Advanced statistical methods

Causal inference

Leonardo Egidi

March 2024

Università di Trieste

Table of contents i

Basic principles of causal inference

Why causal inference?

- So far, we have been interpreting regressions *predictively*: given the values of several inputs, the fitted model allows us to predict y, typically considering the n data points as a simple random sample from a hypothetical infinite "superpopulation" or probability distribution.
- This part of the course considers causal inference, which concerns what would happen to an outcome y as a result of a *treatment*, *intervention*, or *exposure* z, given pre-treatment information x.

Causal inference: basic concepts

- *Causal effects* are conceptualized as a comparison between different potential outcomes of what might have occurred under different scenarios.
- This comparison could be between a factual state (what did happen) and one or more counterfactual states (representing what might have happened), or it could be a comparison among various counterfactuals.
- We will study causal effects through regression and predictors. In a regression framework, the treatment will be:

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

or, for a continuous treatment:

$$z_i = \text{level of the “treatment" assigned to unit } i. \tag{2}$$

Randomized experiments and observational studies

We will study the basics of causal inference in two distinct scenarios:

- *randomized experiments*: experiments with units randomly assigned to receive treatment and control, and with the units in the study considered as a random sample from a population of interest. The random sampling and random treatment assignment-sometimes performed according to *blocks* or *strata*-allow us to estimate the *average causal effect* of the treatment in the population, and regression modeling can be used to refine this estimate.
- *observational studies*: in observational studies treatments are observed rather than assigned (for example, comparisons of smokers to non-smokers), and it is not at all reasonable to consider the observed data under different treatments as random samples from a common population. In these studies there can be systematic differences between groups of units that receive different treatments—differences that are outside the control of the experimenter. Often, however, observational studies refer more broadly to survey data settings where no intervention has been performed.

The fundamental problem of causal inference i

- We begin by considering the problem of estimating the causal effect of a treatment compared to a control, for example in a medical experiment. Formally, the causal effect of a treatment z on an outcome y for an observational or experimental unit i can be defined by comparisons between the outcomes that would have occurred under each of the different treatment possibilities.
- With a binary treatment z taking on the value 0 (control) or 1 (treatment), we can define the *potential outcomes*, y_i^0 and y_i^1 for unit i as the outcomes that would be observed under control and treatment conditions, respectively. (These ideas can also be directly generalized to the case of a treatment variable with multiple levels.)
- For someone assigned to the treatment condition ($z_i = 1$), y_i^1 is observed and y_i^0 is the unobserved counterfactual outcome—it represents what would have happened to the individual if assigned to control. Conversely, for control units, y_i^0 is observed and y_i^1 is counterfactual.
- The *observed outcome* for the unit i is then $y_i = y_i^0(1 - z_i) + y_i^1 z_i$.

The fundamental problem of causal inference ii

- In either case, a simple treatment effect for unit i can be defined as:

$$\text{individual treatment effect: } \tau_i = y_i^1 - y_i^0. \qquad (3)$$

  Causal effects can also be expressed as nonlinear functions of the potential outcomes, but linear functions are conceptually simpler.

- The problem inherent in determining the effect for any given individual, however, is that we can never observe both potential outcomes y_i^0 **and** y_i^1. This is commonly referred to as the *fundamental problem of causal inference*.

The fundamental problem of causal inference iii

- Running example: studying the relationship between fish oil supplements and systolic blood pressure. 8 friends: 4 friends were placed in the "fish oil supplement" treatment group. Members of this group agreed to consume 3 grams of fish oil supplements per day for one year while otherwise maintaining their current diets. The other 4 friends agreed to simply maintain their current diets free from fish oil supplements for the same year. At the end of the study period, blood pressure was measured for each of the eight participants (see next Figure 18.1).
  - y_i^0: blood pressure that would result if the person had no supplement,
  - y_i^1: blood pressure that would result if the person had received the prescribed supplement.

The fundamental problem of causal inference iv

- We cannot observe the blood pressure that would have resulted both if, say, Audrey had taken the supplements and if she had not, thus the *causal effect is impossible to directly measure*!

Unit i	Female, x_{1i}	Age, x_{2i}	Treatment, z_i	Potential outcomes if $z_i = 0$, y_i^0	Potential outcomes if $z_i = 1$, y_i^1	Observed outcome, y_i
Audrey	1	40	0	140	?	140
Anna	1	40	0	140	?	140
Bob	0	50	0	150	?	150
Bill	0	50	0	150	?	150
Caitlin	1	60	1	?	155	155
Cara	1	60	1	?	155	155
Dave	0	70	1	?	160	160
Doug	0	70	1	?	160	160

Figure 18.1 *Hypothetical causal inference data for the effect of fish oil supplements on systolic blood pressure. For each person, we as researchers can only observe the potential outcome corresponding to the treatment actually received. Therefore we cannot directly observe either the individual-level or group-level causal effects. In this example, a simple difference of average outcomes across the two groups,* $157.5 - 145$*, would lead to the estimate that supplements produce a 12.5 mmHg increase in systolic blood pressure. This is a poor estimate because the treatment and control groups here are highly imbalanced.*

Ways of getting around the problem i

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

Ways of getting around the problem ii

- ***Close substitutes***: one might object to the formulation of the fundamental problem of causal inference by noting situations where it appears one can actually measure both y_i^0 and y_i^1 on the same unit—consider, for example drinking tea one evening and milk another evening, and then measuring the amount of sleep each time.
- ***Statistical adjustment***: when treatment and control groups are not similar, modeling or other forms of statistical adjustment can be used to fill in the gap.
- ***Randomization and experimentation***: use the outcomes observed on a sample of units to learn about the distribution of outcomes in the population. The basic idea is that since we cannot compare treatment and control outcomes for the same units, we try to compare them on similar units. Similarity can be attained by using randomization to decide which units are assigned to the treatment group and which units are assigned to the control group. $\Rightarrow$ *Cleanest solution!*

Randomized experiments

Estimate the average treatment effect i

- Although we cannot estimate individual-level causal effects (3) (without making strong assumptions, as discussed previously), we can design studies to estimate the population average treatment effect:

$$\text{average treatment effect} = \text{avg}(y_i^1 - y_i^0) \qquad (4)$$

- The cleanest way to estimate the population average is through a randomized experiment in which *each unit has a positive chance of receiving each of the possible treatments*.
- For example, if n_0 units are selected at random from the population and given the control, and n_1 other units are randomly selected and given the treatment, then the observed sample averages of y for the treated and control units can be used to estimate the corresponding population quantities, $\text{avg}(y^0)$ and $\text{avg}(y^1)$, with their difference estimating the average treatment effect in (4) (and with standard error $\sqrt{s_0^2/n_0 + s_1^2/n_1}$).
- This works because the y_i^0's for the control group are a random sample of the values of y_i^0 in the entire population. Similarly, the y_i^1's for the treatment group are a random sample of the y_i^1's in the population.

Estimate the average treatment effect ii

- The starting point is the comparison of treatment and control groups, and we can run into trouble if these two groups are not sufficiently similar or balanced.
- In Figure 18.1, we see that the people who received treatment were on average older than the controls. This difference could have occurred just by chance or perhaps because those who agreed to take the supplements were more concerned about their blood pressure and the study offered them a chance to try out the supplements for free, while those who agreed to be in the no-supplements group did not care if the supplements might benefit their health.
- The groups whose outcomes we were comparing differed in their pre-treatment characteristics. This difference matters because age is also predictive of the outcome.
- In practice, we can never ensure that treatment and control groups are balanced on all relevant pre-treatment characteristics. However, there are statistical approaches that may bring us closer. At the design stage, we can use randomization to ensure that treatment and control groups are balanced in expectation, and we can use blocking to reduce the variation in any imbalance. At the analysis stage, we can adjust for pre-treatment variables to correct for differences between the two groups to reduce bias in our estimate of the sample average treatment effect.

