10

Causal inference in practice

The previous chapter introduced the conceptual foundations of causality, but there's a lot more to learn about how these concepts play out in practice. In fact, there's a flourishing practice of causal inference in numerous scientific disciplines. Increasingly, ideas from machine learning show up in the design of causal estimators. Conversely, ideas from causal inference can help machine learning practitioners run better experiments.

In this chapter we focus on estimating the average treatment effect, often abbreviated as ATE, of a binary treatment T on an outcome variable Y:

$$\mathbb{E}[Y \mid \mathrm{do}(T := 1)] - \mathbb{E}[Y \mid \mathrm{do}(T := 0)]\,.$$

Causal effects are population quantities that involve two hypothetical actions, one holding the treatment variable constant at the treatment value 1, the other holding the treatment constant at its baseline value 0.

The central question in causal inference is how we can estimate causal quantities, such as the average treatment effect, from data.

Confounding between the outcome and treatment variable is the main impediment to causal inference from observational data. Recall that random variables Y and T are confounded, if the conditional probability distribution of Y given T does not equal its interventional counterpart:

$$\mathbb{P}\{Y = y \mid \mathrm{do}(T := t)\} \neq \mathbb{P}\{Y = y \mid T = t\}$$

If these expressions were equal, we could estimate the average treatment effect in a direct way by estimating the difference $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$ from samples. Confounding makes the estimation of treatment effects more challenging, and sometimes impossible. Note that the main challenge here is to arrive at an expression for the desired causal effect that is free of any causal constructs, such as the do-operator. Once we have a plain probability expression at hand, tools from statistics allow us to relate the population quantity with a finite sample estimate.

Design and inference

There are two important components to causal inference, one is *design*, the other is *inference*.

In short, design is about sorting out various substantive questions about the data generating process. Inference is about the statistical apparatus that we unleash on the data in order to estimate a desired causal effect.

Design requires us to decide on a population, a set of variables to include, and a precise question to ask. In this process we need to engage substantively with relevant scientific domain knowledge in order to understand what assumptions we can make about the data.

Design can only be successful if the assumptions we are able to make permit the estimation of the causal effect we're interested in. In particular, this is where we need to think carefully about potential sources of confounding and how to cope with them.

There is no way statistical estimators can recover from poor design. If the design does not permit causal inference, there is simply no way that a clever statistical trick could remedy the shortcoming. It's therefore apt to think of causal insights as consequences of the substantive assumptions that we can make about the data, rather than as products of sophisticated statistical ideas.

Hence, we emphasize design issues throughout this chapter and intentionally do not dwell on technical statements about rates of estimation. Such mathematical statements can be valuable, but design must take precedence.

Experimental and observational designs

Causal inference distinguishes between *experimental* and *observational* designs. Experimental designs generally are active in the sense of administering some treatment to some set of experimental units. Observational designs do not actively assign treatment, but rather aim to make it possible to identify causal effects from collected data without implementing any interventions.

The most common and well-established experimental design is a randomized controlled trial (RCT). The main idea is to assign treatment randomly. A randomly assigned treatment, by definition, is not influenced by any other variable. Hence, randomization eliminates any confounding bias between treatment and outcome.

In a typical implementation of a randomized controlled trial, subjects are randomly partitioned into a *treatment group* and a *control group*. The treatment group receives the treatment, the control group receives no treatment. It is important that subjects do not know

which group they were assigned to. Otherwise knowledge of their assignment may influence the outcome. To ensure this, subjects in the control group receive what is called a *placebo*, a device or procedure that looks indistinguishable from treatment to the study subject, but lacks the treatment ingredient whose causal powers are in question. Adequate placebos may not exist depending on what the treatment is, for example, in the case of a surgery.

Randomized controlled trials have a long history with many success stories. They've become an important source of scientific knowledge.

Sometimes randomized controlled trials are difficult, expensive, or impossible to administer. Treatment might be physically or legally impossible, too costly, or too dangerous. Nor are they free of issues and pitfalls.[1] In this chapter, we will see observational alternatives to randomized controlled trials. However, these are certainly not without their own set of difficulties and shortcomings.

