard error of 1.3) based on the full dataset.

If we fully believed in the linear model and were confident that it could be extrapolated to the areas of poor overlap, we would use the regression based on all the data. Realistically, however, we prefer to construct comparable groups and restrict our attention to the range of overlap.

Insufficient overlap? What happens if there are observations about which we want to make inferences but there are no observations with similar propensity scores in the other group? For instance, suppose we are interested in the effect of the treatment on the treated but there are some treated observations with propensity scores far from the propensity scores of all the control observations. One option is to accept some lack of comparability (and corresponding level of imbalance in covariates). Another option is to eliminate the problematic treated observations. If the latter choice is made it is important to be clear about the change in the population about whom inferences will now generalize. It is also helpful to try "profile" the observations that are omitted from the analysis.

Matched pairs? Although matching often results in pairs of treated and control units, we typically ignore the pairing in the analysis of the matched data. Propensity score matching works well (in appropriate settings) to create matched groups, but it does not necessarily created closely matched *pairs*. It is not generally appropriate to add the complication of including the pairing in the model, because the pairing in the matching is performed in the analysis, not the data collection. However, pairing in this way does affect variance calculations, as we shall discuss.

The propensity score as a one-number summary used to assess balance and overlap

A quick way of assessing whether matching has achieved increased balance and overlap is to plot histograms of propensity scores across treated and control groups.

Figure 10.7 displays these histograms for unmatched and matched samples. (We plot the propensity scores on the logit scale to better display their variation at the extremes, which correspond to probabilities near 0 and 1.) The decreased imbalance and increased overlap illustrated in the histograms for the matched groups do not ensure that all predictors included in the model will be similarly matched, but they provide some indication that these distribution will have closer balance in general than before matching.

Geographic information

We have excluded some important information from these analyses. We have access to indicators reflecting the state in which each child resides. Given the tremendous variation in test scores and child care quality[3] across states, it seems prudent to control for this variable as well. If we redo the propensity score matching by including state indicators in both the propensity score model and final regression model, we get an estimate of 8.8 (with standard error of 2.1), which is even lower than our original estimate of 10.2. Extending the regression analysis on the full dataset to include state indicators changes the estimate only from 11.7 to 11.6.

We include results from this analyses using classical regression to adjust for states because it would be a standard approach given these data. A better approach would be to include states in a multilevel model, as we discuss in Chapter 23.

Experimental benchmark by which to evaluate our estimates

It turns out that the researchers evaluating this intervention did not need to rely on a comparison group strategy to assess its impact on test scores. The intervention was evaluated using a randomized experiment. In the preceding example, we simply replaced the true experimental control group with a comparison group pulled from the National Longitudinal Survey of Youth. The advantage of this setup as an illustration of propensity score matching is that we can compare the estimates obtained from the observational study that we have "constructed" to the estimates found using the original randomized experiment. For this sample, the experimental estimate is 7.4. Thus, both propensity score estimates are much closer to the best estimate of the true effect than the standard regression estimates.

Subclassification on mother's education alone yields an estimated treatment effect of 7.0, which happens to be close to the experimental benchmark. However, this does not imply that subclassifying on one variable is generally the best strategy overall. In this example, failure to control for all confounding covariates leads to many biases (some negative and some positive—the geographic variables complicate this picture), and unadjusted differences in average outcomes yield estimates that are lower than the experimental benchmark. Controlling for one variable appears to work well for this example because the biases caused by the imbalances in the other variables just happen to cancel. We would not expect this to happen in general.

Other matching methods, matching on all covariates, and subclassification

The method we have illustrated is called *matching without replacement* because any given control observation cannot be used as a match for more than one treatment

[3] Variation in quality of child care is important because it reflects one of the most important alternatives that can be chosen by the parents in the control group.

observation. This can work well in situations when there is a large enough control group to provide adequate overlap. It has the advantage of using each control observation only once, which maximizes our sample size (assuming a constraint of one match per treatment unit) and makes variance calculations a bit easier; see the discussion of standard errors at the end of this section.

However, situations arise when there are not enough controls in the overlapping region to fully provide one match per treated unit. In this case it can help to use some control observations as matches for more than one treated unit. This approach is often called *matching with replacement*, a term which commonly refers to with one-to-one matching but could generalize to multiple control matches for each control. Such strategies can create better balance, which should yield estimates that are closer to the truth on average. Once such data are incorporated into a regression, however, the multiple matches reduce to single data points, which suggests that matching with replacement has limitations as a general strategy.

A limitation of one-to-one matching is that it may end up "throwing away" many informative units if the control group is substantially bigger than the treatment group. One way to make better use of the full sample is simply to subclassify based on values of the propensity score—perhaps discarding some noncomparable units in the tails of the propensity score distribution. Then separate analyses can be performed within each subclass (for example, difference in outcome averages across treatment groups or linear regressions of the outcome on an indicator variable for treatment and other covariates). The estimated treatment effects from each of the subclasses then can either be reported separately or combined in a weighted average with different weights used for different estimands. For instance, when estimating the effect of the treatment on the treated, the number of treated observations in each subclass would be used as the weight, just as we did for the simple subclassification of mother's education in model (10.2) on page 205.

A special case of subclassification called "full matching" can be conceptualized as a fine stratification of the units where each statum has either (1) one treated unit and one control unit, (2) one treated unit and multiple control units, or (3) multiple treated units and one control unit. "Optimal" versions of this matching algorithm have the property of minimizing the average distance between treatment and control units. Strategies with non-overlapping strata such as subclassification and full matching have the advantage of being more easily incorporated into larger models. This enables strata to be modeled as groups in any number of ways.

Other uses for propensity scores

Some researchers use the propensity score in other ways. For instance, the inverse of estimated propensity scores can be used to create a weight for each point in the data, with the goal that weighted averages of the data should look, in effect, like what would be obtained from a randomized experiment. For instance, to obtain an estimate of an average treatment effect, one would use weights of $1/p_i$ and $1/(1-p_i)$ for treated and control observations i, respectively, where the p_i's are the estimated propensity scores. To obtain an estimate of a the effect of the treatment on the treated, one would use weights of 1 for the treated and $p_i/(1-p_i)$ for the controls. These weights can be used to calculate simple means or can be included within a regression framework. In our example, this method yielded a treatment effect estimate of 7.8 (when including state information), which is close to the experimental benchmark.

