Making Decisions with Data:
An Introduction to Causal Inference[1]

Alexander W. Butler, Erik J. Mayer

March 2015

Abstract

This article provides a brief and intuitive introduction to methods used in causal inference, suitable for a classroom setting. The paper formalizes the notion that correlation does not imply causation, and develops familiarity with statistical skills to make business policy decisions that are properly informed by data. An emphasis is on establishing compelling counterfactuals when policy choices are subject to selection bias and other endogeneity concerns.

Does consuming coffee cause you stress? Will hedging your company's commodity risk increase firm value? Firms that hedge appear to be more valuable, on average, and coffee guzzlers appear to be stressed out, so it is tempting to conclude answers in the affirmative to both questions. But correlation (higher stock prices and hedging activity go hand-in-hand) does not imply causation (hedging *causes* the higher stock price). Although it is often straightforward to identify a correlation, it is not so simple to identify actual causation cleanly. This article provides some recommendations for how to do so.[2]

Policy choices—such as whether to drink less coffee, hedge commodity risk, or take an ibuprofen to attempt to rid yourself of a headache—are best informed with evidence on the causal implications of those choices. Causal inference is the identification of a causal relation between A and B. Providing convincing evidence to support causal statements is often challenging because reverse causality, omitted factors, and chance can all create a correlation between A and B without A actually causing B. Several examples will illustrate the challenges we face when trying to make decisions that are properly informed by data.

[1] We are grateful to Grace Butler, Lee Ann Butler, and Alan Crane for conversations that helped us improve the paper.

[2] The article is intentionally intuitive rather than rigorous. Elevated treatments of these topics can be found in "Mastering 'Metrics" by Angrist and Pischke, "Mostly Harmless Econometrics: An Empiricist's Companion" by Angrist and Pischke, and textbook treatments by Wooldridge, and others.

Consider the following hypothetical example. You are not feeling well, and your doctor prescribes you medication, which you take, as directed, every 12 hours for the following week. Your symptoms disappear and you are back to 100%. You might naturally conclude that the medication worked and that it caused you to get better. However, we have no way of knowing what your recovery would have been like in the absence of treatment. The outcome we do not observe—e.g., having foregone medical treatment—is called the *counterfactual outcome*. A counterfactual outcome is what people are referring to when they say things like, "you took the medicine and you got better quickly, but…*compared to what*?" In this case, perhaps the counterfactual is that you would have recovered even faster without medication, or at the same rate, or maybe it would have taken you weeks to recover. The fact that we can never observe *both* the realized outcome and the counterfactual for the same event is the fundamental problem of causal inference.

Pharmaceutical companies face the same problem in evaluating the efficacy of new drugs. For any given test subject, they do not know what outcome the patient would have experienced in the absence of treatment with the drug. To estimate this counterfactual outcome, they conduct large clinical trials of new drugs, during which they administer the drug to many patients. They also administer placebo pills designed to have no effect to a large number of patients who are in a similar condition. These trials are true experiments, because patients are randomly assigned to be in the treatment group (those who receive the drug) or the control group (those who receive the placebo pill). The trials then compare the outcomes of patients in the treated group to those of patients in the control group. If the outcomes for the treated group are notably better, then the drug is deemed to have caused the difference in outcomes.[3]

These trials use the control group outcomes as an estimate of the treated group's counterfactual outcomes. Using this comparison group's (the placebo takers') outcome as an estimate of the treated group's counterfactual outcome is valid when individuals in the two groups are comparable, for instance, because they were randomly assigned to groups. When it comes to providing evidence that a treatment has a causal effect on an outcome, random experiments are the gold standard against which other methodologies are judged. Unfortunately, decision makers usually do not have the luxury of conducting true randomized experiments to

[3] If the researcher suspects the placebo taking control group may show improvement, which is the better counterfactual: a control group taking no pills, or a control group taking fake pills?

answer policy questions. This article will provide an overview of several methods we use to conduct causal inference from observational/non-experimental data.

The use of observational data carries several potential problems that a researcher must address before claiming to document a causal effect. One important concern that a researcher must address is the possibility of *reverse causality*. In the coffee-stress example above, it is possible that stress causes people to drink more coffee, in addition to or instead of the other way around. If stress causes people to drink more coffee, it will generate the same observed pattern (that stress and coffee consumption are correlated) that we would expect to find if coffee caused people to be stressed. But the policy implication is very different.

Additionally, other factors, such as a person's work environment, have the potential to influence both their stress level and their coffee consumption. Consider the case of an energy trading desk versus a yoga studio. It might be the case that the trading desk causes its employees lots of stress, and at the same time provides them with an unlimited supply of free coffee. The yoga studio might provide its instructors with both a low-stress work environment and no free coffee. Therefore, if we find that energy traders drink more coffee and are more stressed than yoga instructors, what are we to conclude? That coffee causes people to be more stressed? That stress causes people to drink coffee? Or that there is no causal relationship—in either direction—between stress and coffee consumption, but stressful work environments include coffee-drinking as a part of their culture? This case shows the possibility of *omitted variable bias*. This omitted variables bias problem arises when the treatment we wish to study (coffee consumption) and the outcome (stress) are both related to a third variable whose role we do not account for (work environment).

Inlay:

Policy question: "Does hedging commodity risk cause firms to be more valuable?"

Observation: Data show that oil companies who hedge using financial derivatives are valued more highly on average than those who do not.

Simplistic conclusion: Hedging causes firms to be more valuable.

Challenge #1, Reverse Causality: It is likely the case that corporate hedging programs are costly to start up. Firms have to hire financial experts and run the hedging program in-house, or contract with a financial services company who has expertise in the area. Profitable and highly

valued firms can better afford hedging programs, whereas less profitable firms avoid paying these startup costs. Therefore, it is possible that higher firm value allows the firm to hedge more.

Challenge #2, Omitted Variable Bias: The oil companies with the most financially astute executives are also the most likely to implement derivative hedging programs. These executives are also likely to have the greatest business acumen and to run some of the most profitable and highly valued companies. Therefore, it might be that smart executives cause their company to have a hedging program, and also cause their company to be highly valued. This would generate the pattern that companies with hedging programs tend to be more highly valued, even if hedging has no direct effect on value.

To sort out whether A causes B, B causes A, or a third element C causes both A and B, we need additional information. A source of *exogenous variation* allows us to see what would happen if we manipulated just one aspect of a larger system (e.g. coffee consumption), while holding all other aspects constant (work environment, personality, etc.). In manipulating just this one aspect, we can learn the causal effect that it has on the outcome (stress levels). In a true experiment, this exogenous variation is built directly into the experimental design because treatment is assigned randomly, and that assignment is not related to any characteristics of the subjects. In contrast, when we are analyzing observational data, we need to recognize that the treatment is generally chosen by individuals instead of being randomly assigned to them, creating the possibility of reverse causality and omitted variables bias.