[1] Deaton and Cartwright, "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science & Medicine* 210 (2018): 2–21.

The machine learning practitioner is likely to encounter randomization in the form of so-called *A/B tests*. In an A/B test we randomly assign one of two treatments to a set of individuals. Such experiments are common in the tech industry to find out which of two changes to a product leads to a better outcome.

The observational basics: adjustment and controls

For the remainder of the chapter we focus on observational causal inference methods. In the previous chapter we saw that there are multiple ways to cope with confounding between treatment and outcome. One of them is to adjust (or control) for the parents (i.e., direct causes) of T via the adjustment formula.

The extra variables that we adjust for are also called *controls*, and we take the phrase *controlling for* to mean the same thing as *adjusting for*.

We then saw that we could use any set of random variables satisfying the graphical backdoor criterion. This is helpful in cases where some direct causes are unobserved so that we cannot use them in the adjustment formula.

Let's generalize this idea even further and call a set of variables *admissible* if it satisfies the adjustment formula.

Definition 1. *We say that a discrete random variable X is* admissible *if it satisfies the adjustment formula:*

$$\mathbb{P}[Y = y \mid \mathrm{do}(T := t)] = \sum_x \mathbb{P}[Y = y \mid T = t, X = x]\, \mathbb{P}[X = x]$$

Here we sum over all values x in the support of X.

The definition directly suggests a basic estimator for the do-intervention:

1. Collect samples n samples $(t_i, y_i, x_i)_{i=1}^n$.
2. Estimate each of the conditional probabilities $\mathbb{P}[Y = y \mid T = t, X = x]$ from the collected samples.
3. Compute the weighted sum.

This estimator can only work if all slices $\{T = t, X = x\}$ have nonzero probability, an assumption often called *overlap* or *positivity* in causal inference.

But the basic estimator also fails when the adjustment variable X can take on too many possible values. In general, the variable X could correspond to a tuple of features, such as, age, height, weight, etc. The support of X grows exponentially with the number of features. This poses an obvious computational problem, but more importantly a statistical problem as well. By a counting argument some of the events $\{T = t, X = x\}$ must have probability as small as the inverse of size of the support X. To estimate a probability $p > 0$ from samples to within small relative error, we need about $O(1/p^2)$ samples.

Much work in causal inference deals with overcoming the statistical inefficiency of the basic estimator. Conceptually, however, most sophisticated estimators work from the same principle. We need to assume that we have an admissible variable X and that positivity holds. Different estimators then use this assumption in different ways.

Potential outcomes and ignorability

The average treatment effect often appears in the causal inference literature equivalently in its potential outcome notation $\mathbb{E}[Y_1 - Y_0]$. This way of going about it is mathematically equivalent and either way works for us.

When talking about potential outcomes, it's customary to replace the assumption that X is admissible with another essentially equivalent assumption called *ignorability* or *unconfoundedness*. To recall from the previous chapter, this assumption requires that the potential outcomes variables are conditionally independent of treatment given X. Formally, $T \perp (Y_0, Y_1) \mid X$. It's not hard to show that ignorability implies that X is admissible.

Reductions to model fitting

Adjustment gives a simple and general way to estimate causal effects given an admissible set of variables. The primary shortcoming that we discussed is the sample inefficiency of the formula in high-dimensional settings.

There's a vast literature of causal estimators that aim to address this central shortcoming in a range of different settings. While the landscape of causal estimators might seem daunting to newcomers, almost all causal inference methods share a fundamental idea. This idea is reduce causal inference to standard supervised machine learning tasks.

Let's see how this central idea plays out in a few important cases.

Propensity scores

Propensity scores are one popular way to cope with adjustment variables that have large support.

Let $T \in \{0, 1\}$ be a binary treatment variable. The quantity

$$e(x) = \mathbb{E}[T \mid X = x]$$

is known as the *propensity score* and gives the likelihood of treatment given in the subpopulation defined by the condition $X = x$.