These strategies have the advantage (in terms of precision) of retaining the full

sample. However, the weights may have wide variability and may be sensitive to model specification which could lead to instability. Therefore, these strategies work best when care is taken to create stable weights and to use robust or nonparametric models to estimate the weights. Such methods are beyond the scope of this book.

More simply, propensity scores can be used in a regression of the outcome on the treatment and the scores rather than the full set of covariates. However, if observations that lie in areas where there is no overlap across treatment groups are not removed, the same problems regarding model extrapolation will persist. Also, this method once again places a great deal of faith in precise and correct estimation of the propensity score.

Finally, generalizations of the binary treatment setup have been formalized to accommodate multiple-category or continuous treatment variables.

Standard errors

The standard errors presented for the analyses fitted to matched samples are not technically correct. First, matching induces correlation among the matched observations. The regression model, however, if correctly specified, should account for this by including the variables used to match. Second, our uncertainty about the true propensity score is not reflected in our calculations. This issue has no perfect solution to date and is currently under investigation by researchers in this field. Moreover, more complicated matching methods (for example, matching with replacement and many-to-one matching methods) generally require more sophisticated approaches to variance estimation. Ultimately, one good solution may be a multilevel model that includes treatment interactions so that inferences explicitly recognize the decreased precision that can be obtained outside the region of overlap.

10.4 Lack of overlap when the assignment mechanism is known: regression discontinuity

Simple regression works to estimate treatment effects under the assumption of ignorable treatment assignment if the model is correct, or if the confounding covariates are well balanced with respect to the treatment variable, so that regression serves as a fine-tuning compared to a simple difference of averages. But if the treated and control groups are very different from each other, it can be more appropriate to identify the subset of the population with overlapping values of the predictor variables for both treatment and control conditions, and to estimate the causal effect (and the regression model) in this region only. Propensity score matching is one approach to lack of overlap.

If the treatment and control groups do not overlap at all in key confounding covariates, it can be prudent to abandon causal inferences altogether. However, sometimes a clean lack of overlap arises from a covariate that itself was used to assign units to treatment conditions. *Regression discontinuity analysis* is an approach for dealing with this extreme case of lack of overlap in which the assignment mechanism is clearly defined.

Regression discontinuity and ignorability

A particularly clear case of imbalance sometimes arises in which there is some pretreatment variable x, with a cutoff value C so that one of the treatments applies for all units i for which $x_i < C$, and the other treatment applies for all units for

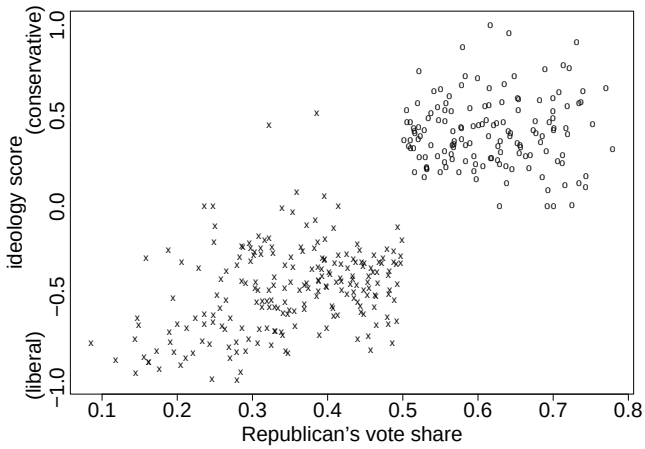

Figure 10.8 *Example of a regression discontinuity analysis: political ideology of members of the 1993–1994 House of Representatives versus Republican share of the two-party vote in the district's congressional election in 1992. Democrats and Republicans are indicated by crosses and circles, respectively. For the purpose of estimating the effect of electing a Democrat or Republican, there is no overlap between the "treatment" (the congressmember's party) and the pre-treatment control variable on the x-axis.*

which $x_i > C$. This could occur, for example, in a medical experiment in which a risky new treatment is only given to patients who are judged to be in particularly bad condition. But the usual setting is in observational studies, where a particular event or "treatment" only occurs under certain specified conditions. For example, in a two-candidate election, a candidate wins if and only if he or she receives more than half the vote.

In a setting where one treatment occurs only for $x < C$ and the other only for $x > C$, it is still possible to estimate the treatment effect for units with x in the neighborhood of C, *if* we assume that the regression function—the average value of the outcome y, given x and the treatment—is a continuous function of x near the cutoff value C.

In this scenario, the mechanism that assigns observations to treatment or control is known, and so we need not struggle to set up a model in which the ignorability assumption is reasonable. All we need to do is control for the input(s) used to determine treatment assignment—these are our confounding covariates. The disadvantage is that, by design, there is no overlap on this covariate across treatment groups. Therefore, to "control for" this variable we must make stronger modeling assumptions because we will be forced to extrapolate our model out of the range of our data. To mitigate such extrapolations, one can limit analyses to observations that fall just above and below the threshold for assignment.

Example: political ideology of congressmembers

Figure 10.8 shows an example, where the goal is to estimate one aspect of the effect of electing a Republican, as compared to a Democrat, in the U.S. House of Representatives. The graph displays political ideologies (as computed using a separate statistical analysis of congressional roll-call votes) for Republican and Democratic congressmembers, plotted versus the vote received by the Republican candidate in the previous election. There is no overlap because the winner in each district nec-

essarily received at least 50% of the vote. (For simplicity, we are only considering districts where an incumbent was running for reelection, so that different districts with the same congressional vote share can be considered as comparable.)

Regression discontinuity analysis. If we wish to consider the effect of the winning party on the political ideology of the district's congressmember, then a simple regression discontinuity analysis would consider a narrow range—for example, among all the districts where x lies between 0.45 and 0.55, and then fit a model of the form,

$$y_i = \beta_0 + \theta T_i + \beta_1 x_i + \text{error}_i$$

where T_i is the "treatment," which we can set to 1 for Republicans and 0 for Democrats.