In the absence of a true experiment, identifying exogenous variation in the assignment of treatment requires some creativity. For instance, if the person responsible for purchasing coffee supplies for the office of the energy traders quit, and there was no coffee for a week, we could use this as a source of exogenous variation in the coffee consumption of the traders. We could compare the stress levels of the traders who drink coffee to those who do not. This hypothetical situation is a *natural experiment*—a naturally occurring event that provides us with a treatment and control group, similar in spirit to a random experiment.[4]

These examples introduce the two most important concepts in causal inference: the need to establish a valid counterfactual outcome and the importance of identifying exogenous variation in the treatment. The rest of this article will discuss these concepts in greater detail, and

[4] Of course, if having one's coffee purchaser quit is itself stress-inducing then we should be concerned as to whether this experiment violates the "only through" condition necessary for an ideal experimental design. We return to this condition later.

introduce techniques for constructing counterfactuals and identifying naturally occurring sources of exogenous variation to conduct causal inference.

1 The Foundation: Treatment Effects, Random Experiments, and Selection Bias

True randomized experiments can provide compelling evidence on the causal relation between two variables. In order to understand true experiments better, we will introduce the potential outcomes notation. The approach here will be intuitive.[5] However, the discussion will use a few basic concepts from mathematics/statistics such as expectation and conditional expectation.

As most textbook discussions of the topic do, we will denote the outcome variable we study as *Y*, and the treatment indicator as *T*. The potential outcomes notation is our way of recognizing that individuals may have different outcomes if they are treated or untreated. This notation will provide us with the framework to discuss how to measure the effect of treatment, and the potential issues we face when measuring the treatment effect in non-experimental settings. In practice, we are often interested in the effect of continuous treatments that can take many values/levels of intensity (e.g., the effects of consuming zero, one, two, or ten cups of coffee in a day). But for the sake of the exposition, we will focus on binary treatments—situations where either the individual receives the treatment (hedging, coffee, drugs) or they do not. That is $T=1$ or $T=0$. In the pharmaceutical study example, participants receiving the real drug have $T=1$ and those receiving the placebo have $T=0$.

Potential Outcomes Notation:

$Y_i(1)$: the outcome of individual *i* if they are treated

> Coffee Ex: The stress level of individual *i* if they consistently drank more than 2 cups of coffee
>
> Hedging Ex: The value of firm *i* if they hedge using derivatives

[5] For more rigorous approaches, see textbook treatments by Angrist and Pischke *Mostly Harmless Econometrics: An Empiricist's Companion,* and Wooldridge *Econometric Analysis of Cross Section and Panel Data*.

$Y_i(0)$: the outcome of individual *i* if they remain untreated

Coffee Ex: The stress level of individual *i* if they consistently drank less than 2 cups

Hedging Ex: The value of firm *i* if they do not hedge

T=1: denotes that the individual was in the treated group

Coffee Ex: *T*=1 means that the individual consistently consumes more than 2 cups

Hedging Ex: *T*=1 means the firm hedges using derivatives

T=0: denotes that the individual was in the control group

Coffee Ex: *T*=0 means that the individual consistently consumes less than 2 cups

Hedging Ex: *T*=0 means the firm does not hedge

Measures of the Treatment Effect (ATE, ATT):

The treatment effect, $Y_i(1)$ - $Y_i(0)$, is the outcome for individual *i* if they were to be treated minus their outcome if they went untreated. Because we never observe both $Y_i(1)$ and $Y_i(0)$, we cannot compute this treatment effect directly. This difficulty is the fundamental challenge of causal inference. Instead, we must choose a method to construct an approximate value for the counterfactual using the data available to us (more on this in the next section).

In order to perform statistical inference and examine whether there is a treatment effect, we need to average the treatment effect over multiple individuals. Whereas idiosyncratic factors may influence an individual's response to treatment, we can draw more reliable conclusions if we have a large sample of observed outcomes. (A study of 10,000 participants is more informative than a similar study of 10 because idiosyncratic characteristics of participants tend to cancel out in large groups.) There are two measures of the treatment effect that decision makers might be interested in: the average treatment effect (ATE), and the average treatment effect on the treated (ATT). The ATE is the average effect for the entire sample, meaning that we use all individuals when computing the mean and testing the statistical significance of the treatment effects $Y_i(1)$ - $Y_i(0)$ (remember that for each individual one of these values is observed, and the

other is the constructed counterfactual). The ATT is the average effect for the subsample of individuals who were actually treated. If the treatment has the same effect on all individuals, then the ATE and the ATT are equivalent. But if treatment is likely to have different effects on different subpopulations, then the ATT and ATE can differ. For instance, imagine we are examining the treatment effect of a job training program. The people who gravitate to such a training program (and get "treated", T=1) are likely to be unemployed, or dissatisfied with their current job. Conversely, top executives at Exxon Mobil, Conoco Phillips, or IBM are unlikely to find such a program worthwhile and choose not to participate (T=0). The ATT measures the effect on the individuals who actually receive treatment, whereas the ATE measures the expected average effect if the treatment were applied to all individuals. In this example, job training, the ATT is the more interesting measure of the treatment effect. The mathematical definitions for the ATE and ATT are provided below.

ATE: $E[Y_i(1) - Y_i(0)]$

ATT: $E[Y_i(1) - Y_i(0) \mid T = 1]$

Randomized Experiments:

Randomized experiments such as clinical trials are the gold standard when it comes to establishing the causal effect of a treatment. The potential outcomes framework and measures of the treatment effect (ATE/ATT) clarify why random experiments give compelling evidence of the causal effect of some treatment. As mentioned in the introduction, random experiments allow us to just take the difference in the average outcome for the treated group minus the average outcome for the control group as the treatment effect. The discussion of why this works focuses on equation (1) below. In order to convince yourself that this equation is true, notice that the first and fourth elements on the right side are the same as the left side. We have just added and subtracted $E[Y_i(0)|T = 1]$ on the right side, so the equality holds.

(1) $E[Y_i(1)|\, T = 1] - E[Y_i(0)\,|\mathrm{T} = 0] \;=\; \{E[Y_i(1)|\mathrm{T} = 1] - E[Y_i(0)|T = 1]\} \;+\; \{E[Y_i(0)|\mathrm{T} = 1] - E[Y_i(0)|\mathrm{T} = 0]\}$

The left hand side of equation (1) is the average outcome in the treated group, minus the average outcome in the control group. The first term in curly brackets {} on the right side of equation (1)

is the ATT. The ATT term cannot be directly estimated from the data because we do not observe $Y_i(0)$ for treated observations. The second term on the right side of equation (1) is the difference in the outcome variable for the individuals in the treated versus control group, if no treatment were to take place. This term captures fundamental differences between the individuals in the two groups, differences that are not caused by the treatment. It is the *selection bias*. The defining characteristic of random experiments is that assignment into the treatment or control group is random (e.g., it does not depend on the outcome variable or characteristics affecting the outcome variable). In a true experiment, because we randomly assign individuals to the treatment or control groups, we expect the individuals in the two groups to be comparable, and more so when we have larger samples. The fact that treatment assignment is random ensures that the selection bias is equal to zero. If the selection bias is equal to zero, we can estimate the ATT by simply computing the difference in means of the outcome variable for the treated observations versus the control observations.

When using observational data, rather than experimental data, to evaluate the effects of policy choices we have to acknowledge that individuals may choose whether to be treated or not. Self-selection into treatment will complicate our identification of the treatment effect, and we can no longer simply compare the means of the treated versus control groups.