The assumption of ignorability in controlled experiments

- In a *completely randomized design*, the probability of being assigned to the treatment is the same for each unit in the sample.
- Moreover, the treatment assignment is a random variable that is independent of the potential outcomes, a statement that can be written formally as,

$$y^0, y^1 \perp z. \tag{5}$$

  As a consequence under repeated randomizations, there will be no differences, on average, in the potential outcomes, comparing treatment and control groups. This property is commonly referred to as *ignorability*.
- Ignorability does not imply that the groups are perfectly balanced. Rather, *it implies that there is no imbalance on average across repeated randomizations.*
- Said another way, ignorability implies that *the value of someone's potential outcomes does not provide any information about his or her treatment group assignment.*

Motivations for blocks

- Sometimes a randomized experiment could benefit from a preliminary blocking structure to which apply randomization.
- The goal when creating *blocks* is to minimize the variation of each type of potential outcome, y^0 and y^1, within the block. In practice researchers only have access to observed pre-treatment variables when making decisions regarding how to define blocks (defined by age, for instance, or even as age$\times$sex).
- So, to the extent possible, *the predictors used to define the blocks should be those that are believed to be predictive of the outcom*e based on either theory or on results from previous studies. The more predictive the blocking variable, the bigger the precision gains at the end of the day.

The regression framework of causal inference i

- In the usual regression context, predictive inference relates to comparisons *between* units, whereas causal inference addresses comparisons of different treatments if applied to the *same* units. More generally, causal inference can be viewed as a special case of prediction in which the goal is to predict *what would have happened* under different treatment options.
- We illustrate the use of regression in the setting of controlled experiments, going through issues of adjustment for pre-treatment predictors, interactions, and pitfalls that can arise when building a regression using experimental data and interpreting coefficients causally.

The regression framework of causal inference ii

For each item i we can have at least three sorts of measurements (see also next Figure 19.1):

- Pre-treatment measurements, also called *covariates*, x_i . As noted above, these are not strictly required for causal inference but in practice can be essential for checking and adjusting for pre-treatment differences between treatment and control groups, estimating treatment interactions, and making inferences about average treatment effects in subpopulations and in the general population.
- The treatment z_i , which equals 1 for treated units and 0 for controls.
- The outcome measurement, y_i , which we label as y_i^1 for units that have been exposed to the treatment and y_i^0 for units that received the control.

The regression framework of causal inference iii

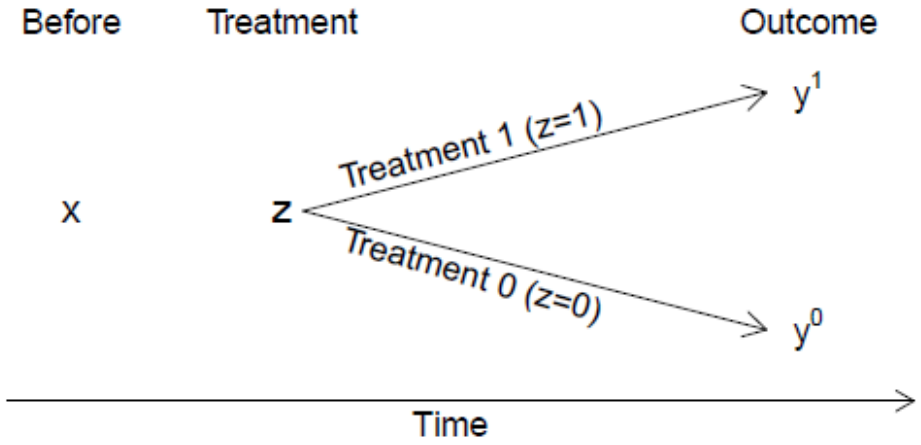

Figure 19.1 *Basic framework for causal inference. The treatment effect is* $y^1 - y^0$*, but we can never observe both* potential outcomes y^1 *and* y^0 *on the same item. For any given unit, the unobserved outcome is called the* counterfactual.

The Electric Company data example

Electric Company Data (ROS book, 19.2; G&H book, 9.3)

The goal is to measure the causal effect of a new educational television program, *The Electric Company*, on children's reading ability. Selected classes of children in grades 1–4 were *randomized into treated and control groups*. At the beginning and the end of the school year, students in all the classes are given a reading test, and the average test score within each class is recorded: our entire analysis will be then at the classroom level, that is, we treat the classes as the observational units in this study.

The experiment was performed around 1970 on a set of 192 elementary school classes in two cities, Fresno and Youngstown. For each city and grade, the experimenters selected a small number of schools (10–20) and, within each school, they selected the two poorest reading classes of that grade. For each pair, one of these classes was randomly assigned to continue with its regular reading course and the other was assigned to view the TV program.

This is an example of a *matched pairs design*: there are characteristics of the school—both observable and, potentially, unobservable—that are predictive of future student outcomes that we would like to adjust for explicitly by forcing balance through our design. For simplicity we shall analyze this experiment *as if the treatment assignment had been completely randomized within each grade*.

Plotting the average post-treatment outcome i

- y_i^0, y_i^1: outcomes at class level.
- $z_i = 1$ if the class undertook the television program (the treatment, here).
- x_i: pre-test score given for each class at the beginning of the school, before the treatment was applied.
- In the next Figures, we plot the distribution of the outcome, the average post-treatments scores, in the control and treatment group for each grade (Figure 19.2), and the same information but arranged in a different orientation to allow easier comparison between treated and control groups in each grade (Figure 19.3; for each histogram, the average is indicated by a vertical line).
- Visual comparisons across treatment groups within grades suggest that watching *The Electric Company* may have led to small increases in average test scores, particularly for the lower grades.