Here is the result of the regression:

R output

```
lm(formula = score1 ~ party + x, subset=overlap)
            coef.est coef.se
(Intercept)    -1.21    0.62
party           0.73    0.07
x               1.65    1.31
  n = 68, k = 3
  residual sd = 0.15, R-Squared = 0.88
```

The effect of electing a Republican (compared to a Democrat) is 0.73 (on a scale in which the most extreme congressmembers are at ± 1; see Figure 10.8) after controlling for the party strength in the district. The coefficient of x is estimated to be positive—congressmembers in districts with higher Republican votes tend to be more conservative, after controlling for party—but this coefficient is not statistically significant. The large uncertainty in the coefficient for x is no surprise, given that we have restricted our analysis to the subset of data for which x lies in the narrow range from 0.45 to 0.55.

Regression fit to all the data. Alternatively, we could fit the model to the whole dataset:

R output

```
lm(formula = score1 ~ party + x)
            coef.est coef.se
(Intercept)    -0.68    0.05
party           0.69    0.04
x               0.64    0.13
  n = 357, k = 3
  residual sd = 0.21, R-Squared = 0.8
```

The coefficient on x is estimated much more precisely, which makes sense given that we have more leverage on x (see Figure 10.8).

Regression with interactions. However, a closer look at the figure suggests different slopes for the two parties, and so we can fit a model interacting x with party:

R output

```
lm(formula = score1 ~ party + x + party:x)
            coef.est coef.se
(Intercept)    -0.76    0.06
party           1.13    0.16
x               0.87    0.15
party:x        -0.81    0.29
  n = 357, k = 4
  residual sd = 0.21, R-Squared = 0.81
```

Everything is statistically significant, but it is difficult to interpret these coefficients.

We shall reparameterize and define,

R code
```
z <- x - 0.5
```

so that when $z = 0$, we are at the point of discontinuity. We then reparameterize the interaction slope as separate slopes for the Democrats (`party==0`) and Republicans (`party==1`):

R output
```
lm(formula = score1 ~ party + I(z*(party==0)) + I(z*(party==1)))
                    coef.est coef.se
(Intercept)            -0.33    0.03
party                   0.73    0.04
I(z * (party == 0))     0.87    0.15
I(z * (party == 1))     0.06    0.24
  n = 357, k = 4
  residual sd = 0.21, R-Squared = 0.81
```

We see a strong positive slope of z among Democrats but not Republicans, and an estimate of 0.73 for the effect of party at the discontinuity point.

Comparison of regression discontinuity analysis to the model with interactions using all the data. In this particular example, the analysis fit to the entire dataset gives similar results (but with a much lower standard error) as compared to the regression discontinuity analysis that focused on the region of near overlap. In general, however, the model fit just to the area of overlap may be considered more trustworthy.

Partial overlap

What happens when the discontinuity is not so starkly defined? This is sometimes called a "fuzzy" discontinuity, as opposed to the "sharp" discontinuity discussed thus far. Consider, for instance, a situation where the decision whether to promote children to the next grade is made based upon results from a standardized test (or set of standardized tests). Theoretically this should create a situation with no overlap in these test scores across those children forced to repeat their grade and those promoted to the next grade (the treatment and control groups). In reality, however, there is some "slippage" in the assignment mechanism. Some children may be granted waivers from the official policy based on any of several reasons, including parental pressure on school administrators, a teacher who advocates for the child, and designation of the child as learning-disabled.

This situation creates partial overlap between the treatment and control groups in terms of the supposed sole confounding covariate, promotion test scores. Unfortunately, this overlap arises from deviations from the stated assignment mechanism. If the reasons for these deviations are well defined (and measurable), then ignorability can be maintained by controlling for the appropriate child, parent, or school characteristics. Similarly, if the reasons for these deviations are independent of the potential outcomes of interest, there is no need for concern. If not, inferences could be compromised by failure to control for important omitted confounders.

10.5 Estimating causal effects indirectly using instrumental variables

There are situations when the ignorability assumption seems inadequate because the dataset does not appear to capture all inputs that predict both the treatment and the outcomes. In this case, controlling for observed confounding covariates

through regression, subclassification, or matching will not be sufficient for calculating valid causal estimates because unobserved variables could be driving differences in outcomes across groups.

When ignorability is in doubt, the method of *instrumental variables* (IV) can sometimes help. This method requires a special variable, the *instrument*, which is predictive of the treatment and brings with it a new set of assumptions.

Example: a randomized-encouragement design

Suppose we want to estimate the effect of watching an educational television program (this time the program is Sesame Street) on letter recognition. We might consider implementing a randomized experiment where the participants are preschool children, the treatment of interest is watching Sesame Street, the control condition is not watching,[4] and the outcome is the score on a test of letter recognition. It is not possible here for the experimenter to force children to watch a TV show or to refrain from watching (the experiment took place while Sesame Street was on the air). Thus *watching* cannot be randomized. Instead, when this study was actually performed, what was randomized was *encouragement* to watch the show—this is called a randomized encouragement design.

A simple comparison of randomized groups in this study will yield an estimate of the effect of *encouraging* these children to watch the show, not an estimate of the effect of actually viewing the show. In this setting the simple randomized comparison is an estimate of a quantity called the *intent-to-treat* (ITT) effect. However, we may be able to take advantage of the randomization to estimate a causal effect for at least some of the people in the study by using the randomized encouragement as an "instrument." An instrument is a variable thought to randomly induce variation in the treatment variable of interest.

Assumptions for instrumental variables estimation

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

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

In addition, the model assumes no interference between units (the stable unit treatment value assumption) as with most other causal analyses, an issue we have already discussed at the end of Section 9.3.

Ignorability of the instrument

The first assumption in the list above is *ignorability of the instrument* with respect to the potential outcomes (both for the primary outcome of interest and the treatment variable). This is trivially satisfied in a randomized experiment (assuming the

[4] Actually the researchers in this study recorded four viewing categories: (1) rarely watched, (2) watched once or twice a week, (3) watched 3-5 times a week, and (4) watched more than 5 times a week on average. Since there is no a category for "never watched," for the purposes of this illustration we treat the lowest viewing category ("rarely watched") as if it were equivalent to "never watched."

randomization was pristine). In the absence of a randomized experiment (or natural experiment) this property may be more difficult to satisfy and often requires conditioning on other predictors.