If we cannot conduct a true experiment, what do we do? The rest of this article will provide an introduction to methods to recover a meaningful treatment effect estimate from observational data and highlight the two overarching approaches to combating selection bias. The first approach, discussed in the next section, is to "control for" the fundamental differences between the individuals in the treated versus control group. The second approach is to find an explicit source of random variation in treatment assignment, and use it to study the causal effect of the treatment.

2 Controlling For Selection Bias: Matching and Regression

How effective is exercise at reducing the risk of high blood pressure? Both genetics and lifestyle choices may complicate analysis of observational data on the relation between exercise and hypertension. Studying siblings who share a genetic background (but who exercise

differently) would help to isolate the effects of exercise from the effects of genetics. But siblings would be different ages, and age might be related to hypertension and/or exercise choices. Fraternal twins would be better to study, because they come from the same parents and are the same age. Identical twins would be better still.

In this section we discuss approaches to "control for" the fact that individuals in the treatment and control groups have different characteristics (e.g., different genetic background, different work environment, different corporate culture). One approach to deal with such heterogeneity is to identify a surrogate "sibling" or "twin" by *matching* on characteristics (e.g., an individual's education, gender, marital status; a corporation's size, state of incorporation, listing status) that might complicate analysis of a policy choice. The approach taken by matching methods is to match each treated individual to a control individual with similar characteristics, and to ignore individuals for which an acceptable match is not found. By virtue of requiring a close match, the treated and control pairs will be comparable along the dimensions used for matching. Some important conditions must be satisfied for these techniques to provide valid causal inference. Foremost among these conditions is the *conditional independence* assumption. Conditional independence means that, conditional on some characteristics, X, that we might control for or match on (like an individual's education, gender, marital status), treatment status (whether T=0 or T=1) is independent of the potential outcomes.

Recall our coffee-stress example and, for now, assume that reverse causality (i.e., the possibility that stress causes people to drink coffee) is not an issue. We still have a risk of omitted variable bias, such as if a person's workplace might affect both their coffee consumption and their stress level. Variables that are correlated with both the treatment (coffee consumption) and the outcome (stress) are only a problem if we can't control for them. In this case, we could control for the effect of workplace on coffee consumption by comparing coffee drinkers to coffee avoiders who work in the same office. The conditional independence assumption states, in essence, that there are no omitted variables. This assumption is bold. For instance, an individual's personal motivation may be correlated with both their coffee consumption, and their stress level, and yet, is not easy to control for in our comparison. The conditional independence assumption is important and recurring throughout this article. Stated formally, the Conditional Independence Assumption is written as follows: $\big(Y(1), Y(0)\big) \perp T \mid X$.

The second assumption we must make is that the treated and control individuals are similar in the sense that they have a comparable probability of receiving treatment. Or in statistical jargon, treated and control observations have overlapping values of the propensity score, i.e., the probability of being treated, conditional on observable characteristics X. Less formally, we are comparing treated apples to control apples, not comparing treated apples to control oranges.

Characteristic Matching:

There are different ways to identify or create matches (pseudo-twins). In this section we discuss two: characteristic matching and propensity score matching. Characteristic matching involves matching an individual from the treated group to an individual from the control group based on particular observable characteristics that we think are important determinants of the outcome variable. We will then use the control individual's values of the outcome variable as our approximation for the treated individual's counterfactual outcome. We can then compute an estimate of the ATT by taking the average treatment effect for the treated individuals. An estimate of the ATE can be obtained by matching both the treated to the controls, and the controls to the treated, and then taking the average of the differences between the treated and control observations in each pair. The general process of matching observations, computing the difference between the treated and matched control, and averaging, is the same across different matching methodologies.

To carry out characteristic matching in the coffee-stress example, we could require potential matches to work in the same office, be the same gender, and be in the same age bracket (e.g. 31-35, 36-40, and so forth). Then, if there are multiple potential matches satisfying these criteria, we could choose from among them the individual who works the most similar number of hours per week. A low-coffee individual, matching in all these dimensions, might give us a convincing counterfactual for the high-coffee individual we are matching to. If we are convinced this is an appropriate counterfactual, we can consider the difference in stress levels of the two individuals as the treatment effect of coffee consumption. We can repeat this procedure for all treated individuals and compute an estimate of the ATT, or for all individuals and compute an estimate of the ATE.

A modification of this approach is to select the characteristics that you deem important, and to use a distance metric to select the nearest match for a given individual. Suppose that to evaluate the coffee-stress hypothesis you have three important characteristics you want to match on: age, hours worked per week, and hours spent exercising per month. You deem all three characteristics to be equally important. To select the best match among the potential controls, you choose the one that is "closest," such as selecting the control individual that is nearest in terms of Euclidean distance in this three dimensional space.[6] If the dimensions on which you are matching are not of comparable magnitudes across individuals (hours worked may have less variation than hours exercised), a common alternative is to use as a match the observation that minimizes the sum of the squared percentage differences in the matching characteristics. Regardless of the exact function one uses to determine which control observation is the best match, the underlying motivation is to select the match that is the closest over all of the matching dimensions combined.

Propensity Score Matching:

Characteristic matching is appealing, but if we try to match on *each* of numerous characteristics we run into the "curse of dimensionality." The curse is that we will obtain fewer and fewer viable matches as we increase the number of criteria on which we match. For example, consider how our pool of potential matching controls decreases as we add, incrementally, a requirement that we match on age, weight, height, gender, marital status, resting heart rate, and mother's maiden name. Although the quality of each match we find is good, there will be many treated observations that have no acceptable matching counterpart. This loss of observations will decrease the statistical power of our tests and the generality of our results. However, if we match on too few dimensions, it will increase the possibility of omitted variable bias. The propensity score matching method is designed to mitigate the curse of dimensionality by allowing us to use information on many characteristics in our matching process, without drastically reducing the number of matches.

[6] Euclidean distance between two points in X-Y-Z space is $\text{sqrt}((X_{\text{treated}} - X_{\text{control}})^2 + (Y_{\text{treated}} - Y_{\text{control}})^2 + (Z_{\text{treated}} - Z_{\text{control}})^2)$.

The propensity score, described more formally below, is the estimated probability that an individual will be treated based on their observable characteristics.[7] The intuition underlying propensity score matching is that if we compare treated individuals to untreated individuals that were, in an ex ante sense, equally likely to be treated based on their characteristics, then the treatment assignment can be considered "as good as random," giving us an experimental design free of selection bias. For our hedging example, we would decide on the characteristics that make an oil company likely to have a hedging program (e.g., company size, company age, whether the CFO has an MBA). We would then estimate the propensity score for each company. Some companies may be, predictably, *likely* to have a hedging program. Of these, some will indeed have hedging programs (T=1), whereas others, though the type of firm expected to hedge, will not (T=0). Like twins separated at birth, these two firms will be a good match, because they are equally *likely* to be treated (to hedge) based on their characteristics, but one is treated and its twin is not. Likewise, some companies may be, predictably, *unlikely* to have a hedging program and of these, some will, surprisingly, have hedging programs (T=1), whereas others, as expected, will not (T=0). These two firms will also be a good match for each other.

Formally, the propensity score matching process is as follows.[8]

1. Estimate the propensity score using a probit or logit regression of the form: $T = \alpha + \beta' X + \epsilon$.
2. Match each individual to the individual of the opposite treatment status with the closest predicted value of the propensity score.
3. Check that the matched pairs have similar values of the characteristics, X, that you used to model the propensity score. This is known as checking for covariate balance.