Plotting the average post-treatment outcome ii

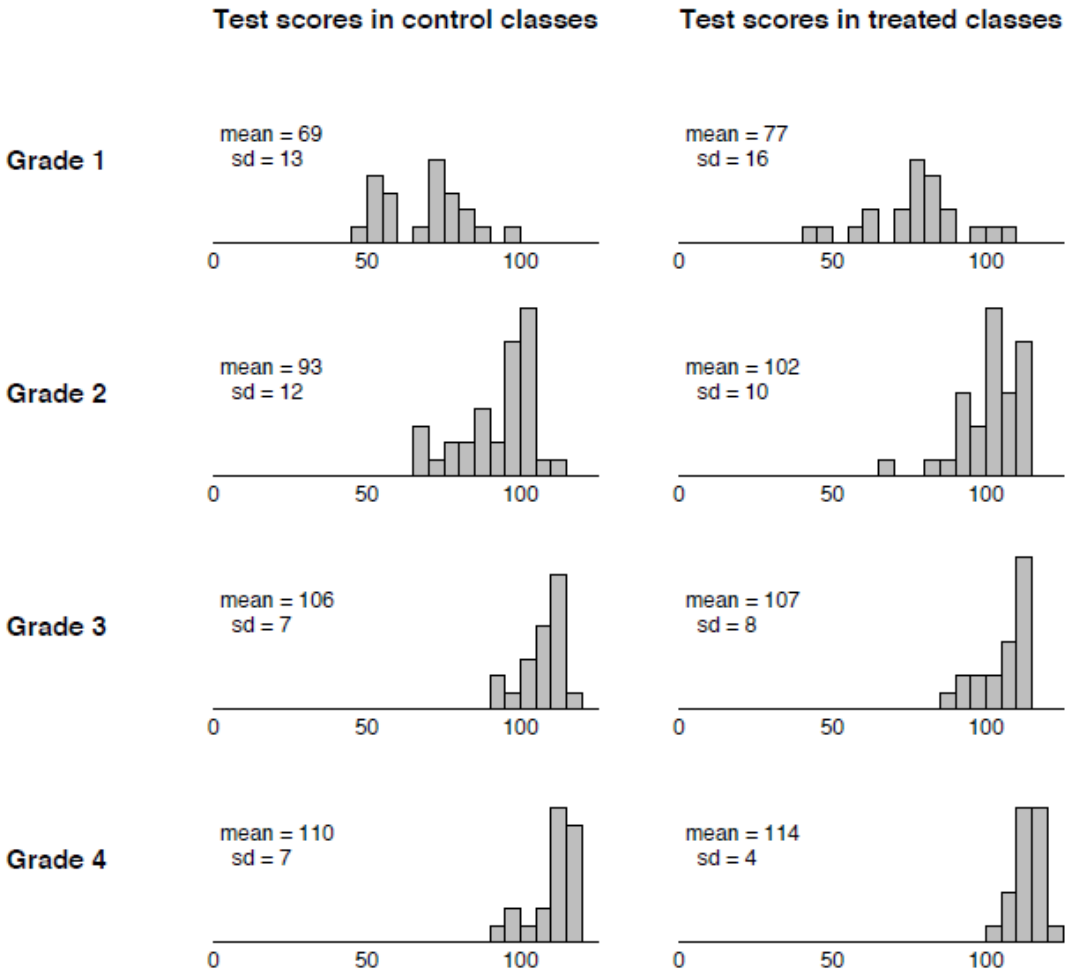

Figure 19.2 *Post-treatment test scores from an experiment measuring the effect of an educational television program, The Electric Company, on children's reading abilities. The experiment was applied on a total of 192 classrooms in four grades. At the end of the experiment, the average reading test score in each classroom was recorded.*

Plotting the average post-treatment outcome iii

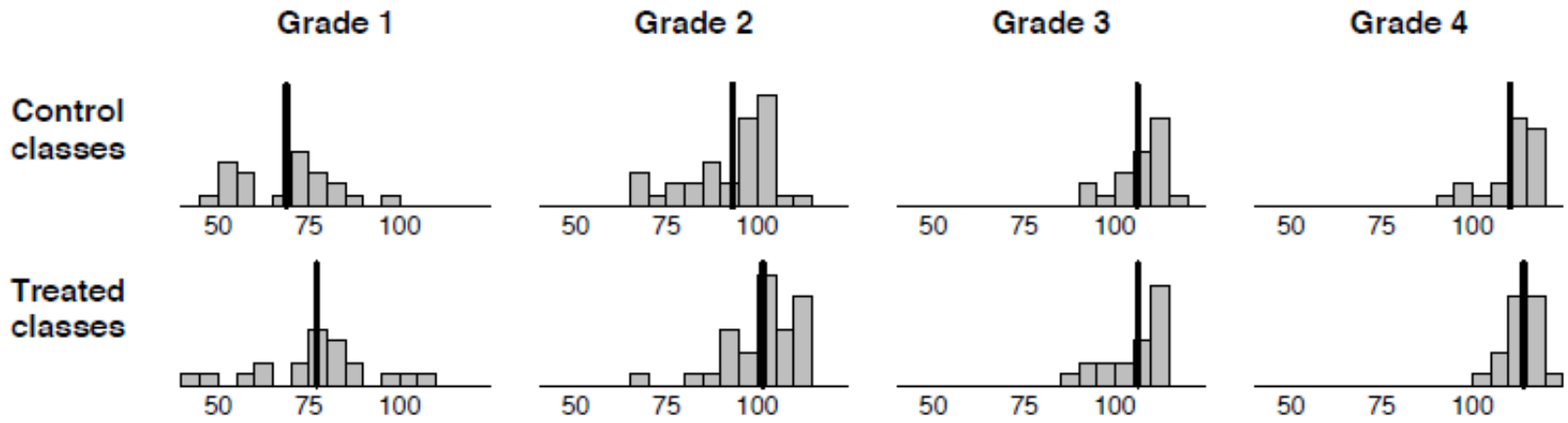

Figure 19.3 *Data from the Electric Company experiment, from Figure 19.2, displayed in a different orientation to allow easier comparison between treated and control groups in each grade. For each histogram, the average is indicated by a vertical line.*

A first regression model

- We start by estimating a single treatment effect using the simplest possible estimate, a *linear regression of post-test on treatment indicator*, which would be the appropriate analysis had the data come from a completely randomized experiment with no available pre-treatment information.
- When treatments are assigned completely at random, we can think of the treatment and control groups as two random samples from a common population. The population average under each treatment, $\text{avg}(y^0)$ and $\text{avg}(y^1)$, can then be estimated by the sample average, and the population average difference between treatment and control, $\text{avg}(y^1) - \text{avg}(y^0)$—that is, the average causal effect—can be estimated by the difference in sample averages, $\bar{y}_1 - \bar{y}_0$.
- Equivalently:

$$y_i = \alpha + \theta z_i + \text{error}_i, \tag{6}$$

where θ corresponds to the average causal effect of the treatment. In fact, linear regression on an indicator variable is a comparison of averages (*try as an exercise!*).

Separate analysis within each grade i

- Given the large variation in test scores from grade to grade, it makes sense to take the next step and perform a separate regression analysis on each grade's data. This is *equivalent to fitting a model in which treatment effects vary by grade*—that is, an interaction between treatment and grade indicators—and where the residual variance can be different from grade to grade as well.
- The resulting estimates and uncertainty intervals for model (6) are given in Figure 19.4(a). The treatment appears to be generally effective, perhaps more so in the low grades, but it is hard to be sure, given the large standard errors of estimation. Sample sizes are approximately the same in each of the grades, but the estimates for higher grades have lower standard errors because the residual standard deviations of the regressions are lower in these grades.

Separate analysis within each grade ii

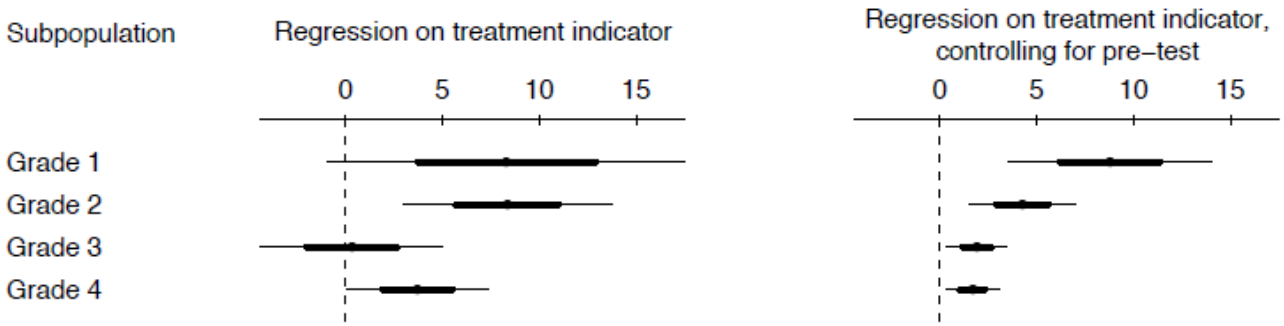

Figure 19.4 *Estimates, 50%, and 95% intervals for the effect of watching The Electric Company (see data in Figures 19.2 and 19.5) as estimated in two ways: (a) from a regression on treatment alone, and second,(b) also adjusting for pre-test data. In both cases, the coefficient for treatment is the estimated causal effect. Including pre-test data as a predictor increases the precision of the estimates.*
Displaying these coefficients and intervals as a graph facilitates comparisons across grades and across estimation strategies (adjusting for pre-test or not). For instance, the plot highlights how adjusting for pre-test scores increases precision and reveals decreasing effects of the program for the higher grades, a pattern that would be more difficult to see in a table of numbers.
Sample sizes are approximately the same in each of the grades. The estimates for higher grades have lower standard errors because the residual standard deviations of the regressions are lower in these grades; see Figure 19.5.

