12
Quasi-experiments and matching

Natural experiments

A randomized, controlled experiment is the gold standard of evidence for a causal hypothesis. Yet many times an experiment is impossible, impractical, unethical, or too expensive in time or money. In these situations, it often pays to look for something called a *natural experiment*, also called a *quasi-experiment*. A natural experiment is not something that you, as the investigator, design. Rather, it is an "experiment" where nature seems to have done the randomization and intervention for you, thereby giving you the same type of balance between treatment and control groups that you'd expect to get out of a real experiment.

This idea is best understood by example. Suppose you want to study the effect of class size on student achievement. You reason that, in smaller classes, students can get more individual attention from the instructor, and that intructors will feel a greater sense of personal connection to their students. All else being equal, you believe that smaller class sizes will help students learn better.

A cheap, naïve way to study this question would be to compare the test scores of students in small classes to those of students in larger classes. Any of these confounders, however, might render such a comparison highly unbalanced, and therefore dubious: (1) students in need of remediation are sometimes put in very small classes; (2) highly gifted students are also sometimes put in very small classes; (3) richer school districts can afford both smaller classes and many other potential sources of instructional advantage; or (4) better teachers successfully convince their bosses to let them teach the smaller classes themselves.

An expensive, intelligent way to study this question would be to design an experiment, in conjunction with a scientifically inclined school district, that randomly assigned both teachers and students to classes of varying size. In fact, a few school systems have done exactly this. A notable experiment is Project STAR in

Question	Problem	Natural experiment	Lingering issues
Does being rich make people happy?	Even if richer people are happier on average, maybe happiness and success are the common effect of a third factor. Or maybe the rich grade on a different curve than the rest of us.	Compare a group of lottery winners with a similar group of people who played the lottery but didn't win.	Lottery winners may play the lottery far more often than people who played the lottery but didn't win, which might correlate with other important differences.
Does smoking increase a person's risk for Type-II diabetes?	People who smoke may also engage in other unhealthy behaviors at systematically different rates than non-smokers.	Compare before-and-after rates of diabetes in cities that recently enacted bans on smoking in public places.	Maybe the incidence of diabetes would have changed anyway.
Do bans on mobile phone use by drivers in school zones reduce the rate of traffic collisions?	Groups of citizens that enact such bans may differ systematically in their attitudes toward risk and behavior on the road.	Go to Texarkana, split by State Line Avenue. Observe what happens when Texas passes a ban and Arkansas doesn't.	There may still be systematic differences between the two halves of the city.

Table 12.1: Three hypothetical examples of natural experiments.

Tennessee—an expensive, lengthy experiment that studied the effect of primary-school class sizes on high-school achievement, and showed that reduced class sizes have a long-term positive impact both on test scores and drop-out rates.[1]

But suppose you are neither naïve nor rich, and yet still want to study the question of whether small class sizes improve test scores. If you're in search of a third way—one that's better than merely looking at correlations, yet cheaper than a full-fledged experiment—you might be interested to know the following fact about the Israeli school system.

> [I]n Israel, class size is capped at 40. Therefore, a child in a fifth grade cohort of 40 students ends up in a class of 40 while a child in a fifth grade cohort of 41 students ends up in a class only half as large because the cohort is split. Since students in cohorts of size 40 and 41 are likely to be similar on other dimensions, such as ability and family background, we can think of the difference between 40 and 41 students enrolled as being "as good as randomly assigned."[2]

This is a lovely example of a natural experiment—something you didn't design yourself, but that is almost as good as if you had. The researchers in this study compared the students in a group of 40 ("control group," in one large class) versus the students in a group of 41 ("treatment group," split into two smaller classes). This is a plausibly random assignment: the "randomiza-

[1] The original study is described in Finn and Achilles (1990). "Answers and Questions about Class Size: a Statewide Experiment." *American Educational Research Journal* 28, pp. 557–77

[2] Angrist and Pischke (2009). *Mostly Harmless Econometrics*, Princeton University Press, p. 21

tion mechanism" is whether a student fell into a peer group of 40 versus a peer group of 41, and we would not expect this difference to be confounded by anything else that might predict test scores. Therefore, if we see a big difference in performance between the two groups, the most likely explanation is that class size caused the difference.