[7] Rosenbaum and Rubin (1983) show that if we make the conditional independence assumption required for characteristic matching, then we need only match on the propensity score, not on all of the characteristics, in order to alleviate selection bias. Rosenbaum and Rubin show that independence conditional on X implies independence conditional on p(X). Mathematically, they show: $\big(Y(1), Y(0)\big) \perp T \mid X \;\Rightarrow\; \big(Y(1), Y(0)\big) \perp T \mid p(X)$.

[8] The process outlined below is called "nearest neighbor propensity score matching with replacement" because it selects as the match the single nearest individual with the opposite treatment status. This process is done "with replacement" because after using an observation as a match, we do not exclude it from the pool of potential matches for the next individual. There are also versions of propensity score matching that use several nearest neighbors, or weighted averages of many neighbors.

4. Compute the differences between the treated and untreated individuals in each matched pair. The average difference for the full sample is an estimate of the ATE. The average difference using only the treated individuals and their matched controls is an estimate of the ATT.

Regression:

Other than a naïve comparison of treated outcomes to control outcomes, regression is the simplest way to evaluate a treatment effect. Although the simplicity of regression is appealing, the approach also has some severe limitations. The goal of the next few paragraphs is to provide a reader who is unfamiliar with regression with enough information to understand regression's strengths, weaknesses, its role in causal inference, and to follow the upcoming discussions of differences in differences and instrumental variables. (Regression experts can safely skip to the next section without loss of continuity.)

A regression equation looks like: $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. *Y* is the outcome variable whose variation we are trying to explain. The *X*s are explanatory variables that we use to explain variation in *Y*. We can use many *X*s if necessary. The βs are coefficients. They tell us how much we expect Y to change if we increase that *X* by one unit. The parameter α is the intercept; its value tells us what we would expect *Y* to be if all of the *X*s were equal to 0. And ϵ is the error term, the variation in *Y* unexplained by the estimated model.

When we specify a regression equation, we are assuming that *Y* is a linear function of the *X*s we choose. A regression equation represents a conditional expectation—it tells us what we expect *Y* to be conditional on knowing the values of the *X*s. Regression models are appealing because, among other reasons:

1. The true conditional expectation function may not be linear, but in these cases regression still provides us with the best linear approximation to the true function.
2. It is convenient to think of relationships as linear.
3. Regression is relatively easy to understand and implement.

Continuous variables take many values (e.g. height, weight, age). Indicator variables take a value of 0 or 1. For instance, the treatment variable *T* we discussed earlier is an indicator variable equal to 1 if the individual is treated and 0 if not. The interpretation of regression coefficients, the βs, is straightforward: they tell us what the effect on *Y* would be if we increased that *X* by one unit, while holding all other *X*s constant. For continuous variables this tells us, for instance, what the average effect of growing one inch, gaining one pound, or aging one year is on the outcome variable *Y*. For indicator variables, the coefficient tells us the difference in the mean for the group with indicator=1 compared to the mean for the group with indicator=0, after controlling for the effect of the other *X*s.

The pivotal assumption that regression relies on to identify the effects of the *X*s on *Y* is a version of the conditional independence assumption discussed earlier. In the regression setting, this assumption is called the *exogeneity assumption*, specifically, that the variation in the *X*s is, in effect, randomly assigned. Stated formally, we need to make the very strong assumption that $E[\epsilon|X] = 0$. This condition is only valid when there is no reverse causality or omitted variable bias, and allows a coefficient to be interpreted as the causal effect of a unit increase in *X* on the outcome *Y*.

In the section on characteristic matching we discussed how we might match on age, sex, and place of employment in an attempt to test the hypothesis that coffee consumption increases stress levels. We can analogously control for these characteristics in a regression setting.

Consider the regression: $Stress = \alpha + \beta_1 I_{Treated} + \beta_2 Age + \beta_3 I_{Male} + \beta_4 I_{EnergyTrader} + \epsilon$, where the notation $I_{Description}$ denotes a variable that equals either one (if in the group described) or zero. Under the exogeneity assumption, the coefficient β_1 in this regression tells us the causal effect of the high-coffee treatment on stress levels if we are willing to make the assumption that this treatment is randomly assigned conditional on age, gender, and workplace.

Regression generally does not require the analog of a "good" match or overlap between treated and controls, allowing the policy makers to extrapolate from existing data to hypothetical situations (what would be the effect of coffee consumption for someone who is 3 years old or 103 years old?) nor does it have the dimensionality problem that matching techniques have.

These characteristics are both strengths and weaknesses of the approach. They are strengths because regression provides *external validity*: we can generalize from our sample of individuals or firms to an external sample not well-represented with our observed data, and infer what would happen in that other setting.

But all causal inference is, to some extent, local. As we generalize away from a narrow and carefully chosen setting, we lose *internal validity*: the ability to conclude that, indeed, A causes B (*and* not that B is caused by A *and also* not that A and B are co-determined by C). Regression often violates this notion of local-ness, whereas matching methods generally do not.

3 Causal Inference Methods Using Explicit Sources of Exogenous Variation

It is clear that the key to identifying the causal effect of a treatment is to have random or as-good-as-random variation in the treatment assignment because random treatment assignment solves the selection bias problem. The previous section discussed methods to control for the selection bias that arises when using observational data, but those methods require the conditional independence assumption to hold: treatment assignment must be random conditional on the characteristics—X_1, X_2, X_3, etc.—that we control for through matching or regression.

This section will discuss causal inference techniques—natural experiments, instrumental variables, and regression discontinuity designs—that use explicit sources of exogenous variation. The strength of these techniques is that they enable us to identify causal effects using more plausible assumptions than conditional independence. The disadvantage of these techniques is that it often requires substantial creativity to recognize the sources of exogenous variation to utilize.

Natural Experiments:

Suppose that poor weather in Colombia, civil unrest in Guatemala, and/or politically inspired export tariffs in Sumatra decrease the world supply of coffee beans, thereby raising the price of coffee substantially. Perhaps this increase in coffee prices would induce some coffee drinkers to reduce their intake, not because of any stress-related reasons, but simply for

budgetary ones. Treatment—whether an individual drinks coffee—changes, but for a reason that is exogenous, or outside of the system we are studying. The concept of a *natural experiment* is to use a shock to the economic environment that affects some individuals, but not others, to study the effect of the shock on outcomes. Natural experiments can provide a source of exogenous variation—a change in the way treatment, T=1 or T=0, is assigned that is independent of the outcome, Y, or characteristics, X—that allows us to make a causal inference.

Similar to random experiments, natural experiments create a treated group (affected by the shock), and a control group (unaffected by the shock). Examples of shocks include: court rulings or law changes that affect companies in one juridiction but not others, regulatory changes that apply to some companies in an industry but not others, natural disasters that affect economic conditions in a city/state but not other geographic regions, etc. The more arbitrary the mechanism for determining which individuals are affected by the shock, the better. As the assignment to affected/unaffected (i.e. treatment/control) becomes more random, the natural experiment design becomes closer to a random experiment.

Natural experiments can be powerful tools to help us answer causal questions, but the challenge facing the researcher is to find viable natural experiments that help answer the question at hand. But we generally cannot change weather patterns, tariffs, or civil unrest to have an ideal source of variation in our observational data. Let's return to our two examples and discuss the type of events that would provide natural experiments to test our hypotheses.