Theorem 1. *Suppose that X is admissible, and the propensity scores are positive $e(x) \neq 0$ for all X. Then,*

$$\mathbb{E}[Y \mid \mathrm{do}(T := 1)] = \mathbb{E}\left[\frac{YT}{e(X)}\right]$$

Proof. Applying the adjustment formula, we have

$$\begin{aligned}
\mathbb{E}[Y \mid \mathrm{do}(T := 1)] &= \sum_y y \, \mathbb{P}[Y = y \mid \mathrm{do}(T := 1)] \\
&= \sum_y y \left(\sum_x \mathbb{P}[Y = y \mid T = 1, X = x] \, \mathbb{P}[X = x] \right) \\
&= \sum_y y \left(\sum_x \frac{\mathbb{P}[Y = y \mid T = 1, X = x] \, \mathbb{P}[X = x] \, \mathbb{P}[T = 1 \mid X = x]}{\mathbb{P}[T = 1 \mid X = x]} \right) \\
&= \sum_y y \left(\sum_x \frac{\mathbb{P}[Y = y, T = 1, X = x]}{\mathbb{P}[T = 1 \mid X = x]} \right) \\
&= \sum_y y \left(\sum_{x, t \in \{0,1\}} \frac{t \, \mathbb{P}[Y = y, T = t, X = x]}{\mathbb{P}[T = 1 \mid X = x]} \right) \\
&= \sum_{y, x, t \in \{0,1\}} \frac{yt \, \mathbb{P}[Y = y, T = t, X = x]}{\mathbb{P}[T = 1 \mid X = x]} \\
&= \mathbb{E}\left[\frac{YT}{e(X)}\right]
\end{aligned}$$

Here, we used the adjustment formula in the second line, then multiplied and divided by $e(x) = \mathbb{P}[T = 1 \mid X = x] \neq 0$ to obtain the third line.

□

The same theorem also shows that

$$\mathbb{E}[Y \mid \mathrm{do}(T := 0)] = \mathbb{E}\left[\frac{Y(1-T)}{1-e(X)}\right]$$

and thus the average treatment effect of X on Y is given by

$$\mathbb{E}[Y \mid \mathrm{do}(T := 1)] - \mathbb{E}[Y \mid \mathrm{do}(T := 0)] = \mathbb{E}\left[Y\left(\frac{T}{e(X)} - \frac{1-T}{1-e(X)}\right)\right].$$

This formula for the average treatment effect is called *inverse propensity score weighting*. Let's understand what it buys us compared with the adjustment formula when working with a finite sample.

One way to approximate the expectation given the theorem above is to collect many samples from which we estimate the propensity score $e(x)$ separately for each possible setting X. However, this way of going about it runs into the very same issues as the basic estimator. Practitioners therefore choose a different route.

In a first step, we fit a model $\widehat{e}$ to the propensity scores hoping that our model $\widehat{e}$ approximates the propensity score function e uniformly well. We approach this step as we would any other machine learning problem. We create a dataset of observations (x_i, e_i) where e_i is an empirical estimate of $e(x_i)$ that we compute from our sample. We then fit a model to these data points using our favorite statistical technique, be it logistic regression or something more sophisticated.

In a second step, we then use our model's estimated propensity scores in our sample estimate instead of the true propensity scores:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{t_i y_i}{\widehat{e}(x_i)}.$$

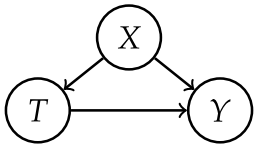
The appeal of this idea is that we can use the entire repertoire of model fitting to get a good function approximation of the propensity scores. Depending on what the features are we could use logistic regression, kernel methods, random forests, or even deep models. Effectively we're reducing the problem of causal inference to that of model fitting, which we know how to do.

Double machine learning

Our previous reduction to model fitting has a notable shortcoming. The propensity score estimate $\widehat{e}(x_i)$ appears in the denominator of our estimator. This has two consequences. First, unbiased estimates

of propensity scores do not imply an unbiased estimate of the causal effect. Second, when propensity scores are small and samples aren't too plentiful, this can lead to substantial variance.

There's a popular way to cope, called *double machine learning*, that works in a partially linear structural causal model:



$$Y = \tau T + g(X) + U$$
$$T = f(X) + V$$

In this model, the variable X is an observed confounder between treatment and outcome. We allow the functions g and f to be arbitrary, but note that g only depends on X but not on T as it could in general. The random variables U, V are independent exogenous noise variables with mean 0. In this model, the effect of treatment on the outcome is linear and the coefficient τ is the desired average treatment effect.