Nonzero association between instrument and treatment variable

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

Monotonicity and the exclusion restrictions

Those children whose viewing patterns could be altered by encouragement are the only participants in the study for whom we can conceptualize counterfactuals with regard to viewing behavior—under different experimental conditions they might have been observed either viewing or not viewing, so a comparison of these potential outcomes (defined in relation to randomized encouragement) makes sense. We shall label these children "induced watchers"; these are the only children for whom we will make inferences about the effect of watching Sesame Street.

For the children who were encouraged to watch but did not, we might plausibly assume that they also would not have watched if not encouraged—we shall label this type of child a "never-watcher." We cannot directly estimate the effect of viewing for these children since in this context they would never be observed watching the show. Similarly, for the children who watched Sesame Street even though not encouraged, we might plausibly assume that if they had been encouraged they would have watched as well, again precluding an estimate of the effect of viewing for these children. We shall label these children "always-watchers."

Monotonicity. In defining never-watchers and always-watchers, we assumed that there were no children who would watch if they were not encouraged but who would *not* watch if they *were* encouraged. Formally this is called the *monotonicity assumption*, and it need not hold in practice, though there are many situations in which it is defensible.

Exclusion restriction. To estimate the effect of viewing for those children whose viewing behavior would have been affected by the encouragement (the induced watchers), we must make another important assumption, called the *exclusion restriction*. This assumption says for those children whose behavior would not have been changed by the encouragement (never-watchers and always-watchers) there is no effect of encouragement on outcomes. So for the never-watchers (children who would not have watched either way), for instance, we assume encouragement to watch did not affect their outcomes. And for the always-watchers (children who

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

Figure 10.9 *Hypothetical complete data in a randomized encouragement design. Units have been ordered for convenience. For each unit, the students are encouraged to watch Sesame Street* ($z_i = 1$) *or not* ($z_i = 0$). *This reveals which of the potential viewing outcomes* (T_i^0, T_i^1) *and which of the potential test outcomes* (y_i^0, y_i^1) *we get to observe. The observed outcomes are displayed in boldface. Here, potential outcomes are what we would observe under either encouragement option. The exclusion restriction forces the potential outcomes to be the same for those whose viewing would not be affected by the encouragement. The effect of watching for the "induced watchers" is equivalent to the intent-to-treat effect (encouragement effect over the whole sample) divided by the proportion induced to view, that is,* $3.4/0.4 = 8.5$.

would have watched either way), we assume encouragement to watch did not affect their outcomes.[5]

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

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

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

Figure 10.9 illustrates with hypothetical data based on the concepts in this real-life example by displaying for each study participant not only the observed data (encouragement and viewing status as well as observed outcome test score) but also the unobserved categorization, c_i, into always-watcher, never-watcher, or induced watcher based on potential watching behavior as well as the counterfactual test outcomes (the potential outcome corresponding to the treatment not received). Here, potential outcomes are the outcomes we would have observed under either *encouragement* option. Because of the exclusion restriction, for the always-watchers and the never-watchers the potential outcomes are the same no matter the encouragement (really they need not be *exactly* the same, just distributionally the same, but this simplifies the exposition).

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

$$\begin{aligned} \text{ITT} &= \frac{9+8+7+10+9+8+7+10+0+\cdots+0}{20} \\ &= 8.5 \cdot \frac{8}{20} + 0 \cdot \frac{12}{20} \\ &= 8.5 \cdot 0.4. \end{aligned} \tag{10.3}$$

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

Instrumental variables estimate

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

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

```
fit.1a <- lm (watched ~ encouraged)
```
R code

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

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

```
fit.1b <- lm (y ~ encouraged)
```
R code

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

[5] Technically, the assumptions regarding always-watchers and never-watchers represent distinct exclusion restrictions. In this simple framework, however, the analysis suffers if either assumption is violated. Using more complicated estimation strategies, it can be helpful to consider these assumptions separately as it may be possible to weaken one or the other or both.

R code
```
iv.est <- coef(fit.1a)[,"encouraged"]/coef(fit.1b)[,"encouraged"]
```

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

Local average treatment effects

The instrumental variables strategy here does not estimate an overall causal effect of watching Sesame Street across everyone in the study. The exclusion restriction implies that there is no effect of the instrument (encouragement) on the outcomes for always-watchers and for never-watchers. Given that the children in these groups cannot be induced to change their watching behavior by the instrument, we cannot estimate the causal effect of watching Sesame Street for these children. Therefore the causal estimates apply only to the "induced watchers."

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

10.6 Instrumental variables in a regression framework

Instrumental variables models and estimators can also be derived using regression, allowing us to more easily extend the basic concepts discussed in the previous section. A general instrumental variables model with continuous instrument, z, and treatment, d, can be written as,

$$\begin{aligned} y &= \beta_0 + \beta_1 T + \epsilon_i \\ T &= \gamma_0 + \gamma_1 z + \nu_i \end{aligned} \tag{10.4}$$

The assumptions can now be expressed in a slightly different way. The first assumption is that z_i is uncorrelated with both ϵ_i and ν_i, which translates informally into the ignorability assumption and exclusion restriction (here often expressed informally as "the instrument only affects the outcome *through* its effect on the treatment"). Also the correlation between z_i and t_i must be nonzero (parallel to the monotonicity assumption from the previous section). We next address how this framework identifies the causal effect of T on y.

Identifiability with instrumental variables

Generally speaking, *identifiability* refers to whether the data contain sufficient information for unique estimation of a given parameter or set of parameters in a particular model. For example, in our formulation of the instrumental variables model, the causal parameter is not identified without assuming the exclusion restriction (although more generally the exclusion restriction is not the only assumption that could be used to achieve identifiability).

What if we did not impose the exclusion restriction for our basic model? The model (ignoring covariate information, and switching to mathematical notation for

simplicity and generalizability) can be written as,

$$\begin{aligned} y &= \beta_0 + \beta_1 T + \beta_2 z + \text{error} \\ T &= \gamma_0 + \gamma_1 z + \text{error}, \end{aligned} \tag{10.5}$$

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

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

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

$$\begin{aligned} y &= \beta_0 + \beta_1 T + \beta_2 z + \text{error} \\ &= \beta_0 + \beta_1(\gamma_0 + \gamma_1 z) + \beta_2 z + \text{error} \\ &= (\beta_0 + \beta_1\gamma_0) + (\beta_1\gamma_1 + \beta_2)z + \text{error}. \end{aligned} \tag{10.6}$$

We now show how to estimate β_1, the causal effect of interest, using the slope of this regression, along with the regressions (10.5) and the exclusion restriction.