Coffee/Stress Example:

In order to study the effect of coffee consumption on stress levels, we need exogenous variation in coffee consumption (the treatment). Therefore, for a natural experiment, we need a shock that either increases or decreases coffee consumption for a subset of individuals. For instance, if the person who purchases coffee supplies for the office quits, and the office has no coffee for a week, we could use this as an exogenous decrease in the coffee consumption of the coffee-drinkers. We can then compare the stress levels of these employees to those of employees who do not drink coffee and are unaffected by the shock. As long as employees in the two groups are comparable ex ante, the assignment to the treatment and control groups is fairly random, allowing us to identify the causal effect of coffee consumption on stress.

Oil Company Derivative Hedging Example:

In order to study whether derivative hedging increases firm value, we need exogenous variation in the ability of oil companies to hedge. A hypothetical example of a shock that could generate this variation is a regulatory change that decreases the cost of hedging for oil companies incorporated in a certain state, perhaps through relaxing some state-mandated reporting requirements. Assuming that this new policy increases the hedging activity of the treated companies, we can then study the effect of hedging on firm value by comparing the values of the treated firms to the untreated. This approach is invalid if the companies in the treatment state are fundamentally different than the control companies. Acknowledging that the treatment and control groups generated by natural experiments are not always comparable leads us into our discussion of differences-in-differences (DiD).

Differences-in-Differences:

The differences-in-differences methodology is the common approach to dealing with the fact that the treatment assignment in natural experiments is not completely random. The differences-in-differences method compares the treated group's *difference* in the outcome variable from before to after treatment to the *difference* experienced by the control group over the same period. The first difference is the outcome after treatment minus the outcome before treatment. We compute this difference for each group—treated and control—separately. The second difference is the first difference for the treatment group, minus the first difference for the control group. The difference-in-differences is a single number that tells us how much the treatment group changed in excess of how much the control group changed over the same period. The difference-in-differences is an estimate of the ATT.

The differences-in-differences method provides a good estimate of the causal effect of treatment when a few conditions are met. Foremost is that the treated and control groups are similar, in the sense that the outcomes for each would be expected to evolve in the same way, except for the fact that one group is assigned treatment due to the exogenous shock. In the parlance of econometricians, the treated and control groups have parallel trends prior to the

shock. In other words, if there were no treatment, the first differences of the two groups would be the same.

Consider how we could use the differences-in-differences methodology with the proposed natural experiment for our coffee/stress example — the employee in charge of purchasing coffee for the office of the energy traders is out sick for one week, cutting off the office's coffee supply. Recall that the treatment group is the traders working in the office who drink coffee and the control group is those who do not. In order to test whether coffee causes stress, we need data on the stress levels of the traders, perhaps from a weekly survey that assigns employees a stress score from 1 to 10. To implement the differences-in-differences method, we need data on stress levels for the week prior to the coffee shortage, and the week of the shortage. We compute the average stress level for both the treated and control groups during the prior week, and during the coffee-shortage week. Then we compute the first difference for each group as the second week's average stress level minus the first week's average stress level. Finally, we compute the difference-in-differences as the change in average stress level for the treated group (coffee drinkers) minus the change for the control group (coffee avoiders). The difference-in-differences gives us an estimate of the causal effect of the treatment (decreased coffee consumption) on stress levels.

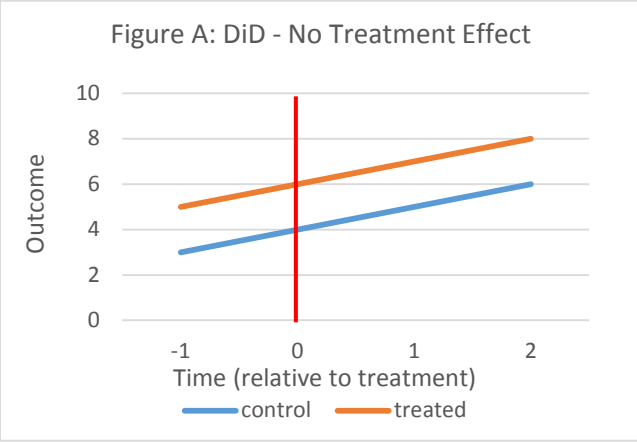
The differences-in-differences method has several advantages compared to documenting a correlation between coffee consumption and stress levels. For instance, if traders who drink coffee are just more stressed in general than traders who do not (a time-invariant difference between the groups), taking the first difference will prevent this from affecting the difference-in-differences estimate. Let's say average stress levels for the coffee drinkers during the week prior to, and the week of the coffee-shortage were 8 and 6 respectively, whereas, the stress levels for traders who don't drink coffee were 4 and 2 during the same weeks. Clearly there is a difference in the general level of stress felt by the two groups, but it is unaffected by reducing the coffee consumption of coffee drinkers. Therefore, the difference in stress levels is unlikely to be *caused* by coffee consumption. This pattern in the data would generate a strong correlation between coffee consumption and stress. However, the differences-in-differences method will account for these time-invariant differences, and produce an accurate estimate of the treatment effect $(6-8) - (2-4) = 0$.

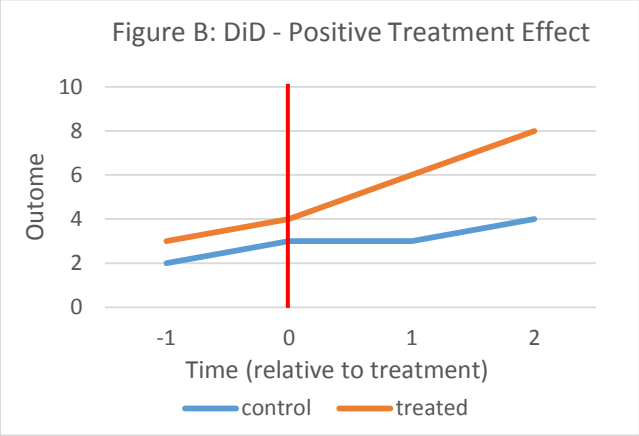
A second advantage of the differences-in-differences methodology is that the results are not confounded by a time trend in the data that is common to all individuals. Using our hypothetical coffee data, we can see that there is a common time trend because average stress levels decreased by two points for both groups from the first to second week in our sample (perhaps due to a favorable change in the world price of oil). Taking the second difference is the reason that common time trends do not confound the analysis. If we only examined data for the treated group, we would find a positive correlation between their coffee consumption and stress levels. However, the differences-in-differences method accounts for the fact that the control group also experienced this decrease in stress over the same period, suggesting it was not caused by the treatment, but by other factors (e.g., a change in oil prices). As before, the difference-in-differences is equal to zero ((6-8) – (2-4) = 0), showing that this common time trend does not affect the estimate of the ATT. The fact that neither time-invariant differences between the treatment and control groups, nor a time trend common to all individuals confound the results of the differences-in-differences method makes it a powerful tool for causal inference.