The trick behind double machine learning is to subtract $\mathbb{E}[Y \mid X]$ from each side of the first equation and to use the fact that $\mathbb{E}[Y \mid X] = \tau\,\mathbb{E}[T \mid X] + g(X)$. We therefore get the equation

$$Y - \mathbb{E}[Y \mid X] = \tau(T - \mathbb{E}[T \mid X]) + U\,.$$

Denoting $\tilde{Y} = Y - \mathbb{E}[Y \mid X]$ and $\tilde{T} = T - \mathbb{E}[T \mid X]$ we can see that the causal effect τ is the solution to the regression problem $\tilde{Y} = \tau\tilde{T} + U$.

The idea now is to solve *two* regression problems to find good function approximations of the conditional expectations $\mathbb{E}[Y \mid X]$ and $\mathbb{E}[T \mid X]$, respectively. We can do this using data drawn from the joint distribution of (X, T, Y) by solving two subsequent model fitting problems, hence the name double machine learning.

Suppose then that we find two function approximations $q(X, Y) \approx \mathbb{E}[Y \mid X]$ and $r(X, T) \approx \mathbb{E}[T \mid X]$. We can define the random variables $\widehat{Y} = Y - q(X, Y)$ and $\widehat{T} = T - r(X, T)$. The final step is to solve the regression problem $\widehat{Y} = \widehat{\tau}\widehat{T} + U$ for the parameter $\widehat{\tau}$.

Compared with inverse propensity score weighting, we can see that finite sample errors in estimating the conditional expectations have a more benign effect on the causal effect estimate $\widehat{\tau}$. In particular, unbiased estimates of the conditional expectations lead to an unbiased estimate of the causal effect.

Heterogeneous treatment effects

In many applications, treatment effects can vary by subpopulation. In such cases we may be interested in the *conditional average treatment effect* (CATE) in the subpopulation defined by $X = x$:

$$\tau(x) = \mathbb{E}[Y \mid \mathrm{do}(T := 1), X = x] - \mathbb{E}[Y \mid \mathrm{do}(T := 0), X = x]\,.$$

We're in luck, because the same proof we saw earlier shows that we can estimate these so-called heterogeneous treatment effects with the propensity score formula:

$$\tau(x) = \mathbb{E}\left[Y\left(\frac{T}{e(X)} - \frac{1-T}{1-e(X)}\right) \mid X = x\right]$$

We can also extend double machine learning easily to the heterogeneous case by replacing the coefficient τ in the first structural equation with a function $\tau(X)$ that depends on X. The argument remains the same except that in the end we need to solve the problem $\widehat{Y} = \widehat{\tau}(X)\widehat{T} + Y$, which amounts to optimizing over a function $\widehat{\tau}$ in some model family rather than a constant $\widehat{\tau}$.

Both inverse propensity score weighting and the double machine learning can, in principle, estimate heterogeneous treatment effects. These aren't the only reductions to model fitting, however. Another popular method, called *causal forests*, constructs decision trees whose leaves correspond covariate settings that deconfound treatment and outcome.[2]

[2] Wager and Athey, "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association* 113, no. 523 (2018): 1228–42.

Quasi-experiments

The idea behind quasi-experimental designs is that sometimes processes in nature or society are structured in a way that enables causal inference. The three most widely used quasi-experimental designs are *regression discontinuities*, *instrumental variables*, and *differences in differences*. We will review the first two briefly to see where machine learning comes in.

Regression discontinuity

Many consequential interventions in society trigger when a certain score R exceeds a threshold value t. The idea behind a regression discontinuity design is that units that fall just below the threshold are indistinguishable from units just above threshold. In other words, whether or not a unit is just above or just below the threshold is a matter of pure chance. We can then hope to identify a causal effect of an intervention by comparing units just below and just above the threshold.