The first step is to express (10.6) in the form,

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

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

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

which we can rearrange to get,

$$\beta_1 = (\delta_1 - \beta_2)/\gamma_2. \tag{10.7}$$

We can directly estimate the denominator of this expression, γ_2, from the regression of T on z in (10.5)—this is not a problem since we are assuming that the instrument, z, is randomized.

The only challenge that remains in estimating β_1 from (10.7) is to estimate β_2, which in general cannot simply be estimated from the top equation of (10.5) since, as already noted, the error in that equation can be correlated with T. However, under the exclusion restriction, we know that β_2 is zero, and so $\beta_1 = \delta_1/\gamma_1$, leaving us with the standard instrumental variables estimate.

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

Two-stage least squares

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

To illustrate we return to our Sesame Street example. The first step is to regress

the "treatment" variable—an indicator for regular watching (`watched`)—on the randomized instrument, encouragement to watch (`encouraged`). Then we plug predicted values of `encouraged` into the equation predicting the letter recognition outcome, `y`:

R code

```
fit.2a <- lm (watched ~ encouraged)
watched.hat <- fit.2a$fitted
fit.2b <- lm (y ~ watched.hat)
```

The result is,

R output

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

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

This second-stage regression does not give the correct standard error, however. We discuss calculation of an appropriate standard error on page 223.

Adjusting for covariates in an instrumental variables framework

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

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

R code

```
fit.3a <- lm (watched ~ encouraged + pretest + as.factor(site) + setting)
watched.hat <- fit.3a$fitted
fit.3b <- lm (y ~ watched.hat + pretest + as.factor(site) + setting)
display (fit.3b)
```

yielding,

R output

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

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

Since the randomization took place within each combination of site (five categories) and setting (two categories), it would be appropriate to interact these variables in our equations. Moreover, it would probably be interesting to estimate variation of effects across sites and settings. However, for simplicity of illustration (and also due to the complication that one site × setting combination has no observations) we only include main effects for this discussion. We return to this example using multilevel models in Chapter 23. It turns out that the estimated average treatment effect changes only slightly (from 14.0 to 14.1) with the model that includes site × setting interactions.

Standard errors for instrumental variables estimates

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

The regression of compliance on treatment and other covariates (model `fit.3a`) is unchanged. We then regress the outcome on predicted compliance and covariance, this time saving the predictor matrix, X, from this second-stage regression (which we do using the `x=TRUE` option in the `lm` call):

```
fit.3b <- lm (y ~ watched.hat+pretest+as.factor(site)+setting, x=TRUE)
```
R code

We next compute the standard deviation of the adjusted residuals, $r_i^{\text{adj}} = y_i - X_i^{\text{adj}}\hat{\beta}$, where X^{adj} is the predictor matrix from `fit.3b` but with the column of predicted treatment values replaced by observed treatment values:

```
X.adj <- fit.2$x
X.adj[,"watched.hat"] <- watched
residual.sd.adj <- sd (y - X.adj %*% coef(fit.3b))
```
R code

Finally, we compute the adjusted standard error for the two-stage regression estimate by taking the standard error from `fit.3b` and scaling by the adjusted residual standard deviation, divided by the residual standard deviation from `fit.3b` itself:

```
se.adj <-se.coef(fit.3b)["watched.hat"]*residual.sd.adj/sigma.hat(fit.3b)
```
R code

So the adjusted standard errors are calculated as the square roots of the diagonal elements of $(X^tX)^{-1}\hat{\sigma}^2_{\text{TSLS}}$ rather than $(X^tX)^{-1}\hat{\sigma}^2$, where $\hat{\sigma}$ is the residual standard deviation from `fit.3b` and $\hat{\sigma}_{\text{TSLS}}$ is calculated using the residuals from an equation predicting the outcome from `watched` (not `watched.hat`) using the two-stage least squares estimate of the coefficient, not the coefficient that would have been obtained in a least squares regression of the outcome on `watched`).

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

Performing two-stage least squares automatically using the tsls function

We have illustrated the key concepts in our instrumental variables discussion using basic R commands with which you were already familiar so that the steps were transparent. There does exist, however, a package available in R called `sem` that has a function, `tsls()`, that automates this process, including calculating appropriate standard errors.

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

R code

```
iv1 <- tsls (postlet ~ regular, ~ encour, data=sesame)
display (iv1)
```

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

R output

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

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

R code

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

yielding,

R output

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

The point estimate of the treatment calculated this way is the same as with the preceding step-by-step procedure, but now we automatically get correct standard errors.

More than one treatment variable; more than one instrument

In the experiment discussed in Section 10.3, the children randomly assigned to the intervention group received several services ("treatments") that the children in the control group did not receive, most notably, access to high-quality child care and home visits from trained professionals. Children assigned to the intervention group did not make full use of these services. Simply conceptualized, some children participated in the child care while some did not, and some children received home visits while others did not. Can we use the randomization to treatment or control groups as an instrument for these two treatments? The answer is no.

Similar arguments as those used in Section 10.6 can be given to demonstrate that a single instrument cannot be used to identify more than one treatment variable. In fact, as a general rule, we need to use at least as many instruments as treatment variables in order for all the causal estimates to be identifiable.

Continuous treatment variables or instruments

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

Researchers must be careful, however, in the context of binary instruments and continuous treatment variables. A binary instrument cannot in general identify a continuous treatment or "dosage" effect (without further assumptions). If we map this back to a randomized experiment, the randomization assigns someone only to be encouraged or not. This encouragement may lead to different dosage levels, but for those in the intervention group these levels will be chosen by the subject (or subject's parents in this case). In essence this is equivalent to a setting with many different treatments (one at each dosage level) but only one instrument—therefore causal effects for all these treatments are not identifiable (without further assumptions). To identify such dosage effects, one would need to randomly assign encouragement levels that lead to the different dosages or levels of participation.