The two key features of a natural experiment, also called a quasi-experiment, are the following:

1. There are treatment and control groups.
2. The allocation to treatment and control groups is outside the control of either the investigators or the people actually in the study, and plausibly "resembles a random assignment," in the sense that it balances nuisance variables between the two groups.

Some natural experiments, of course, are better than others. The key question in evaluating the quality of a natural experiment almost always boils to the quality of the randomization mechanism, i.e. how effective it is at balancing nuisance variables. Consider the examples in Table 12.1, on page 116. For each one, ask yourself two questions. (1) What are the "treatment" and "control" groups? (2) How balanced are these two groups? (Said another way: how good is the quasi-randomization of cases to these groups?) Think carefully about each one, and you may begin to see "experiment" versus "non-experiment" as the black and white ends of a spectrum, with many shades of grey in between.

Matching

Matching is a strategy for attempting to estimate a causal effect from observational data, rather than experimental data. To estimate a causal effect by matching, we artificially construct a balanced data set out of an unbalanced one, by explicitly matching treated cases with similar control cases. We then compare the outcomes in treatment versus control groups, using only the balanced data set and discarding the cases without good matches.

Matching is closely related to the idea of blocking in randomized experiments. The big difference is that blocking, where units are explicitly paired or grouped in an attempt to hold known nuisance variables constant, is something that's done *before* randomization, in the context of a controlled (or natural) experiment.

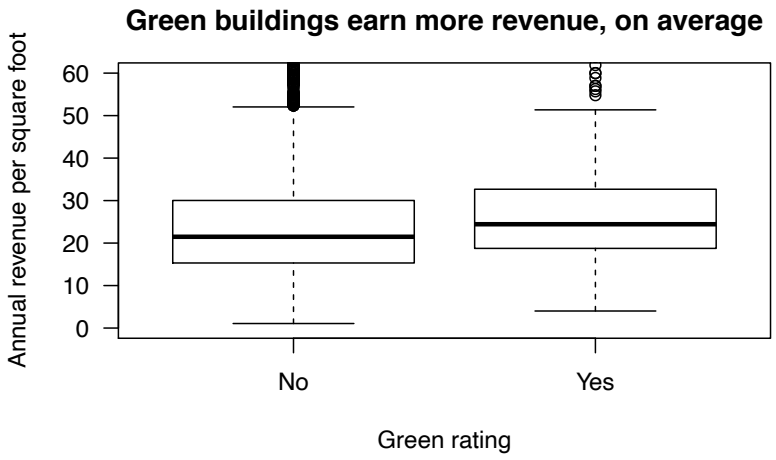

Figure 12.1: Green buildings seem to earn more revenue per square foot, on average, than non-green buildings.

Matching involves a similar process—pairing or grouping units to hold known nuisance variables constant—but it is done *in the absence of* randomization.

This is most readily seen by example.

An example: the value of going green

For many years now, both investors and the general public have paid increasingly close attention to the benefits of environmentally conscious ("green") buildings. There are both ethical and economic forces at work here. To quote a recent report by Mercer, an investment-consulting firm, entitled "Energy efficiency and real estate: Opportunities for investors":

> Investing in energy efficiency has two intertwined virtues that make it particularly attractive in a world with a changing climate and a destabilized economy: It cuts global-warming greenhouse gas emissions and saves money by reducing energy consumption. Given that the built environment accounts for 39 percent of total energy use in the US and 38 percent of total indirect CO_2 emissions, real estate investment represents one of the most effective avenues for implementing energy efficiency.