To illustrate the idea, consider an intervention in a hospital setting that is assigned to newborn children just below a birth weight of 1500g. We can ask if the intervention has a causal effect on wellbeing of the child at a later age as reflected in an outcome variable, such as, mortality or cumulative hospital cost in their first year. We expect various factors to influence both birth weight and outcome variable. But

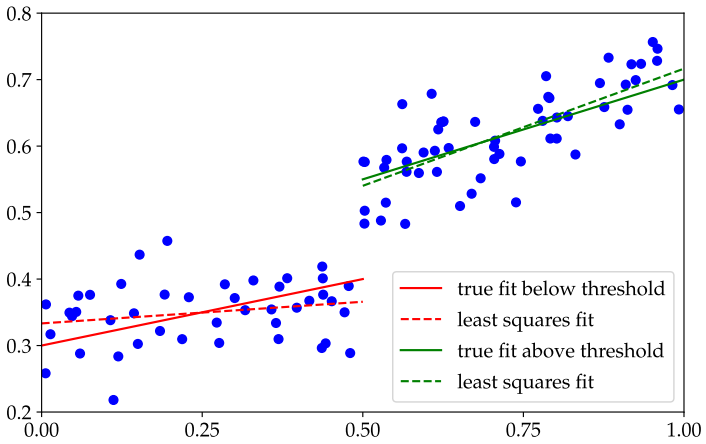


Figure 1: Illustration of an idealized regression discontinuity. Real examples are rarely this clear cut.

we hope that these confounding factors are essentially held constant right around the threshold weight of 1500g. Regression discontinuity designs have indeed been used to answer such questions for a number of different outcome variables.[3]

Once we have identified the setup for a regression discontinuity, the idea is to perform two regressions. One fits a model to the data below the threshold. The other fits the model to data above the threshold. We then take the difference of the values that the two models predict at the threshold as our estimate of the causal effect. As usual, the idea works out nicely in an idealized linear setting and can be generalized in various ways.

There are numerous subtle and not so subtle ways a regression discontinuity design can fail. One subtle failure mode is when intervention incentivizes people to strategically make efforts to fall just below or above the threshold. Manipulation or *gaming* of the running variable is a well-known issue for instance when it comes to social program eligibility.[4] But there are other less obvious cases. For example, school class sizes in data from Chile exhibit irregularities that void regression discontinuity designs.[5] In turn, researchers have come up with tests designed to catch such problems.

[3] Almond et al., "Estimating Marginal Returns to Medical Care: Evidence from at-Risk Newborns," *The Quarterly Journal of Economics* 125, no. 2 (2010): 591–634; Bharadwaj, Løken, and Neilson, "Early Life Health Interventions and Academic Achievement," *American Economic Review* 103, no. 5 (2013): 1862–91.

[4] Camacho and Conover, "Manipulation of Social Program Eligibility," *American Economic Journal: Economic Policy* 3, no. 2 (2011): 41–65.

[5] Urquiola and Verhoogen, "Class-Size Caps, Sorting, and the Regression-Discontinuity Design," *American Economic Review* 99, no. 1 (2009): 179–215.

Instrumental variables

Instrumental variables are a popular quasi-experimental method for causal inference. The starting point is confounding between a treatment T and our outcome of interest Y. We are in a situation where we're unable to resolve confounding via the adjustment formula.

However, what we have is the existence of a special variable Z called an *instrument* that will help us estimate the treatment effect.

What makes Z a valid instrument is nicely illustrated with the following causal graph.

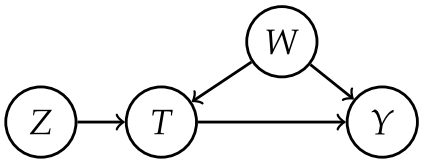


Figure 2: Typical graphical model for an instrumental variable setup

The graph structure encodes two key assumptions:

1. The instrument Z and the outcome Y are unconfounded.
2. The instrument Z has no direct effect on the outcome Y.

Let's walk through how this works out in the one-dimensional linear structural equation for the outcome:

$$Y = \alpha + \beta T + \gamma W + N$$

Here, N is an independent noise term. For convenience, we denote the *error term* $U = \gamma W + N$. What we're interested in is the coefficient β since we can easily verify that it corresponds to the average treatment effect:

$$\beta = \mathbb{E}[Y \mid \mathrm{do}(T := 1)] - \mathbb{E}[Y \mid \mathrm{do}(T := 0)]$$

To find the coefficient β, we cannot directly solve the regression problem $Y = \alpha + \beta T + U$, because the error term U is not independent of T due to the confounding influence of W.