Have we really avoided the ignorability assumption? Natural experiments and instrumental variables

We have motivated instrumental variables using the cleanest setting, within a controlled, randomized experiment. The drawback of illustrating instrumental variables using this example is that it de-emphasizes one of the most important assumptions of the instrumental variables model, *ignorability of the instrument*. In the context of a randomized experiment, this assumption should be trivially satisfied (assuming the randomization was pristine). However, in practice an instrumental variables strategy potentially is more useful in the context of a *natural experiment*, that is, an observational study context in which a "randomized" variable (instrument) appears to have occurred naturally. Examples of this include:

- The draft lottery in the Vietnam War as an instrument for estimating the effect of military service on civilian health and earnings,
- The weather in New York as an instrument for estimating the effect of supply of fish on their price,
- The sex of a second child (in an analysis of people who have at least two children) as an instrument when estimating the effect of number of children on labor supply.

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

Assessing the plausibility of the instrumental variables assumptions

How can we assess the plausibility of the assumptions required for causal inference from instrumental variables? As a first step, the "first stage" model (the model that predicts the treatment using the instrument) should be examined closely to ensure both that the instrument is strong enough and that the sign of the coefficient makes sense. This is the only assumption that can be directly tested. If the association between the instrument and the treatment is weak, instrumental variables can yield incorrect estimates of the treatment effect even if all the other assumptions are satisfied. If the association is not in the expected direction, then closer examination

is required because this might be the result of a mixture of two different mechanisms, the expected process and one operating in the opposite direction which could in turn imply a violation of the monotonicity assumption.

Another consequence of a weak instrument is that it exacerbates the bias that can result from failure to satisfy the monotonicity and exclusion restrictions. For instance, for a binary treatment and instrument, when the exclusion restriction is not satisfied, our estimates will be off by a quantity that is equal to the effect of encouragement on the outcomes of noncompliers (in our example, never-watchers and always-watchers) multiplied by the ratio of noncompliers to compliers (in our example, induced watchers). The bias when monotonicity is not satisfied is slightly more complicated but also increases as the percentage of compliers decreases.

The two primary assumptions of instrumental variables (ignorability, exclusion) are not directly verifiable, but in some examples we can work to make them more plausible. For instance, if unconditional ignorability of the instrument is being assumed, yet there are differences in important pre-treatment characteristics across groups defined by the instrument, then these characteristics should be included in the model. This will not ensure that ignorability is satisfied, but it removes the *observed* problem with the ignorability assumption.

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

Structural equation models

A goal in many areas of social science is to infer causal relations among many variables, a generally difficult problem (as discussed in Section 9.8). *Structural equation modeling* is a family of methods of multivariate data analysis that are sometimes used for causal inference.[6] In that setting, structural equation modeling relies on conditional independence assumptions in order to identify causal effects, and the resulting inferences can be sensitive to strong parametric assumptions (for instance, linear relationships and multivariate normal errors). Instrumental variables can be considered to be a special case of a structural equation model. As we have just discussed, even in a relatively simple instrumental variables model, the assumptions needed to identify causal effects are difficult to satisfy and largely untestable. A structural equation model that tries to estimate many causal effects at once multiplies the number of assumptions required with each desired effect so that it

[6] Structural equation modeling is also used to estimate latent factors in noncausal regression settings with many inputs, and sometimes many outcome variables, which can be better understood by reducing to a smaller number of linear combinations.

quickly becomes difficult to justify all of them. Therefore we do not discuss the use of structural equation models for causal inference in any greater detail here. We certainly have no objection to complicated models, as will become clear in the rest of this book; however we are cautious about attempting to estimate complex causal structures from observational data.

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

Comparisons within groups—so-called "fixed effects" models

What happens when you want to make a causal inference but no valid instrument exists and ignorability does not seem plausible? Do alternative strategies exist? Sometimes repeated observations within groups or within individuals over time can provide a means for controlling for unobserved characteristics of these groups or individuals. If comparisons are made across the observations within a group or persons, implicitly such comparisons "hold constant" all characteristics intrinsic to the group or individual that do not vary across observations (across members of the group or across measures over time for the same person).

For example, suppose you want to examine the effect of low birth weight on children's mortality and other health outcomes. One difficulty in establishing a causal effect here is that children with low birth weight are also typically disadvantaged in genetic endowments and socioeconomic characteristics of the family, some of which may not be easy or possible to measure. Rather than trying to directly control for all of these characteristics, however, one could implicitly control for them by comparing outcomes across twins. Twins share many of the same genetic endowments (all if identical) and, in most cases, live in exactly the same household. However, there are physiological reasons (based, for instance, on position in the uterus) why one child in the pair may be born with a markedly different birth weight than the sibling. So we may be able to consider birth weight to be randomly assigned (ignorable) *within* twin pairs. Theoretically, if there is enough variation in birth weight, within sets of twins, we can estimate the effect of birth weight on subsequent outcomes. In essence each twin acts as a counterfactual for his or her sibling.

A regression model that is sometimes used to approximate this conceptual comparison simply adds an indicator variable for each of the groups to the standard regression model that might otherwise have been fit. So, for instance, in our twins example one might regress outcomes on birth weight (the "treatment" variable) and one indicator variable for each pair of twins (keeping one pair as a baseline category to avoid collinearity). More generally, we could control for the groups using a multilevel model, as we discuss in Part 2. In any case, the researcher might want to control for other covariates to improve the plausibility of the ignorability assumption (to control for the fact that the treatment may not be strictly randomly assigned even within each group—here, the pair of twins). In this particular example, however, it is difficult to find child-specific predictors that vary across children within a pair but can still be considered "pre-treatment."

In examples where the treatment is dichotomous, a substantial portion of the data may not exhibit any variation at all in "treatment assignment" within groups. For instance, if this strategy is used to estimate the effect of maternal employment on child outcomes by including indicators for each family (set of siblings) in the dataset, then in some families the mother may not have varied her employment status across children. Therefore, no inferences about the effect of maternal employment status can be made for these families. We can only estimate effects for the type of family

where the mother varied her employment choice across the children (for example, working after her first child was born but staying home from work after the second).

Conditioning on post-treatment outcomes. Still more care must be taken when considering variation over time. Consider examining the effect of marriage on men's earnings by looking at data that follows men over time and tracks marital status, earnings, and predictors of each (confounding covariates such as race, education, and occupation). Problems can easily arise in a model that includes an indicator for each person and also controls for covariates at each time point (to help satisfy ignorability). In this case the analysis would be implicitly conditioning on post-treatment variables, which, as we know from Section 9.8, can lead to bias.