This only scratches the surface. In commercial real estate, issues of eco-friendliness are intimately tied up with ordinary decisions about how to allocate capital. Every new project involves negotiating a trade-off between costs incurred and benefits realized over

the lifetime of the building. In this context, the decision to invest in an eco-friendly building could pay off in at least four ways.

(1) Every building has the obvious list of recurring costs: water, climate control, lighting, waste disposal, and so forth. Almost by definition, these costs are lower in green buildings.

(2) Green buildings are often associated with indoor environments that are full of sunlight, natural materials, and various other humane touches. Such environments, in turn, might result in higher employee productivity and lower absenteeism, and might therefore be more coveted by potential tenants. The financial impact of this factor, however, is rather hard to quantify *ex ante*; you cannot simply ask an engineer in the same way that you could ask a question such as, "How much are these solar panels likely to save on the power bill?"

(3) Green buildings make for good PR. They send a signal about social responsibility and ecological awareness, and might therefore command a premium from potential tenants who want their customers to associate them with these values. It is widely believed that a good corporate image may enable a firm to charge premium prices, to hire better talent, and to attract socially conscious investors.

(4) Finally, sustainable buildings might have longer economically valuable lives. For one thing, they are expected to last longer, in a direct physical sense. (One of the core concepts of the green-building movement is "life-cycle analysis," which accounts for the high front-end environmental impact of acquiring materials and constructing a new building in the first place.) Moreover, green buildings may also be less susceptible to market risk—in particular, the risk that energy prices will spike, driving away tenants into the arms of bolder, greener investors.

Of course, much of this is mere conjecture. At the end of the day, tenants may or may not be willing to pay a premium for rental space in green buildings. We can only find out by carefully examining data on the commercial real-estate market and comparing "green" versus "non-green" buildings. By "green," we mean that a commercial property has received some official certification, because its energy efficiency, carbon footprint, site selection, and

	Original data	
	Non-green buildings	Green buildings
Sample size	6928	678
Mean revenue/sq ft.	24.51	26.97
Mean age	49.2	23.9
Class A	37%	80%
Class B	48%	19%
Class C	15%	1%

Table 12.2: Covariate balance for the original data. Class A, B, and C are relative classifications within a specific real-estate market. Class A buildings are generally the highest-quality properties in a given market. Class B buildings are a notch down, but still of reasonable quality. Class C buildings are the least desirable properties in a given market.

building materials meet certain environmental benchmarks, as certified by outside engineers.[3]

Let's look at some data on 678 green-certified buildings in the United States, together with 6,298 non-green buildings in similar geographic areas. The boxplot above shows that, when we measure revenue by a building's rental rate per square foot per year, green buildings tend to earn noticeably higher revenue (mean = 26.97) than non-green buildings (mean = 24.51). That's a difference of $2.46 per square foot, or nearly a 10% market premium.

However, there's a problem with this comparison. As Table 12.2 shows, the green buildings tend to be newer than the non-green buildings, and are more likely to be "Class A" buildings.

So the important question is: do green buildings command a market premium *because* they are green, or simply because they are newer, better buildings in the first place? We can't tell by simply computing the average revenue in each group, because the green ("treatment") and non-green ("control") groups are highly unbalanced with respect to some important confounders.

This is where matching comes in. Matching means constructing a balanced data set from an unbalanced one. It involves three steps:

(1) For each case in the treatment group, find the case in the control group that is the closest match in terms of confounding variables, and pair them up. Put these matched pairs into a new matched data set, and discard the cases in the original data set for which there are no close matches.

(2) Verify covariate balance for the matched data set, by checking that the confounders are well balanced between the treatment and control groups.

(3) Assuming that the confounders are approximately balanced,

[3] The two most common certifications are LEED and EnergyStar; you can easily find out more about these rating systems on the web, e.g. at www.usgbc.org.