Adjusting for pre-test i

- We can use the information about pre-test to improve our treatment effect estimates, using a regression model such as:

$$y_i = \alpha + \theta z_i + \beta x_i + \text{error}_i, \tag{7}$$

where x_i denotes the average pre-test scores of the students in classrom i. The coefficient θ still represents the average treatment effect in the grade, but *adjusting for pre-treatment score, x_i, can reduce the uncertainty in the estimate*-see Figure 19.4(b).

- Figure 19.5 (next slide) shows the before-after data for the Electric Company experiment. For grades 2–4, the same test was given to the students at the beginning and the end of the year, and so it is no surprise that all the classes improved whether treated or not. For grade 1, the pre-test was a subset of the longer test, which explains why the pre-test scores for grade 1 are so low. We can also see that the distribution of post-test scores for each grade is similar to the next grade's pre-test scores, which makes sense.

Adjusting for pre-test ii

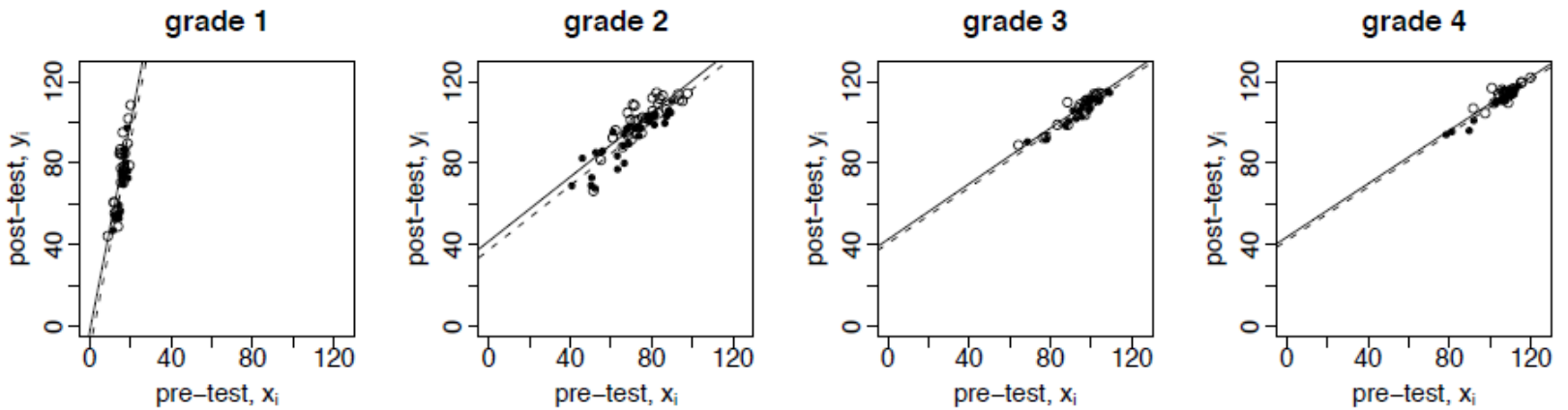

Figure 19.5 *Pre-test/post-test data for the Electric Company experiment. Treated and control classes are indicated by circles and dots, respectively, and the solid and dotted lines represent parallel regression lines fit to the treatment and control groups, respectively. The solid lines are slightly higher than the dotted lines, indicating slightly positive estimated treatment effects. Compare to Figure 19.2, which displays only the post-test data.*

Adjusting for pre-test iii

There are some benefits of adjusing for pre-treatment scores:

- If the predictor has a strong association with the outcome it can help to bring each estimate closer (on average) to the truth, and if the randomization was less than pristine, the addition of predictors to the equation may help us adjust for systematically unbalanced characteristics across groups: potential to adjust for both random and systematic differences between the treatment and control groups-see Figure 19.4(b).
- This reasoning applies not just to pre-test but to any pre-treatment variables that help to predict the outcome. In practice, we can neither collect nor analyze everything, and *our models (linear and otherwise) are themselves only approximations*, so we can only try to reduce our errors of estimation, not bring them all the way to zero.
- Crucially, when fitting such a regression *it is only appropriate to adjust for pre-treatment predictors*, or, more generally, predictors that would not be affected by the treatment.

Treatment interactions and poststratification i

- Once we include pre-test in the model, it is natural to *interact* it with the treatment effect. The treatment effect is then allowed to vary with the level of the pre-test. Then:

$$y_i = \alpha + \theta z_i + \beta x_i + \tau z_i x_i + \text{error}_i, \tag{8}$$

  where the treatment effect is $\theta + \tau x$, and the summary treatment effect in the sample is $\frac{1}{n}\sum_{i=1}^{n}(\theta + \tau x_i)$.

- Figure 19.6 shows the Electric Company data with separate intercepts and slopes estimated for the treatment and control groups. As with Figure 19.5, for each grade the difference between the regression lines is the estimated treatment effect as a function of pre-test score.

Treatment interactions and poststratification ii

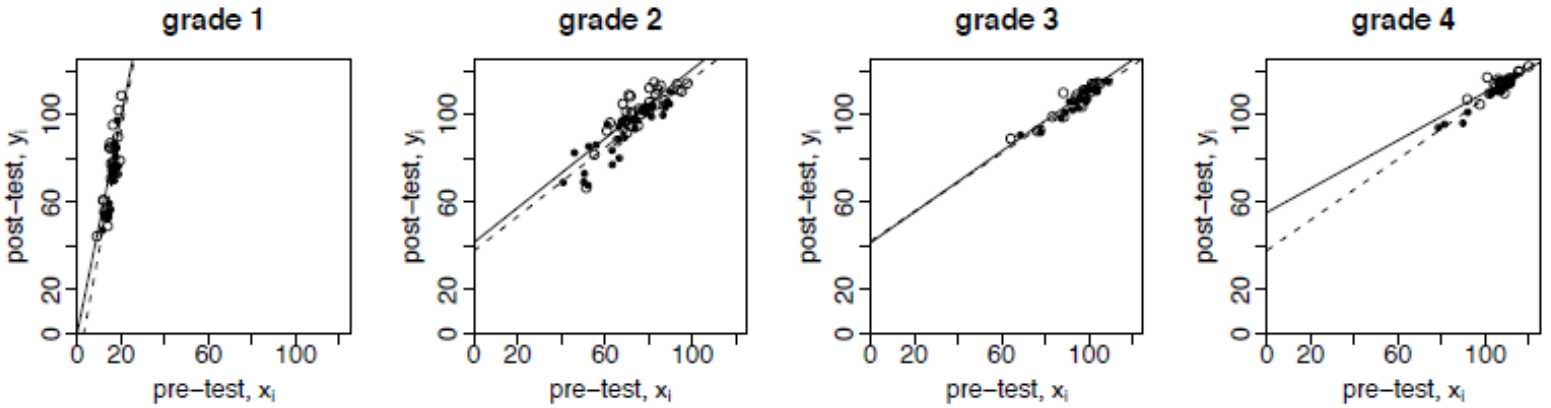

Figure 19.6 *Pre-test/post-test data for the Electric Company experiment. Treated and control classes are indicated by circles and dots, respectively, and the solid and dotted lines represent separate regression lines fit to the treatment and control groups, respectively—that is, the model interacts treatment with pre-test score. For each grade, the difference between the solid and dotted lines represents the estimated treatment effect as a function of pre-test score. Compare to Figure 19.5, which displays the same data but with parallel regression lines in each graph. The non-parallel lines in the current figure represent interactions in the fitted model.*