Better suited for a multilevel model framework? This model with indicators for each group is often (particularly in the economics literature) called a "fixed effects" model. We dislike this terminology because it is interpreted differently in different settings, as discussed in Section 11.4. Further, this model is hierarchically structured, so from our perspective it is best analyzed using a multilevel model. This is not completely straightforward, however, because one of the key assumptions of a simple multilevel model is that the individual-level effects are independent of the other predictors in the model—a condition that is particularly problematic in this setting where we are expecting that unobserved characteristics of the individuals may be associated with observed characteristics of the individuals. In Chapter 23 we discuss how to appropriately extend this model to the multilevel framework while relaxing this assumption.

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

Almost all causal strategies make use of comparisons across groups: one or more that were exposed to a treatment, and one or more that were not. *Difference-in-difference* strategies additionally make use of another source of variation in outcomes, typically time, to help control for potential (observed and unobserved) differences across these groups. For example, consider estimating the effect of a newly-introduced school busing program on housing prices in a school district where some neighborhoods were affected by the program and others were not. A simple comparison of housing prices across affected and unaffected areas some time after the busing program went into effect might not be appropriate because these neighborhoods might be different in other ways that might be related to housing prices. A simple before-after comparison of housing prices may also be inappropriate if other changes that occurred during this time period (for example, a recession) might also be influencing housing prices. A difference-in-differences approach would instead calculate the difference in the before-after *change* in housing prices in exposed and unexposed neighborhoods. An important advantage of this strategy is that the units of observation (in this case, houses) need not be the same across the two time periods.

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

Panel data. A special case of difference-in-differences estimation occurs when the same set of units are observed at both time points. This is also a special case of the so-called fixed effects model that includes indicators for treatment groups and for time periods. A simple way to fit this model is with a regression of the outcome on an indicator for the groups, an indicator for the time period, and the interaction between the two. The coefficient on the interaction is the estimated treatment effect.

In this setting, however, the advantages of the difference-in-differences strategy are less apparent because an alternative model would be to include an indicator for treatment exposure but then simply regress on the pre-treatment version of the outcome variable. In this framework it is unclear if the assumption of randomly assigned *changes* in potential outcome is truly weaker than the assumption of randomly assigned potential outcomes for those with the same value of the pre-treatment variable.[7]

Do not condition on post-treatment outcomes. Once again, to make the (new) ignorability assumption more plausible it may be desirable to condition on additional predictor variables. For models where the variation takes place over time—for instance, the differences-in-differences estimate that includes both pre-treatment and post-treatment observations on the same units—a standard approach is to include changes in characteristics for each observation over time. Implicitly, however, this conditions on post-treatment variables. If these predictors can be reasonably assumed to be unchanged by the treatment, then this is reasonable. However, as discussed in Section 9.8, it is otherwise inappropriate to control for post-treatment variables. A better strategy would be to control for pre-treatment variables only.

10.8 Bibliographic note

We have more references here than for any of the other chapters in this book because causal inference is a particularly contentious and active research area, with methods and applications being pursued in many fields, including statistics, economics, public policy, and medicine.

Imbalance and lack of complete overlap have been discussed in many places; see, for example, Cochran and Rubin (1973), and King and Zeng (2006). The intervention for low-birth-weight children is described by Brooks-Gunn, Liaw, and Klebanov (1992) and Hill, Brooks-Gunn, and Waldfogel (2003). Imbalance plots such as Figure 10.3 are commonly used; see Hansen (2004), for example.

Subclassification and its connection to regression are discussed by Cochran (1968). Imbens and Angrist (1994) introduce the local average treatment effect. Cochran and Rubin (1973), Rubin (1973), Rubin (1979), Rubin and Thomas (2000), and Rubin (2006) discuss the use of matching, followed by regression, for causal inference. Dehejia (2003) discusses an example of the interpretation of a treatment effect with interactions.

Propensity scores were introduced by Rosenbaum and Rubin (1983a, 1984, 1985). A discussion of common current usage is provided by D'Agostino (1998). Examples across several fields include Lavori, Keller, and Endicott (1995), Lechner (1999), Hill, Waldfogel, and Brooks-Gunn (2002), Vikram et al. (2003), and O'Keefe (2004). Rosenbaum (1989) and Hansen (2004) discuss full matching. Diamond and Sekhon (2005) present a genetic matching algorithm. Drake (1993) discusses robustness of treatment effect estimates to misspecification of the propensity score model. Joffe

[7] Strictly speaking, we need not assume actual random manipulation of treatment assignment for either assumption to hold, only results that would be consistent with such manipulation.

and Rosenbaum (1999), Imbens (2000), and Imai and van Dyk (2004) generalize the propensity score beyond binary treatments. Rubin and Stuart (2005) extend to matching with multiple control groups. Imbens (2004) provides a recent review of methods for estimating causal effects assuming ignorability using matching and other approaches.

Use of propensity scores as weights is discussed by Rosenbaum (1987), Ichimura and Linton (2001), Hirano, Imbens, and Ridder (2003), and Frolich (2004) among others. This work has been extended to a "doubly-robust" framework by Robins and Rotnitzky (1995), Robins, Rotnitzsky, and Zhao (1995), and Robins and Ritov (1997).

As far as we are aware, Lalonde (1986) was the first use of so-called constructed observational studies as a testing ground for non-experimental methods. Other examples include Friedlander and Robins (1995), Heckman, Ichimura, and Todd (1997), Dehejia and Wahba (1999), Michalopoulos, Bloom, and Hill (2004), and Agodini and Dynarski (2004). Dehejia (2005a, b), in response to Smith and Todd (2005)) provides useful guidance regarding appropriate uses of propensity scores (the need to think hard about ignorability and to specify propensity score models that are specific to any given dataset). The constructed observational analysis presented in this chapter is based on a more complete analysis presented in Hill, Reiter, and Zanutto (2004).

Interval estimation for treatment effect estimates obtained via propensity score matching is discussed in Hill and Reiter (2006). Du (1998) and Tu and Zhou (2003) discuss intervals for estimates obtained via propensity score subclassification. Hill and McCulloch (2006) present a Bayesian nonparametric method for matching.