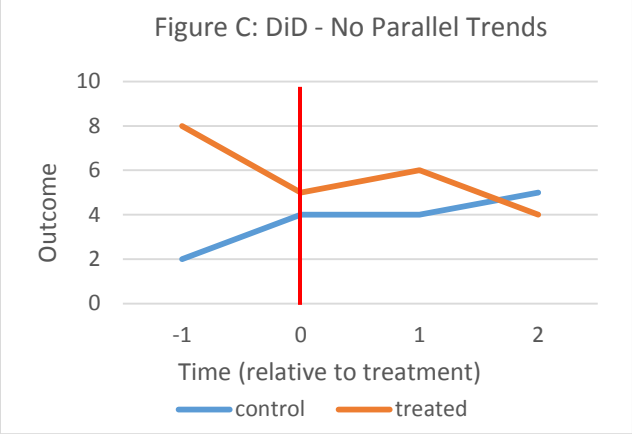
A shortcoming of the differences-in-differences method is that its results can be confounded by a difference between the treatment and control groups that causes the groups to have different time trends for the outcome variable. In order for the results to be valid, we must assume that if there were no treatment, the two groups would have parallel trends. If the two groups' trends in the outcome variable are parallel in the pre-treatment period, it lends credibility to this key assumption differences-in-differences relies on. On the other hand, if the trends in the outcome variable are *not* parallel in the pre-period, it casts serious doubt on the validity of the results.

The three figures below present hypothetical paths of the outcome variable average for treated and control groups around a shock (time periods -1 and 0 are pre-treatment, whereas periods 1 and 2 are post-treatment.) Figure A is an example of a differences-in-differences design that has the parallel pre-treatment trends necessary for valid inference and that will find no treatment effect. Figure B is an example that has parallel pre-treatment trends and will find a positive treatment effect. Figure C is an example of a differences-in-differences design that violates the requisite parallel trends assumption, and should not be trusted. Non-parallel pre-treatment trends like those in Figure C often arise when the treatment and control groups are not

comparable, for instance if we used yoga instructors as a control group for energy traders in the test of our coffee/stress hypothesis.







This table presents the computation of the difference-in-differences for Figures A, B, and C above. In columns 1 and 2, the first difference (FD) is computed for the treated and control groups by taking the average of the post-treatment values minus the average of the pre-treatment values. Column 3 computes the estimate of the treatment effect for each figure (the difference-in-differences).

	Treated FD	Control FD	Treatment Effect (DiD)
Figure A	(7.5-5.5)	(5.5-3.5)	(7.5-5.5) - (5.5-3.5) = 0
Figure B	(7-3.5)	(3.5-2.5)	(7-3.5) - (3.5-2.5) = 2.5
Figure C	(5-6.5)	(4.5-3)	(5-6.5) - (4.5-3) = -3

Instrumental Variables:

Let us return to the regression we discussed earlier to test our hypothesis that coffee consumption causes increased stress:

$$Stress = \alpha + \beta_1 I_{Treated} + \beta_2 Age + \beta_3 I_{Male} + \beta_4 I_{TradingDesk} + \epsilon.$$

In this regression, the coefficient β_1 tells us the causal effect of being a high coffee individual on stress, as long as the conditional independence assumption is satisfied (i.e. being a high-coffee individual is random among people of similar age, sex, and workplace). This assumption, however, is unlikely to hold and our variable of interest ($I_{Treated}$) is likely endogenous for at least two reasons. First, there is a potential omitted variable bias: it is possible that being a coffee drinker is correlated with personal motivation, which is unobserved and likely correlated with stress. Second, reverse causality is a potential issue: it is possible that increased stress causes people to drink more coffee, perhaps because the stress is making them tired. It is clear that omitted variable bias and/or reverse causality could confound our estimate of β_1. Instrumental variables is a method designed to address endogeneity concerns like these and provide an accurate estimate of β_1.

The key idea behind instrumental variables is that although the explanatory variable of interest ($I_{Treated}$) may be endogenous, perhaps there is some portion of its variation that is exogenous. For instance, an individual's coffee consumption is likely determined by their unobserved characteristics like personality type and/or personal motivation (endogenous to stress levels), by their current stress levels (endogenous), and by the price of coffee (exogenous to stress levels). The instrumental variables approach seeks to isolate the variation in the explanatory variable that is exogenous and use it to study the variable's effect on outcomes. In order to accomplish this, the instrumental variables approach requires the researcher to find a variable (e.g., world price of coffee) that induces variation in the explanatory variable (coffee consumption) for reasons unrelated to the outcome variable (stress levels). This source of exogenous variation is called an instrumental variable, or an instrument.

Good instruments are often challenging to find. In order for the world price of coffee to induce exogenous variation in coffee consumption, the price of coffee must be unrelated to stress levels, except through its effect on coffee consumption. Does an increase in coffee prices have implications for average stress levels of people? If the answer is "no, not directly" then the world price of coffee is potentially a valid instrument, because the only way we expect it to affect stress levels is if coffee consumption affects stress levels, and the price of coffee affects stress levels through this specific channel.[9]

Formalizing the previous discussion, there are two requirements in order for an instrument to be valid. First, the instrument must cause variation in the endogenous variable of interest. This requirement is known as the relevance condition, and can be tested by checking that the endogenous variable of interest and the instrument are correlated in the data. Second, the instrument itself must be exogenous. This second condition is known as the exclusion restriction, and as with any claim of conditional independence/exogeneity, it is inherently untestable. The challenge of using the instrumental variables method is to find instruments for which you can make a convincing logical argument for exogeneity. The mathematical statements of the two conditions are given below (the instrumental variable is labeled Z).

[9] The instrument may fail the exclusion restriction (see below) if the subjects of the stress-coffee experiment are coffee farmers or coffee roasters.

1. Relevance Condition: The instrumental variable, Z, must be correlated with the endogenous variable, X: $Cov(X, Z) \neq 0$

   Coffee Example: coffee consumption needs to be correlated with the instrument (world price of coffee)

2. Exclusion Restriction: The instrumental variable, Z, is unrelated to the outcome variable, Y, except through the instrument's impact on the endogenous variable, X: $Cov(Z, \epsilon) = 0$ (i.e. Z is exogenous)

   Coffee Example: the only way that the world price of coffee could affect stress is that it decreases coffee consumption, and coffee consumption might affect stress

Given our valid instrument (world price of coffee), we are prepared to use the instrumental variables method to estimate the causal effect of coffee consumption on stress. The most common implementation of the instrumental variables method is through two-stage least squares (2SLS), which implements instrumental variables through two regressions. Most statistical packages can do this procedure automatically. In the first stage, we regress the endogenous explanatory variable (coffee consumption) on the instrument (world price of coffee) and any exogenous explanatory variables (age, gender, workplace). The predicted value of coffee consumption from this regression is the "clean/exogenous part" since its variation is driven purely by exogenous variables. We will use only this exogenous part of coffee consumption to examine its causal effect on stress levels. In order to examine this causal effect, we run the second stage regression. In the second stage we regress stress levels on the exogenous part of coffee consumption, and the other explanatory variables. The coefficient (β_1) on the predicted value of coffee consumption in the second stage regression is the instrumental variables estimate of coffee consumption's causal effect on stress levels.

Implementing the Instrumental Variables method in our coffee example:

Let $Coffee$ denote the number of cups consumed per day, $CoffeePrice$ the world price of coffee, and $\widehat{Coffee}$ the predicted value from the first stage regression.