Treatment interactions and poststratification iii

- For frther illustration, let's focus on grade 4 classes. We fit the model (8) with the following instruction:

```
stan_glm(post_test ~ treatment + pre_test +
                     treatment:pre_test,
                     data=electric, subset=(grade==4))
```

yielding

```
                    Median MAD_SD
(Intercept)          39.41   4.90
treatment            11.61   7.98
pre_test              0.68   0.05
treatment:pre_test   -0.09   0.07
Auxiliary parameter(s):
      Median MAD_SD
sigma 2.17   0.25
```

- The estimated treatment effect is now $11.61 - 0.09x$. Centering x before including it in the model allows the treatment coefficient to represent that treatment effect for classes with the mean pre-test score for the sample.

Treatment interactions and poststratification iv

- To get a sense of the uncertainty, we can plot the 20 random simulation draws of the estimated treatment effect, as displayed in next Figure 19.7.

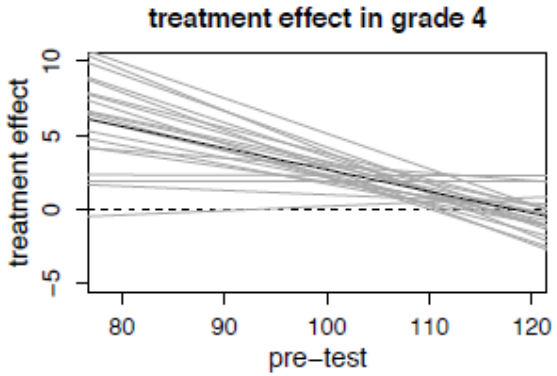

Figure 19.7 *Uncertainty about the effect of viewing The Electric Company (compared to the control) for fourth-graders. Compare to the rightmost plot in Figure 19.6. The dark line here—the estimated treatment effect as a function of pre-test score—is the difference between the two regression lines in the grade 4 plot in Figure 19.6. The gray lines represent 20 random draws from the uncertainty distribution of the treatment effect.*

Treatment interactions and poststratification v

- We can compute the estimated average treatment across the model's simulations to represent the uncertainty in the average treatment effect previously introduced. The result is 1.8 with a standard deviation of 0.7—similar to the result from the model adjusting for pre-test but with no interactions (*try as exercise to produce the results!*).
- In general, for a linear regression model, the estimate obtained by including the interaction, and then averaging over the data, reduces to the estimate with no interaction. The motivation for including the interaction is thus to get a better idea of how the treatment effect varies with pre-treatment predictors, not to simply estimate an average effect.
- Identification of treatment interactions is also important when we want to generalize experimental results to a broader population. If treatment effects vary with pre-treatment characteristics and the distribution of these characteristics varies between the experimental sample and the population of interest, the average treatment effects will typically be different.

Treatment interactions and poststratification vi

- In survey sampling, *stratification* refers to the procedure of dividing the population into disjoint subsets (strata), sampling separately within each stratum, and then combining the stratum samples to get a population estimate.
- *Poststratification* is the analysis of an unstratified sample, breaking the data into strata and reweighting as would have been done had the survey actually been stratified.
- Stratification can adjust for potential differences between sample and population using the survey design; poststratification makes such adjustments in the data analysis.
- Modeling interactions is important when we care about differences in the treatment effect for different groups, and poststratification then arises naturally if a population average estimate is of interest.

Interpreting regression coefficients as treatment effects

- It can be tempting to take the coefficients of a fitted regression model and give them a causal interpretation, but this can be a mistake—even if the data come from randomized experiments.
- Sometimes in fact the data structure could be not suited to suggest causal explanations in the regression coefficients.
- As an imaginary example, suppose you conduct a meta-analysis of studies of incentive in sample surveys. You find 39 randomized experiments on the effects of incentives on survey response rates. In each experiment, respondents were randomly assigned to two or more conditions (for example, no incentive, an incentive of 2 euros, or an incentive of 5 euros).
- (Continue) A regression was then fit, predicting change in response rate compared to several predictors. However, *cautions need to be attached to the estimates*: they only apply to the sorts of surveys in the meta-analysis, which might not be representative of future surveys of interest, and it could be a mistake to extrapolate beyond the data. Because the incentive conditions were not randomly assigned to surveys, we have to be open to this sort of non-causal interpretation.

Adjusting for post-treatment variables

- We recommend adjusting for pre-treatment covariates when estimating causal effects in experiments and observational studies. However, it is generally not a good idea to adjust for variables measured *after* the treatment.
- As discussed in Section 21.1 of the ROS book, information on post-treatment variables can be included in the more complicated framework of instrumental variables or in more general mediator strategies.
- In brevity, adjusting for a post-treatment variable can bias the estimate of the treatment effect, *even when the treatment has been randomly assigned to study participants*. For a detailed explanation, see Section 19.6 of the ROS book.

Observational studies

Causal inference in observational studies i

- So far, we introduced a statistical formalization of causal effects using *potential outcomes*, focusing on the estimation of average causal effects and interactions using data from randomized controlled experiments.
- In theory, the simplest solution to the **fundamental problem of causal inference** is, as we have described, to randomly sample a different set of units for each treatment group assignment from a common population, and then apply the appropriate treatments to each group.
- In practice, however, we often work with observational data because, compared to experiments, observational studies can be more practical to conduct and can have more realism with regard to how the program or treatment is likely to be "administered" in practice-due to logistic, ethical, or financial constraints which can make it difficult or impossible to externally assign treatments.

Causal inference in observational studies ii

- As we have discussed, however, in observational studies treatments are *observed rather than assigned* (for example, comparisons of smokers to nonsmokers), and it is not at all reasonable to consider the observed data under different treatments as random samples from a common population.
- In an observational study, there can be systematic differences between groups of units that receive different treatments—differences that are outside the control of the experimenter—and they can affect the outcome, y.
- In this case we need to rely on more data than just treatments and outcomes and implement a more complicated analysis strategy that will rely upon stronger assumptions. The strategy discussed in this section, however, is relatively simple and relies on controlling for confounding covariates through linear regression.

The challenge of causal inference i

- When we are not able to randomly assign study participants to treatments, we present two simple examples in which predictive comparisons with observational data do not yield appropriate causal inferences.
- Consider a hypothetical medical experiment in which 100 patients receive the treatment and 100 receive the control condition. In this scenario, the causal effect represents a comparison between what would have happened to a given patient had he or she received the treatment compared to what would have happened under control. We suppose that *the treatment effect is zero*.

The challenge of causal inference ii

- Now let us further suppose that treated and control groups systematically differ, with healthier patients receiving the treatment and sicker patients receiving the control. This scenario is illustrated in Figure 20.1, where the distribution of previous health status is different for the two groups. This scenario leads to a positive predictive comparison between the treatment and control groups, even though the causal effect is zero. This sort of discrepancy between the predictive comparison and the causal effect is sometimes called selection bias.
- Conversely, it is possible for a truly nonzero treatment effect to be erased in the predictive comparison. Figure 20.2 illustrates: the treatment has a positive effect for all patients, whatever their previous health status, as displayed by outcome distributions.