	Matched data	
	Non-green buildings	Green buildings
Sample size	678	678
Mean revenue/sq ft.	25.94	26.97
Mean age	23.9	23.9
Class A	80%	80%
Class B	19%	19%
Class C	1%	1%

Table 12.3: Covariate balance for the matched data.

then compare the treatment-group outcomes with the control-group outcomes, using *only* the matched pairs.

Matching relies on a simple principle: compare like with like. In this example, that means if we have a 25-year-old, Class A building with a green rating, we try to find another 25-year old, Class A building without a green rating to compare it to.

In this particular example, once we've constructed the data set of matched pairs, the confounder variables are much more closely balanced between the treatment and control groups (see Table 12.3). A comparison of revenue rates for this matched data set makes the premium for green buildings look a lot smaller: $26.97 versus $25.94, or about a 4% premium. Compare that with the 10% green premium we estimated from the original, unmatched data.

How do we actually find matches? The nitty-gritty algorithmic details of actually finding good matched pairs of cases are best left to the experts who write the software for these things, and we're not concerned with those details here. As an aside, the two most common types of matching are called *nearest-neighbor search* and *propensity-score matching*; follow the links if you'd like to know more. In R, the package `MatchIt` uses propensity-score matching as a default; this is a very commonly used algorithm in real-world data analysis. If you're interested in exploring these ideas on your own, the paper linked here[4] has a much more detailed overview of different matching methods.

[4] "Matching Methods for Causal Inference: A Review and a Look Forward." Elizabeth A. Stuart, *Statistical Science*, 2010.

Matching isn't a silver bullet: a bigger example

If you've ever been admitted to the intensive-care unit at a hospital, you may have undergone a diagnostic procedure called right heart catheterization, or RHC. RHC is used to see how well a pa-

	Original data		Matched data	
	No RHC	RHC	No RHC	RHC
Sample size	3551	2184	2184	2184
180-day survival rate	0.370	0.320	0.354	0.320
mean APACHE score	50.934	60.739	57.643	60.739
Trauma	0.005	0.016	0.008	0.016
Heart attack	0.030	0.043	0.036	0.043
Congestive heart failure	0.168	0.195	0.209	0.195
Sepsis	0.148	0.321	0.24	0.321

Table 12.4: A before-and-after table of summary statistics showing covariate balance for the observational study on right-heart catheterization. The entries for trauma, heart attack, etc. show rates of these complications in the two groups. The left half of the table shows the original data set, while the right half shows the matched data set.

tient's heart is pumping, and to measure the pressures in that patient's heart and lungs. RHC is widely believed to be helpful, since it allows the doctor to directly measure what's going on inside a patient's heart. But it is an invasive procedure, since it involves inserting a small tube (the catheter) into the right side of your heart, and then passing that tube through into your pulmonary artery. It therefore poses some risks—for example, excessive bleeding, partial collapse of a lung, or infection.

A natural question is: do the diagnostic benefits of RHC outweigh the possible risks? But this turns out to be tricky to answer. The reason is that doctors would not consider it ethical to run a randomized, controlled trial to see if RHC improves patient outcomes. As the authors of one famous study from the 1990s pointed out:[5]

> Many cardiologistics and critical care physicians believe that the direct measurement of cardiac function provided by right heart catheterization (RHC) . . . is necessary to guide therapy for certain critically ill patients, and that such management leads to better patient outcomes. While the benefit of RHC has not been demonstrated in a randomized controlled trial (RCT), the popularity of this procedure, and the widespread belief that it is beneficial, make the performance of an RCT difficult. Physicians cannot ethically participate in such a trial or encourage a patient to participate if convinced the procedure is truly beneficial.

[5] "The effectiveness of right heart catheterization in the initial care of critically ill patients." Connors et. al. *Journal of the American Medical Association*. 1996 Sep 18; 276(11):889-97.