First stage:

$$Coffee = \alpha + \lambda_1 CoffeePrice + \lambda_2 Age + \lambda_3 I_{Male} + \lambda_4 I_{EnergyTrader} + u$$

Second stage:

$$Stress = \alpha + \beta_1 \widehat{Coffee} + \beta_2 Age + \beta_3 I_{Male} + \beta_4 I_{EnergyTrader} + \epsilon$$

Implementing the IV/2SLS method in a general setting:

Let X denote the endogenous variable, $\hat{X}$ its predicted value, Z the instrumental variable, the vector A the exogenous explanatory variables, and Y the outcome variable.

First stage:

$X = \alpha + \lambda_1 Z + \lambda_A' A + u$ …the predicted value from this regression is $\hat{X}$

Second stage:

$Y = \alpha + \beta_1 \hat{X} + \beta'_A A + \epsilon$

The coefficient (β_1) on the predicted value of the endogenous explanatory variable is the estimate of the causal effect of X on Y. This instrumental variables coefficient estimate can be interpreted in a similar fashion to an ordinary least squares regression coefficient for a variable that is exogenous—it estimates the effect of a one unit increase in X on Y. The important difference is that the instrumental variables method can accurately estimate this effect even when X is endogenous, so long as the instrument Z is truly exogenous. By contrast, the coefficient of an ordinary least squares regression of Y on an endogenous X reflects not only the causal effect of X on Y, but also the reverse causality effects (how Y causes changes in X) and any omitted variables bias.

Regression Discontinuity Designs:

Suppose government regulations are important to your industry and you want to assess whether investing in lobbying efforts and political donations to the incumbent political party will

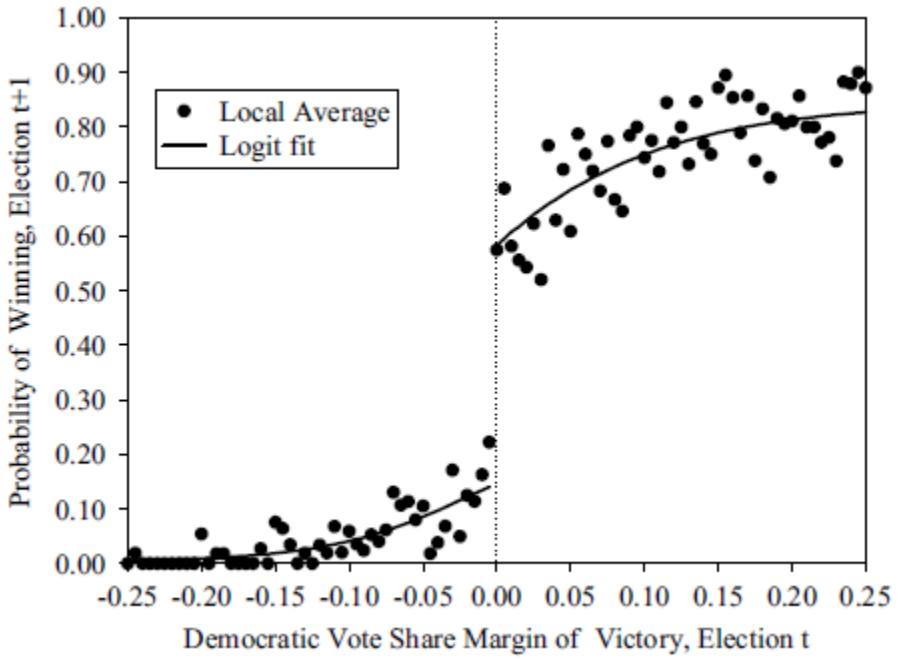
have a long term payoff. To estimate the costs and benefits, it would be useful to know the likelihood the incumbent party will stay in power. One way of evaluating the treatment effect of incumbency on subsequent election outcomes is to start with considering the elections that are *close*. In a close election, candidates with small differences in the percentage of votes earned receive very different treatments—the winner takes public office, while the loser looks for a job.

The world we live in is full of laws, rules, and regulations, and the regression discontinuity approach exploits them. In many cases these rules feature an arbitrary cutoff or threshold for whom they affect. Political elections typically have a threshold of a 0% + one vote margin to be considered the victor. Recreationally, many theme park rides have rules requiring would-be riders to meet a certain height criterion (say 48 inches). This rule allows individuals who are 48 or 49 inches tall to experience the ride, but not those who are 46 or 47 inches tall.

The regression discontinuity approach to causal inference takes advantage of the arbitrary nature of thresholds imbedded in rules. Any regression discontinuity research design begins with a threshold for which individuals on one side are treated, while individuals on the other side are not. The basic idea of the regression discontinuity method is that for individuals near the threshold, it is "as good as random" on which side of the cutoff they end up. Compelling economic or political conditions did not lead to a landslide; instead, the outcome of a very close election is, all things considered, almost like being decided on a coin flip. The threshold itself gives us some exogenous variation in which observations get treated that we can use to study the causal effect of the treatment on outcomes.

We will discuss regression discontinuity designs using the results from Lee (2008), a paper that documents the advantage of being the incumbent party in elections to the U.S. House of Representatives. The first requirement of the method is that we have a *forcing variable* that can take many possible values. The forcing variable is the variable that the threshold is based on (percentage of votes relative to the other political party, height of the roller coaster patron). The second requirement is that there is a threshold that determines whether subjects receive the treatment (holding political office, riding the roller coaster). Lee (2008) examines the effect of incumbency on the probability of the party winning the next election for the same seat in the U.S. House of Representatives. Essentially, the paper compares the chances of a political party winning the next election when they win the current election by a narrow margin, versus when

they barely lose the current election. Because the standing of political parties is likely similar in districts where they barely won/lost the previous lection, the difference in future election outcomes can be attributed to the causal effect of being the incumbent. An attractive feature of the regression discontinuity approach is that the main result can be displayed in a single picture.



The horizontal axis of the graph is the forcing variable—the difference in the percentage of votes between the Democratic Party and the Republican Party. The vertical axis of the graph is the outcome variable—the percentage of the time the Democratic Party wins the next election for that political office. The dots on the graph are local averages of the outcome variable. For instance, the first dot to the right of the threshold is the average of the indicator variable for whether the Democratic Party wins the next election, given that the Democratic Party won the current election by less than half a percent. This average of the indicator function tells us the percentage of times that a Democrat won the next election in these close-win cases, approximately 60% compared to the close losses, where Democrats win only approximately 25%.

Given a forcing variable, like margin of victory or roller coaster patron height, a threshold that determines treatment (0% margin + one vote, 48 inches tall), and an outcome variable, how should we determine the treatment effect? A direct approach might be to take the average of the outcome variable for individuals close to, but on opposite sides of the threshold

(i.e. 60% - 25% = 35% advantage for the incumbent). Pointing out that for these individuals treatment assignment is quasi-random, we could take the simple difference between the treated group and untreated group as the treatment effect.

We can examine the characteristics of observations on either side of the threshold and check that they are similar and that the differences on either side really are arbitrary: nearly treated apples being compared to nearly control apples. One concern is that individuals are "gaming the threshold," that is, they are aware of the threshold value and can manipulate their value of the forcing variable with enough precision to determine which side of the threshold they fall on, like standing on tiptoes before a roller coaster ride or stuffing a ballot box. If this manipulation occurs, individuals are choosing to be treated or not, and we should expect selection bias to confound our results. Evidence that units on both sides of the threshold are comparable in terms of characteristics, and that there are not dramatically more units just to one side of the threshold would provide confidence in the experimental design.