The challenge of causal inference iii

- So, for any given unit, we would expect the outcome to be better under treatment than control. However, suppose that this time, sicker patients are given the treatment and healthier patients are assigned to the control condition, as illustrated by the different heights of these distributions. It is then possible to see equal average outcomes of patients in the two groups, with sick patients who received the treatment canceling out healthy patients who received the control.
- Previous health status *plays an important role in both these scenarios* because it is related both to treatment assignment and future health status. Simple comparisons of average outcomes across groups that ignore this variable will be misleading because the effect of the treatment will be "confounded" with the effect of previous health status. For this reason, such predictors are sometimes called *confounding covariates*.

Zero causal effect but positive predictive comparison

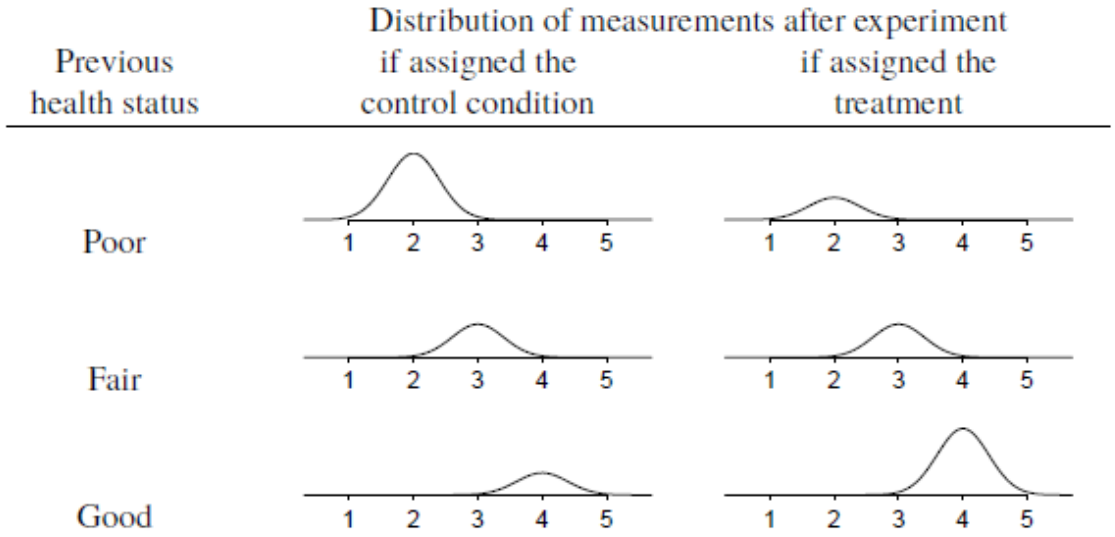

Figure 20.1 *Hypothetical scenario of* zero causal effect *of treatment: for any value of previous health status, the distributions of potential outcomes are* identical *under control and treatment. The heights of these distributions reflect the relative frequency of treatment receipt. Therefore, the* predictive comparison *between treatment and control is positive, because more of the healthier patients receive the treatment and more of the sicker patients receive the control condition.*

Positive causal effect but zero positive predictive comparison

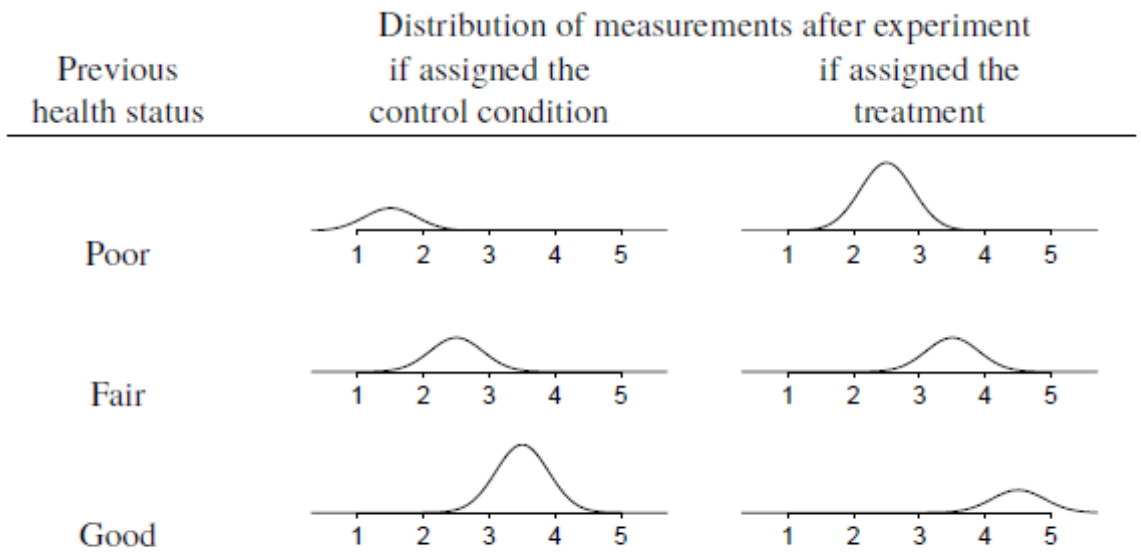

Figure 20.2 *Hypothetical scenario of* positive causal effect *of treatment: for any value of previous health status, the distributions of potential outcomes are centered at* higher *values for the treatment group than for the control group. Once again, the heights of these distributions reflect the relative frequency of treatment receipt. Therefore, the* predictive comparison *between treatment and control can be zero, because more of the sicker patients receive the treatment and more of the healthier patients receive the control condition. Compare to Figure 20.1.*

Adding regression predictors

- The preceding examples illustrate how a simple *predictive comparison* is not necessarily an appropriate estimate of a causal effect. However, we could compare treated and control units conditional on previous health status.
- Another way to estimate the causal effect in this scenario is to regress the outcome on two inputs: the treatment indicator and previous health status.
- If health status is the only confounding covariate—that is, *the only variable that predicts both the treatment and the outcome*—and if the regression model is properly specified, then the coefficient of the treatment indicator corresponds to the average causal effect in the sample.
- In general, then, causal effects can be estimated using regression *if the model includes all confounding covariates* (*predictors that can affect treatment assignment or the outcome*) and if the model is correct.
- Confounders can be *observed*, and everything is ok, or not observed, then they are "omitted" or "lurking" variables that complicate the quest to estimate causal effects.

Omitted variable bias i

- We can quantify the bias incurred by excluding a confounding covariate in the context where a simple linear regression model is appropriate and there is only one confounding covariate.
- Suppose the "correct" model is:

$$y_i = \beta_0 + \beta_1 z_i + \beta_2 x_i + \epsilon_i, \tag{9}$$

  where z_i is the treatment and x_i is the covariate for unit i.
- If instead the confounding covariate, x_i, is ignored, one can fit the model,

$$y_i = \beta_0^* + \beta_1^* z_i + \epsilon_i^*. \tag{10}$$

- To understand the relationship between (9) and (10), it helps to define a third regression:

$$x_i = \gamma_0 + \gamma_1 z_i + \nu_i. \tag{11}$$

Omitted variable bias ii

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

$$y_i = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)z_i + \epsilon_i + \beta_2\nu_i. \tag{12}$$

- Equating the coefficients of z in (10) and (12) we get:

$$\beta_1^* = \beta_1 + \beta_2\gamma_1. \tag{13}$$

- This correspondence helps demonstrate the definition of a confounding covariate. If there is no association between the treatment and the possible confounder (that is, $\gamma_1 = 0$) or if there is no association between the outcome and the confounder (that is, $\beta_2 = 0$), then the variable is not a confounder because there will be no bias ($\beta_2\gamma_1 = 0$).
- We explained the bias that can be incurred if a model is specified incorrectly, and we provided some intuition for the types of problems that can arise when we fail to account for all confounders.