We're therefore left with only observational data on the effectiveness of RHC—which, on the surface, doesn't look good! Here's the data from the study quoted above, showing that critically ill patients undergoing RHC actually have a *worse* 180-day survival rate (698/2184, or 32%) than patients not undergoing RHC

(1315/3551, or 37%):

	No RHC	RHC
Survived 180 days	1315	698
Died within 180 days	2236	1486

What's going on here? Should we conclude that right heart catheterization is actually killing people, and that the doctors are all just plain wrong about its putative benefits?

Not so fast. The problem with this conclusion is that the treatment (RHC) and control (no RHC) groups are heavily unbalanced with respect to baseline measures of health. Put simply, the patients who received RHC were a lot sicker to begin with, so it's no surprise that they have a lower 6-month survival rate. To cite a few examples: the RHC patients were three times more likely to have suffered acute trauma, 50% more likely to have had a heart attack, and 16% more likely to be suffering from congestive heart failure. The RHC patients also had an average APACHE score that was 10 points higher than the non-RHC patients.[6] The left half of Table 12.4 shows these rates of various complications for the two groups in the original data set. They're quite different, implying that the survival rates of these two groups cannot be fairly compared.

[6] The APACHE score is a composite severity-of-disease score used by hospital ICUs to estimate which patients have a higher risk of death. Patients with higher numbers have a higher risk of death.

And what about after matching? Unfortunately, Table 12.4 shows that, even after matching treatment cases with controls having similar complications, the RHC group still seems to have a lower survival rate. The gap looks smaller than it did before, on the unmatched data—a 32% survival rate for RHC patients, versus a 35.4% survival rate for non-RHC patients—but it's still there.

Again we find ourselves asking: what's going on? Is the RHC procedure actually killing patients? Well, it might be, at least indirectly! The authors of the study speculate that one possible explanation for this finding is "that RHC is a marker for an aggressive or invasive style of care that may be responsible for a higher mortality rate." Given the prevalance of overtreatment within the American health-care system, this is certainly plausible.

But we can't immediately jump to that conclusion on the basis of the matched data. In fact, this example points to a couple of basic difficulties with using matching to estimate a causal effect.

The first (and most important) difficulty is that *we can't match on what we haven't measured.* If there is some confounder that we don't know about, then we'll never be able to make sure that it's

balanced between the treatment and control groups within the matched data. This is why randomized experiments are so much more persuasive: because they *also* ensure balance for unmeasured confounders. The authors of the study acknowledge as much, writing:

> A possible explanation is that RHC is actually beneficial and that we missed this relationship because we did not adequately adjust for some confounding variable that increased both the likelihood of RHC and the likelihood of death. As we found in this study, RHC is more likely to be used in sicker patients who are also more likely to die.

Another possible explanation is that we simply haven't been able to match treatment cases with control cases very effectively. The right half of Table 12.4 shows that covariate balance for the matched data is noticeably better than for the unmatched data, but it's not perfect. We still see some small differences in complication rates and APACHE scores between the treatment and control group.

The reason for this is simple: although finding a match on one or two variables is relatively easy, finding a match on several variables is harder. Think of this in terms of your own life experience—for example, in seeking a spouse or partner. It may not be that hard to find someone who's a good match for you in terms of your interests in movies. But if you require that this person *also* match you in terms of age, career, education, home town, height, weight, looks, and favorite sport, then you're a lot less likely to find a match. *Picky people are less likely to find a satisfying match.* For this same reason, it's unlikely that we'll be able to find an exact "control" match for each "treated" case if we're forced to be picky by the presence of many possible confounders. Finding matches for cases with rare confounders is especially hard—by definition, since the confounder is rare!

This point underlines a basic difficulty with matching: perfect matches usually don't exist, and we have no choice but to accept approximate matches. In practice, therefore, we give up on the requirement that every single pair of matched observations is similar in terms of all possible confounders, and settle for having matched groups that are similar in their confounders, *on average*. That's why it's so important to check the covariate balance after finding matched pairs, to make sure that there's nothing radically different between the two groups.