Several packages exist that automate different combinations of the propensity score steps described here and are available as supplements to R and other statistical software. We mention some of these here without intending to provide a comprehensive list. There is a program available for R called `MatchIt` that is available at `gking.harvard.edu/matchit/` that implements several different matching methods including full matching (using software called `OptMatch`; Hansen, 2006). Three packages available for Stata are `psmatch2`, `pscore`, and `nnmatch`; any of these can be installed easily using the "net search" (or comparable) feature in Stata. Additionally, `nnmatch` produces valid standard errors for matching. Code is also available in SAS for propensity score matching or subclassification; see, for example, `www.rx.uga.edu/main/home/cas/faculty/propensity.pdf`.

Regression discontinuity analysis is described by Thistlethwaite and Campbell (1960). Recent work in econometrics includes Hahn, Todd, and van der Klaauw (2001) and Linden (2006). The political ideology example in Section 10.4 is derived from Poole and Rosenthal (1997) and Gelman and Katz (2005); see also Lee, Moretti, and Butler (2004) for related work. The example regarding children's promotion in school was drawn from work by Jacob and Lefgren (2004).

Instrumental variables formulations date back to work in the economics literature by Tinbergen (1930) and Haavelmo (1943). Angrist and Krueger (2001) present an upbeat applied review of instrumental variables. Imbens (2004) provides a review of statistical methods for causal inference that is a little less enthusiastic about instrumental variables. Woolridge (2001, chapter 5) provides a crisp overview of instrumental variables from a classical econometric perspective; Lancaster (2004, chapter 8) uses a Bayesian framework. The "always-watcher," "induced watcher," and "never-watcher" categorizations here are alterations of the "never-taker," "complier," and "always-taker" terminology first used by Angrist, Imbens, and Rubin (1996), who reframe the classic econometric presentation of instrumental variables

in statistical language and clarify the assumptions and the implications when the assumptions are not satisfied. For a discussion of all of the methods discussed in this chapter from an econometric standpoint, see Angrist and Krueger (1999).

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

For models that link instrumental variables with the potential-outcomes framework described in Chapter 9, see Angrist, Imbens, and Rubin (1996). Glickman and Normand (2000) derive an instrumental variables estimate using a latent-data model; see also Carroll et al. (2004).

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

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

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

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

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

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

10.9 Exercises

1. Constructed observational studies: the folder `lalonde` contains data from an observational study constructed by LaLonde (1986) based on a randomized experiment that evaluated the effect on earnings of a job training program called National Supported Work. The constructed observational study was formed by replacing the randomized control group with a comparison group formed using data from two national public-use surveys: the Current Population Survey (CPS) and the Panel Study in Income Dynamics.

   Dehejia and Wahba (1999) used a subsample of these data to evaluate the potential efficacy of propensity score matching. The subsample they chose removes men for whom only one pre-treatment measure of earnings is observed. (There is

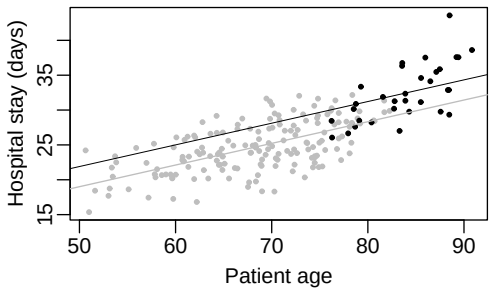

Figure 10.10 *Hypothetical data of length of hospital stay and age of patients, with separate points and regression lines plotted for each treatment condition: the new procedure in gray and the old procedure in black.*

substantial evidence in the economics literature that controlling for earnings from only one pre-treatment period is insufficient to satisfy ignorability.) This exercise replicates some of Dehejia and Wahba's findings based on the CPS comparison group.

(a) Estimate the treatment effect from the experimental data in two ways: (i) a simple difference in means between treated and control units, and (ii) a regression-adjusted estimate (that is, a regression of outcomes on the treatment indicator as well as predictors corresponding to the pre-treatment characteristics measured in the study).

(b) Now use a regression analysis to estimate the causal effect from Dehejia and Wahba's subset of the constructed observational study. Examine the sensitivity of the model to model specification (for instance, by excluding the employed indicator variables or by including interactions). How close are these estimates to the experimental benchmark?

(c) Now estimate the causal effect from the Dehejia and Wahba subset using propensity score matching. Do this by first trying several different specifications for the propensity score model and choosing the one that you judge to yield the best balance on the most important covariates.

Perform this propensity score modeling *without* looking at the estimated treatment effect that would arise from each of the resulting matching procedures.

For the matched dataset you construct using your preferred model, report the estimated treatment effects using the difference-in-means and regression-adjusted methods described in part (a) of this exercise. How close are these estimates to the experimental benchmark (about $1800)?

(d) Assuming that the estimates from (b) and (c) can be interpreted causally, what causal effect does each estimate? (Hint: what populations are we making inferences about for each of these estimates?)

(e) Redo both the regression and the matching exercises, excluding the variable for earnings in 1974 (two time periods before the start of this study). How important does the earnings-in-1974 variable appear to be in terms of satisfying the ignorability assumption?

2. Regression discontinuity analysis: suppose you are trying to evaluate the effect of a new procedure for coronary bypass surgery that is supposed to help with the postoperative healing process. The new procedure is risky, however, and is rarely performed in patients who are over 80 years old. Data from this (hypothetical) example are displayed in Figure 10.10.

(a) Does this seem like an appropriate setting in which to implement a regression discontinuity analysis?

(b) The folder `bypass` contains data for this example: `stay` is the length of hospital stay after surgery, `age` is the age of the patient, and `new` is the indicator variable indicating that the new surgical procedure was used. Preoperative disease severity (`severity`) was unobserved by the researchers, but we have access to it for illustrative purposes. Can you find any evidence using these data that the regression discontinuity design is inappropriate?

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

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

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

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

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

4. In Exercise 9.13, you estimated the effect of incumbency on votes for Congress. Now consider an additional variable: money raised by the congressional candidates. Assume this variable has been coded in some reasonable way to be positive in districts where the Democrat has raised more money and negative in districts where the Republican has raised more.

(a) Explain why it is inappropriate to include money as an additional input variable to "improve" the estimate of incumbency advantage in the regression in Exercise 9.13.

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

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

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

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