Observational studies and confounding covariates i

- We used the term *observational study* to refer to any nonexperimental research design. This study could be *prospective* or *retrospective*, and it may or may not involve direct manipulation of the treatment or potential causal variable of interest.
- Advantage: more practical to conduct and may more accurately reflect how the treatment is likely to be administered in practice or the population that might be likely to be exposed to it.
- However, in observational studies, treatment exposure is observed rather than manipulated and it is not reasonable to consider the observed data as reflecting a random allocation across treatment groups.
- Thus, in an observational study, there can be systematic differences between groups of units that receive different treatments with respect to key covariates, x, that can affect the outcome, y.

Observational studies and confounding covariates ii

- Such covariates that are associated with the treatment and the potential outcomes are typically called confounders or confounding covariates because if we observe differences in average outcomes across these groups, we can't separately attribute these differences to the treatment or the confounders—the effect of the treatment is thus "confounded" by these variables.
- We briefly explore both the dangers and possibilities for using observational data to infer causal effects. The approaches discussed all involve direct or indirect attempts to address *imbalance* and *lack of overlap* in potential outcomes by adjusting for potential confounding covariates that act, in essence, as proxies for the potential outcomes. These approaches include regression adjustments, stratification, matching, and weighting, and combinations of these.

Electric Company data example i

We begin again with the Electric Company data example.

- Once the treatments had been assigned in this experiment, the teacher for each class assigned to the treatment group had the choice of *replacing* or *supplementing* the regular reading program with the television show. That is, all the classes in the treatment group watched the show, but some watched it instead of the regular reading program and others received it in addition.
- This procedural detail reveals that the treatment for the randomized experiment is more subtle than described earlier. As implemented, the experiment estimated the effect of making the program available, either as a supplement or replacement for the current curriculum.
- We now consider something slightly different: the effect of using the show to complement versus substitute for the existing curriculum.

Electric Company data example ii

- Given our inferential goal, it would be naive to simply compare outcomes across the children assigned to these two new treatment options—Replace or Supplement—and expect this to estimate the treatment effect. This is an example of the concern that prompts the advice "*correlation is not causation.*"
- What if we could envision a conditional random assignment similar to the randomized block designs discussed in the previous chapter? For instance suppose that the probability of assignment to the Replace or Supplement groups was determined by the average pre-test scores in that classroom, plus, potentially, some "noise," or variables unrelated to the potential outcomes.
- Ignorability of the Replace or Supplement decision implies that pre-test score is the only confounding covariate.
- When the probability of assignment to the treatment varies with the level of a pre-treatment variable, *it is important to account for that variable when estimating the treatment effect.*

Electric Company data example iii

- Perhaps the easiest way to do this with a continuous covariate is by including it as a regression predictor. For our example, we add to our data frame a variable called supp that equals 0 for the replacement form of the treatment, 1 for the supplement, and NA for the controls; this last drops the control observations from the analysis.
- Then we estimate the following regression for each grade:

$$y_i = \beta_0 + \beta_1 \mathtt{supp}_i + \beta_2 x_i + \epsilon_i, \tag{14}$$

where x_i is the pre-test score. Estimates are reported in Figure 20.3 (next slide). The uncertainties are high enough that the comparison is inconclusive except in grade 2, but on the whole the pattern is consistent with the reasonable hypothesis that supplementing is more effective than replacing in the lower grades.

Electric Company data example iv

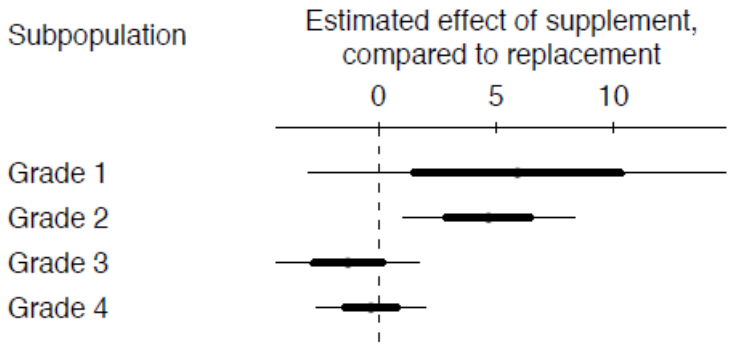

Figure 20.3 *Estimates, 50%, and 95% intervals for the effect of The Electric Company as a supplement rather than a replacement, as estimated by a regression on the supplement/replacement indicator also adjusting for pre-test data. For each grade, the regression is performed only on the treated classes; this is an observational study embedded in an experiment.*

The assumption of ignorability i

- In randomized experiments, as previously mentioned, it is usual to assume a *ignorability* hypothesis, meaning that

$$y^0, y^1 \perp z, \tag{15}$$

  which says that the distribution of the potential outcomes, (y^0, y^1), is the same across levels of the treatment variable, z.

- In observational studies, unlike completely randomized studies, it is typically implausible to assume independence between the potential outcomes and the treatment indicator. Instead, the key structural assumption is that, conditional on the covariates, x, the distribution of potential outcomes is the same across levels of the treatment variable, z.

$$y^0, y^1 \perp z \mid x, \tag{16}$$

The assumption of ignorability ii

which is even called *conditional ignorability*. In other words, the treatment has (or varieties of the treatment have) been assigned at random conditional on the inputs in the regression analysis (in this case, pre-test score) with respect to the potential outcomes.

- In the randomized block experiment, units are randomly assigned to treatment conditions within strata defined by the blocking variables, w. This design ensures that, within blocks defined by w, the distribution of the potential outcomes is the same across treatment groups—just as is true for all pre-treatment variables (or variables unaffected by the treatment).
- Recall that potential outcomes are conceptualized as existing before the treatment even occurs. *In the randomized block experiment, the blocking variables are the only confounding covariates by design.*
- In the observational studies we consider in this section, however, no actual randomized assignment has taken place.

The assumption of ignorability iii

- Crucially, however, we must make the leap of faith that we have conditioned on the appropriate set of confounders, x, such that the distribution of potential outcomes for observations who have the same level of these confounders is the same across treatment groups.
- As with the randomized block experiment, this assumption is called *ignorability of the treatment assignment* in the statistics literature. It is also called *selection on observables* or the *conditional independence* assumption in econometrics. The same assumption is often referred to as all *confounders measured* or *exchangeability* in the epidemiology literature. Failure to satisfy the assumption is sometimes referred to as *hidden bias or omitted variable bias*, a term that is also used more generally in statistics outside the causal context.

The assumption of ignorability iv

- The term ignorability reflects that this assumption *allows the researcher to ignore the model for the treatment assignment as long as analyses regarding the causal effects condition on the predictors needed to satisfy it*. If ignorability holds, causal inference does not, in theory, require modeling the treatment assignment mechanism.
- Recall that analyses of data resulting from completely randomized experiments need not condition on any pre-treatment variables to be unbiased—this is why we can use a simple difference in means to estimate causal effects. Analyses of such data can benefit from conditioning on pre-treatment variables, however, by achieving more precise estimates through a reduction in unexplained variation in the response variable.