However, there's a way forward after we make a few additional assumptions:

1. The error term is zero mean: $\mathbb{E}[U] = 0$
2. The instrument is uncorrelated with the error term: $\mathrm{Cov}(Z, U) = 0$
3. Instrument and treatment have nonzero correlation: $\mathrm{Cov}(Z, T) \neq 0$

The first two assumptions directly imply

$$\mathbb{E}[Y - \alpha - \beta T] = 0$$
$$\mathbb{E}[Z(Y - \alpha - \beta T)] = 0$$

This leaves us with two linear equations in α and β so that we can solve for both parameters. Indeed, $\alpha = \mathbb{E}[Y] - \beta\,\mathbb{E}[T]$. Plugging this into the second equation, we have

$$\mathbb{E}[Z((Y - \mathbb{E}[Y]) - \beta(T - \mathbb{E}[T]))] = 0,$$

which implies, via our third assumption $\mathrm{Cov}(T, Z) \neq 0$,

$$\beta = \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(T, Z)}\,.$$

There's a different intuitive way to derive this solution by solving a *two step least squares* procedure:

1. Predict the treatment from the instrument via least squares regression, resulting in the predictor $\widehat{T} = cZ$.
2. Predict the outcome from the predicted treatment using least squares regression, resulting in the predictor $\widehat{Y} = \beta'\widehat{T}$.

A calculation reveals that indeed $\beta' = \beta$, the desired treatment effect. To see this note that

$$c = \frac{\mathrm{Cov}(Z, T)}{\mathrm{Var}(Z)}$$

and hence

$$\beta' = \frac{\mathrm{Cov}(Y, \widehat{T})}{\mathrm{Var}(\widehat{T})} = \frac{\mathrm{Cov}(Y, Z)}{c\,\mathrm{Var}(Z)} = \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(T, Z)} = \beta\,.$$

This solution directly generalizes to the multi-dimensional linear case. The two stage regression approach is in fact the way instrumental variables is often introduced operationally. We see that again instrumental variables is a clever way of reducing causal inference to prediction.

One impediment to instrumental variables is a poor correlation between the instrument and the treatment. Such instruments are called *weak instruments*. In this case, the denominator $\mathrm{Cov}(T, Z)$ in our expression for β is small and the estimation problem is ill-conditioned. The other impediment is that the causal graph corresponding to instrumental variables is not necessarily easy to come by in applications. What's delicate about the graph is that we want the instrument to have a significant causal effect on the treatment, but at the same time have no other causal powers that might influence the outcome in a way that's not mediated by the treatment.

Nonetheless, researchers have found several intriguing applications of instrumental variables. One famous example goes by the name *judge instruments*. The idea is that within the United States, at least in certain jurisdictions and courts, defendants may be assigned randomly to judges. Different judges then assign different sentences, some perhaps more lenient, others harsher. The treatment here could be the sentence length and the outcome may indicate whether or not the defendant went on to commit another crime upon serving the prison sentence. A perfectly random assignment of judges implies

that the judge assignment and the outcome are unconfounded. Moreover, the assignment of a judge has a causal effect on the treatment, but plausibly no direct causal effect on the outcome. The assignment of judges then serves as an instrumental variable. The observation that judge assignments may be random has been the basis of much causal inference about the criminal justice system. However, the assumption of randomness in judge assignments has also been challenged.[6]

[6] Chilton and Levy, "Challenging the Randomness of Panel Assignment in the Federal Courts of Appeals," *Cornell L. Rev.* 101 (2015): 1.

Limitations of causal inference in practice

It's worth making a distinction between causal modeling broadly speaking and the practice of causal inference today. The previous chapter covered the concepts of causal modeling. Structural causal models make it painfully clear that the model necessarily specifies strong assumptions about the data generating process. In contrast, the practice of causal inference we covered in this chapter seems almost *model-free* in how it reduces to pattern classification via technical assumptions. This appears to free the practitioner from difficult modeling choices.