Synthetic Controls:

An alternative way to create a good counterfactual is *synthetic controls*. The synthetic control approach is like propensity score matching, but instead of matching to one control observation, you match to several, each with potentially different weights. Like propensity score matching, you determine a set of characteristics that are likely to explain the outcome of interest, Y. You identify a *donor pool* of control observations (T=0) to be used as potential matches. Then, you determine what combination of observations from the donor pool, when matched on characteristics, looks most like the treated observation (T=1) in terms of the outcomes before treatment. Thus, a logistical limitation of the approach is that it requires a long time series for both the treated observation and the donor pool observations. In essence, you are forming a portfolio from the donor pool that tracks the pre-treatment outcome for the treated observation. Or, loosely, you are taking bits of control group oranges and putting them together to resemble a treated apple.

A recent paper, Berger, et al. (2015), uses the synthetic controls method to examine the effect of intrastate bank branching deregulation on economic growth. Prior to the 1970s most

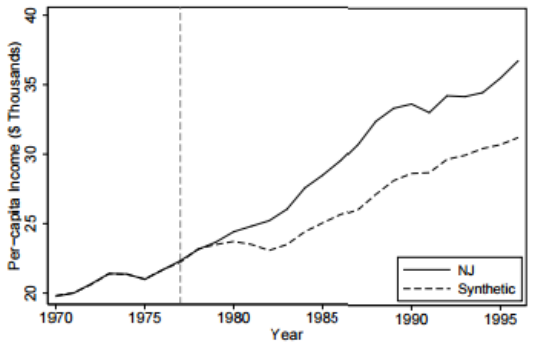
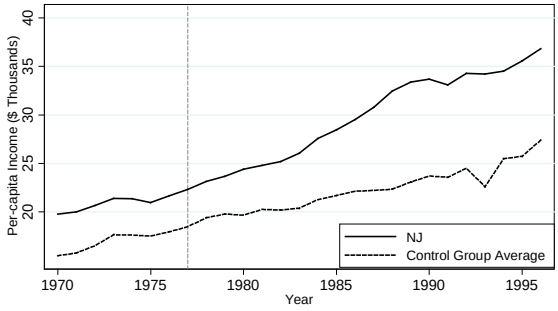
states had laws that heavily restricted banks' ability to acquire competing banks and open new branches throughout the state. Throughout the 1970s, 80s, and 90s almost all states deregulated intrastate banking, leading to increased competition among local banks. Previous academic articles document that deregulation has a positive effect on economic growth (measured in per-capita income) using a differences-in-differences design to compare deregulating states (treated) to states that had not yet deregulated (controls).

A simple differences-in-differences methodology calculates the average treatment effect across all treated units. This method is appropriate when we expect the treatment effect to be similar for all treated units. However, if there is significant heterogeneity in the treatment effect across units, then we would like to know for which type of unit the treatment had a large effect, and for which type it had no effect. The synthetic controls method is conducive to examining heterogeneous treatment effects because it constructs unique synthetic controls/ counterfactual outcomes for each treated unit, and the difference between the observed outcome and the synthetic control is an estimate of the treatment effect for that unit. The Berger, et al. (2015) paper finds that deregulation had a positive effect on economic growth, but only in states where bank consolidation was most aggressive and where the credit supply increased the most as a result.

Figure A below from Berger, et al. (2015) presents the observed economic growth for New Jersey and the growth of the state's synthetic control in the years surrounding New Jersey's deregulation. A key advantage of the method is that *if* the synthetic version is able to closely track the treated unit over a long pre-treatment period, it is unlikely that the two entities differ on important observable or unobservable dimensions. As we can see in Figure A, the synthetic version of New Jersey[10] tracks the real New Jersey's economic growth very closely prior to deregulation. Following deregulation, the real New Jersey experiences greater economic growth than the synthetic version, suggesting that deregulation had a positive effect on economic growth. Figure B presents the economic growth for New Jersey versus the simple average of the states in the donor pool (non-deregulating states). This simple average is the counterfactual outcome that the differences-in-differences method implicitly uses. In the case of New Jersey,

[10] The synthetic version of New Jersey is a portfolio containing the District of Columbia, Florida, Hawaii, Michigan, North Dakota, and Nevada.

the pre-treatment parallel trends condition appears to be satisfied, i.e. the control group average provides a reasonable counterfactual for the differences-in-differences method. However, it is clear from a comparison of Figure A and B that the synthetic controls method does an even better job of matching the pre-treatment trend of economic growth in New Jersey. In studies with a small number of treated units whose counterfactuals need to be very precisely estimated, the payoff to using synthetic controls to construct counterfactuals can be large.

Figure A	Figure B
	

4 Concluding Remarks

Many of the most interesting policy questions are of the form "if we change A, what will happen to B?" Credible answers to these questions *must* involve causal inference. As illustrated by our coffee/stress and hedging/firm value examples, correlation simply does not imply causation. Fundamental problems such as reverse causality and omitted variable bias can cause severe selection bias that can even reverse the direction of the relationships you find. First and foremost, the goal of this article is to equip you with the perspective to think critically about empirical findings before coming to the conclusion that "A causes B". The second purpose is to introduce you to some of the most widely used and powerful causal inference tools.

In order to conduct careful causal inference, you must first identify potential sources of selection bias, and then select a method that addresses the concerns while making the most plausible assumptions possible in your context. The two general approaches to addressing selection bias are to control for it (e.g. regression or matching), or to find sources of exogenous variation in the treatment (e.g. natural experiments, instrumental variables, etc.). The first set of

methods rely on the strong conditional independence/exogeneity assumption to do causal inference. This approach of "controlling" for observable differences in order to minimize the selection bias is clearly better than computing correlations. However, the tenuous assumptions are a weakness of these methods. The strength of regression and matching is that they are (relatively) straightforward to implement and do not require tremendous creativity on the part of the researcher. The second set of methods use explicit sources of exogenous variation, and while they still require assumptions, they are in most cases more plausible ones. Therefore, these methods are the center of many of the most convincing arguments, based on observational data, that A causes B. A drawback of these methods is that they require creativity on the part of the researcher, and the right situation in order to implement – identifying natural experiments, relevant discontinuities, and valid instruments is not always easy. From a pure causal inference perspective, however, this challenge is worth taking on.

References

Angrist, Joshua D., and Jörn-Steffen Pischke, (2008), Mostly Harmless Econometrics: An Empiricists' Companion, Princeton University Press, Princeton, NJ.

Angrist, Joshua D., and Jörn-Steffen Pischke, (2015), Mastering 'Metrics, Princeton University Press, Princeton, NJ.

Berger, Elizabeth Anne, Alexander W. Butler, Edwin Hu, and Morad Zekhnini, 2015. Credit be Dammed: The Impact of Banking Deregulation on Economic Growth. *Working Paper*. Link: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2139679

Lee, David S., 2008. Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics* 142, 675-697.

Wooldridge, Jeffrey, (2010), Econometric Analysis of Cross Section and Panel Data, vol. 1, 2 ed., The MIT Press.