The assumption of ignorability v

- Randomized experiments that block or match satisfy ignorability conditional on those design variables used to block or match (this assumes that the blocking variables are associated with the outcome). *One should therefore include these blocking or matching variables when estimating causal effects, both for concerns of bias and efficiency.*
- The same holds for observational studies that satisfy ignorability. If the probability of treatment varies with a covariate that also predicts the outcome (a confounder), then estimation of treatment effects must condition on this confounding covariates in order to be unbiased. *If a variable is related to the outcome but not the treatment, then we can include it to increase efficiency.*

Imbalance and lack of complete overlap i

- Causal inference is cleanest when the units receiving the treatment are comparable to those receiving the control. However, in an observational study, the treatment and control groups are likely to be different in multiple ways.
- If these differences across groups are with respect to *unobserved confounders*, then ignorability cannot be satisfied and the methods in this section are not appropriate.
- If these differences are with respect to *observed confounders*, then we can try to create comparability using the approaches discussed.
- To better understand the tradeoffs between these approaches, it will help to have a clearer understanding of two sorts of departures from comparability—*imbalance* and *lack of complete overlap*.

Imbalance and lack of complete overlap ii

- *Imbalance* with measured confounders occurs when the distributions of confounders differ for the treatment and control groups.
- When treatment and control groups suffer from imbalance, 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 the pre-treatment differences between the groups.
- See Figure 20.4: two examples of imbalance with respect to a single covariate, x. In Figure 20.4a, the groups have different means (dotted vertical lines) and different skews. In Figure 20.4b, groups have the same mean but different skews. In both examples, the standard deviations are the same across groups.

Imbalance and lack of complete overlap iii

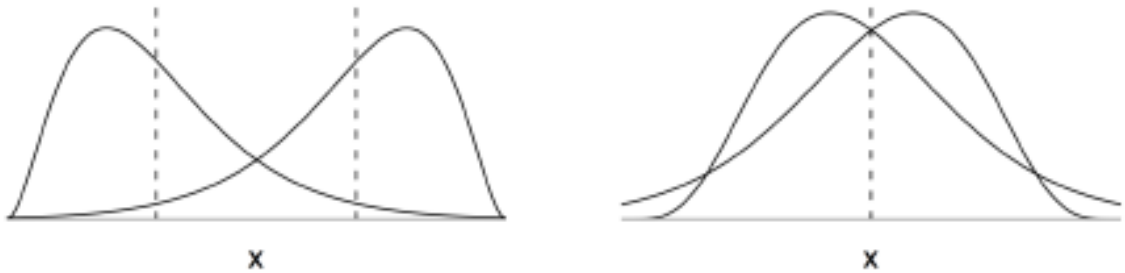

Figure 20.4 *Imbalance in distributions across treatment and control groups. (a) In the left panel, the groups differ in their averages (dotted vertical lines) but cover the same range of x. (b) The right panel shows a more subtle form of imbalance, in which the groups have the same average but differ in their distributions.*

- To see the consequences of imbalance on modeling, we try to make inferences about the effect of a treatment variable (for instance, a new reading program) on test score, y, while adjusting for a crucial confounding covariate, pre-test score, x. Given the true treatment effect is θ and the relationship between the response variable, y, and the sole confounding covariate, x, is quadratic:

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

Imbalance and lack of complete overlap iv

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

$$\theta = \bar{y}_1 - \bar{y}_0 - \beta_1(\bar{x}_1 - \bar{x}_0) - \beta_2(\bar{x^2}_1 - \bar{x^2}_0),$$

  where $\bar{y}_1, \bar{y}_0$ denote the average of the outcome test scores in the treatment and control groups, respectively, $\bar{x}_1, \bar{x}_0$ represent average pre-test scores for treatment and control groups, respectively.

- Ignoring x and simply using the raw treatment/control comparison, $\bar{y}_1 - \bar{y}_0$, will yield a poor estimate of the treatment effect. It will be off by the amount $\beta_1(\bar{x}_1 - \bar{x}_0) - \beta_2(\bar{x^2_1} - \bar{x^2}_0)$, which corresponds to systematic pre-treatment differences between groups 0 and 1.

Imbalance and lack of complete overlap v

- 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 distributions of pre-test scores are across treatment and control groups, the smaller this bias will be.
- *Overlap* or *common support* describes the extent to which the support of the covariate data is the same between the treatment and control groups.
- There is complete overlap when there exist both treatment and control units in all neighborhoods of the covariate space. Lack of complete overlap in the confounders creates problems, because in that setting there are treatment observations for which we have no empirical counterfactuals (that is, control observations with the same covariate distribution) or vice versa.

Imbalance and lack of complete overlap vi

- Figure 20.5 displays several scenarios of lack of complete overlap with respect to one confounder. It becomes increasingly difficult to visualize overlap as the dimension of the confounder space gets larger. Areas with no overlap represent conditions under which we may not want to make causal inferences.

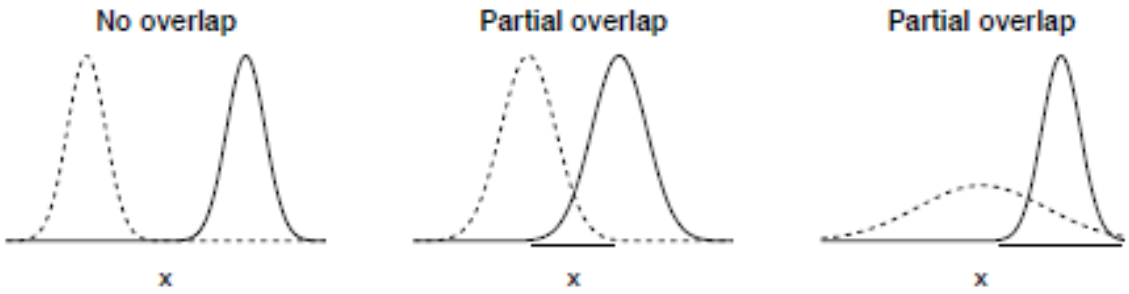

Figure 20.5 *Lack of complete overlap in distributions across treatment and control groups. Dashed lines indicate distributions for the control group; solid lines indicate distributions for the treatment group. (a) Two distributions with no overlap; (b) two distributions with partial overlap; (c) a scenario in which the* range *of one distribution is a subset of the range of the other.*

What should be covered next...

We broadly focused on causal inference strategies that assume ignorability of exposure or treatment assignment. However, when are we really confident that we have measured all confounders? There are some extensions/alternatives.

- Estimating causal effects indirectly using instrumental variables: in some situations when the argument for ignorability of the treatment assignment seems weak, there may exist another variable that does appear to be randomly assigned or can be considered as such. If this variable, called the *instrument*, I, is predictive of the treatment, z, then we may be able to use it to isolate a particular kind of targeted causal estimand. The instrument should only affect the treatment assignment but not have a direct effect on the outcome.
- Regression discontinuity: often the assumption of ignorability is not plausible; that is, it does not make sense to assume that treatment assignment depends only on observed pre-treatment predictors. However, we can design observational studies for which the assignment mechanism is entirely known (as with a controlled experiment) but for which no explicit randomization is involved.
- For more details, see Chapter 21 of the ROS book.

Sum-up about the Electric Company data example

- You find the data in the official Moodle course page.
- For further open discussion in class:
  - Repeat all the analysis.
  - Fit the models (6), (7) and (8) from a Bayesian and a frequentist point of view, provide the estimates (also from a graphical perspective) and comment the results.
  - Fit a multilevel/hierarchical model on these data and compare the results with those previously obtained.

Further reading

To properly capture the contents and the details about causal inference modeling, we strongly suggest the following further reading:

- Chapter 18, 19, and 20 from *Regression and Other Stories*, by A. Gelman, J. Hill, and A. Vehtari.
- Chapter 9 and 10 from *Data Analysis using Regression and Multilevel/Hierarchical models*, by A. Gelman and J. Hill.