The assumptions that make this all work, however, are not verifiable from data. Some papers seek assurance in statistical robustness checks, but these too are sample-based estimates. Traditional robustness checks, such as resampling methods or leave-one-out estimates, may get at issues of generalization, but cannot speak to the validity of causal assumptions.

As a result, a certain pragmatic attitude has taken hold. If we cannot verify the assumption from data anyway, we might as well make it in order to move forward. But this is a problematic position. Qualitative and theoretical ways of establishing substantive knowledge remain relevant where the limitations of data set in. The validity of a causal claim cannot be established solely based on a sample. Other sources of substantive knowledge are required.

Validity of observational methods

The empirical evidence regarding the validity of observational causal inference studies is mixed and depends on the domain of application.

A well known article compared observational studies in the medical domain between 1985 and 1998 to the results of randomized controlled trials.[7] The conclusion was good news for observational methods:

> We found little evidence that estimates of treatment effects in observational studies reported after 1984 are either consistently larger than or

[7] Benson and Hartz, "A Comparison of Observational Studies and Randomized, Controlled Trials," *New England Journal of Medicine* 342, no. 25 (2000): 1878–86.

> qualitatively different from those obtained in randomized, controlled trials.

Another study around the same time came to a similar conclusion:

> The results of well-designed observational studies (with either a cohort or a case–control design) do not systematically overestimate the magnitude of the effects of treatment as compared with those in randomized, controlled trials on the same topic.[8]

[8] Concato, Shah, and Horwitz, "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs," *New England Journal of Medicine* 342, no. 25 (2000): 1887–92.

One explanation, however, is that medical researchers may create observational designs with great care on the basis of extensive domain knowledge and prior investigation.

Freedman's paper *Statistical Models and Shoe Leather* illustrates this point through the famous example of Jon Snow's discovery from the 1850s that cholera is a waterborne disease.[9] Many associate Snow with an early use of quantitative methods. But the application of those followed years of substantive investigation and theoretical considerations that formed the basis of the quantitative analysis.

[9] Freedman, "Statistical Models and Shoe Leather," *Sociological Methodology*, 1991, 291–313.

In other domains, observational methods have been much less successful. Online advertising, for example, generates hundreds of billions of dollars in yearly global revenue, but the causal effects of targeted advertising remain a subject of debate.[10] Randomized controlled trials in this domain are rare for technical and cultural reasons. Advertising platforms are highly optimized toward a particular way of serving ads that can make true randomization difficult to implement. As a result, practitioners rely on a range of observational methods to determine the causal effect of showing an ad. However, these methods tend to perform poorly as a recent large-scale study reveals:

[10] Hwang, *Subprime Attention Crisis* (Farrar, Strauss; Giroux, 2020).

> The observational methods often fail to produce the same effects as the randomized experiments, even after conditioning on extensive demographic and behavioral variables. We also characterize the incremental explanatory power our data would require to enable observational methods to successfully measure advertising effects. Our findings suggest that commonly used observational approaches based on the data usually available in the industry often fail to accurately measure the true effect of advertising.[11]

[11] Gordon et al., "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook," *Marketing Science* 38, no. 2 (2019): 193–225.

Interference, interaction, and spillovers

Confounding is not the only threat to the validity of causal studies. In a medical setting, it's often relatively easy to ensure that treatment of one subject does not influence the treatment assignment or outcome of any other unit. We called this the Stable Unit Treatment Value Assumption (SUTVA) in the previous chapter and noted that it holds by default for the units in a structural causal models. Failures

of SUTVA, however, are common and go by many names, such as, interference, interaction, and spill-over effects.

Take the example of an online social network. Interaction between units is the default in all online platforms, whose entire purpose is that people interact. Administering treatment to a subset of the platform's users typically has some influence on the control group. For example, if our treatment exposes a group of users to more content of a certain kind, those users might share the content with others outside the treatment group. In other words, treatment *spills over* to the control group. In certain cases, this problem can be mitigated by assigning treatment to a cluster in the social network that has a boundary with few outgoing edges thus limiting bias from interaction.[12]

Interference is also common in the economic development context. To borrow an example from economist John Roemer,[13] suppose we want to know if better fishing nets would improve the yield of fishermen in a town. We design a field experiment in which we give better fishing nets to a random sample of fishermen. The results show a significantly improved yield for the treated fishermen. However, if we scale the intervention to the entire population of fishermen, we might cause overfishing and hence reduced yield for everyone.

[12] Eckles, Karrer, and Ugander, "Design and Analysis of Experiments in Networks: Reducing Bias from Interference," *Journal of Causal Inference* 5, no. 1 (2016).

[13] Roemer, *How We Cooperate: A Theory of Kantian Optimization* (Yale University Press, 2019).

Chapter notes

Aside from the introductory texts from the previous chapter, there are a few more particularly relevant in the context of this chapter.

The textbook by Angrist and Pischke[14] covers causal inference with an emphasis on regression analysis an applications in econometrics. See Athey and Imbens[15] for a more recent survey of the state of causal inference in econometrics.

Marinescu et al.[16] give a short introduction to quasi-experiments and their applications to neuroscience with a focus on regression discontinuity design, instrumental variables, and differences in differences.

[14] Angrist and Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton university press, 2008).

[15] Athey and Imbens, "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives* 31, no. 2 (2017): 3–32.

[16] Marinescu, Lawlor, and Kording, "Quasi-Experimental Causality in Neuroscience and Behavioural Research," *Nature Human Behaviour* 2, no. 12 (2018): 891–98.

References

Almond, Douglas, Joseph J Doyle Jr, Amanda E Kowalski, and Heidi Williams. "Estimating Marginal Returns to Medical Care: Evidence from at-Risk Newborns." *The Quarterly Journal of Economics* 125, no. 2 (2010): 591–634.

Angrist, Joshua D., and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton university press, 2008.

Athey, Susan, and Guido W Imbens. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31, no. 2 (2017): 3–32.

Benson, Kjell, and Arthur J Hartz. "A Comparison of Observational Studies and Randomized, Controlled Trials." *New England Journal of Medicine* 342, no. 25 (2000): 1878–86.

Bharadwaj, Prashant, Katrine Vellesen Løken, and Christopher Neilson. "Early Life Health Interventions and Academic Achievement." *American Economic Review* 103, no. 5 (2013): 1862–91.

Camacho, Adriana, and Emily Conover. "Manipulation of Social Program Eligibility." *American Economic Journal: Economic Policy* 3, no. 2 (2011): 41–65.

Chilton, Adam S, and Marin K Levy. "Challenging the Randomness of Panel Assignment in the Federal Courts of Appeals." *Cornell L. Rev.* 101 (2015): 1.

Concato, John, Nirav Shah, and Ralph I Horwitz. "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs." *New England Journal of Medicine* 342, no. 25 (2000): 1887–92.

Deaton, Angus, and Nancy Cartwright. "Understanding and Misunderstanding Randomized Controlled Trials." *Social Science & Medicine* 210 (2018): 2–21.

Eckles, Dean, Brian Karrer, and Johan Ugander. "Design and Analysis of Experiments in Networks: Reducing Bias from Interference." *Journal of Causal Inference* 5, no. 1 (2016).

Freedman, David A. "Statistical Models and Shoe Leather." *Sociological Methodology*, 1991, 291–313.

Gordon, Brett R, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook." *Marketing Science* 38, no. 2 (2019): 193–225.

Hwang, Tim. *Subprime Attention Crisis*. Farrar, Strauss; Giroux, 2020.

Marinescu, Ioana E, Patrick N Lawlor, and Konrad P Kording. "Quasi-Experimental Causality in Neuroscience and Behavioural Research." *Nature Human Behaviour* 2, no. 12 (2018): 891–98.

Roemer, John E. *How We Cooperate: A Theory of Kantian Optimization*. Yale University Press, 2019.

Urquiola, Miguel, and Eric Verhoogen. "Class-Size Caps, Sorting, and the Regression-Discontinuity Design." *American Economic Review* 99, no. 1 (2009): 179–215.

Wager, Stefan, and Susan Athey. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." *Journal of the American Statistical Association* 113, no. 523 (2018): 1228–42.