ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Causal Inference in the Social Sciences

Guido W. Imbens

Department of Economics and Graduate School of Business, Stanford University, Stanford, California, USA; email: imbens@stanford.edu

Annu. Rev. Stat. Appl. 2024. 11:123–52

First published as a Review in Advance on November 17, 2023

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-033121-114601

Keywords

causal inference, experiments, observational studies, unconfoundedness, instrumental variables, synthetic controls, difference-in-differences, regression discontinuity, double robustness

Abstract

Knowledge of causal effects is of great importance to decision makers in a wide variety of settings. In many cases, however, these causal effects are not known to the decision makers and need to be estimated from data. This fundamental problem has been known and studied for many years in many disciplines. In the past thirty years, however, the amount of empirical as well as methodological research in this area has increased dramatically, and so has its scope. It has become more interdisciplinary, and the focus has been more specifically on methods for credibly estimating causal effects in a wide range of both experimental and observational settings. This work has greatly impacted empirical work in the social and biomedical sciences. In this article, I review some of this work and discuss open questions.

1. INTRODUCTION

Knowledge of causal effects is of great importance to decision makers in a wide variety of settings, including policy makers in government and nongovernment organizations and decision makers in the private sector. In many cases, these causal effects are not known to the decision makers and need to be estimated from data. This fundamental problem has been known and studied for many years in many disciplines. In the past thirty years, however, the amount of methodological and empirical research has increased substantially. Its scope has also changed dramatically. Just to illustrate how much the area has grown in the past thirty years, consider **Figure 1**, similar to figures for different methodological terms in the work of Currie et al. (2020). This figure shows the fraction of working papers in empirical economics published by the National Bureau of Economic Research (a widely used working paper series in economics) as well as the fraction of published papers in leading economics journals that use the term "causality" or related terms such as "causal." Whereas before 1990, the percentage of papers using the term "causality" in empirical papers in economics was modest—between 10% and 15%—and relatively constant, starting around 1990, the percentage began to increase rapidly, so that by 2015, a full 50% of empirical papers in economics used the term "causality."

During these past thirty years, the study of statistical problems related to the estimation of causal effects has become more interdisciplinary, with methodological contributions and insights from statistics, econometrics, political science, computer science, epidemiology, biomedical science, and others. The focus in this literature has been specifically on methods for credibly estimating causal effects in both experimental and observational settings. This work has greatly impacted empirical work in a variety of disciplines, including social and biomedical sciences. In this article, I review some of this work and discuss open questions.

This review focuses on four areas. The first is the work on the analysis and design of randomized controlled trials (RCTs) (Section 3). This is the traditional setting where researchers in statistics have studied the estimation of causal effects since the seminal contributions by Fisher and Neyman in the 1920s and 1930s, often in biomedical and agricultural settings. More

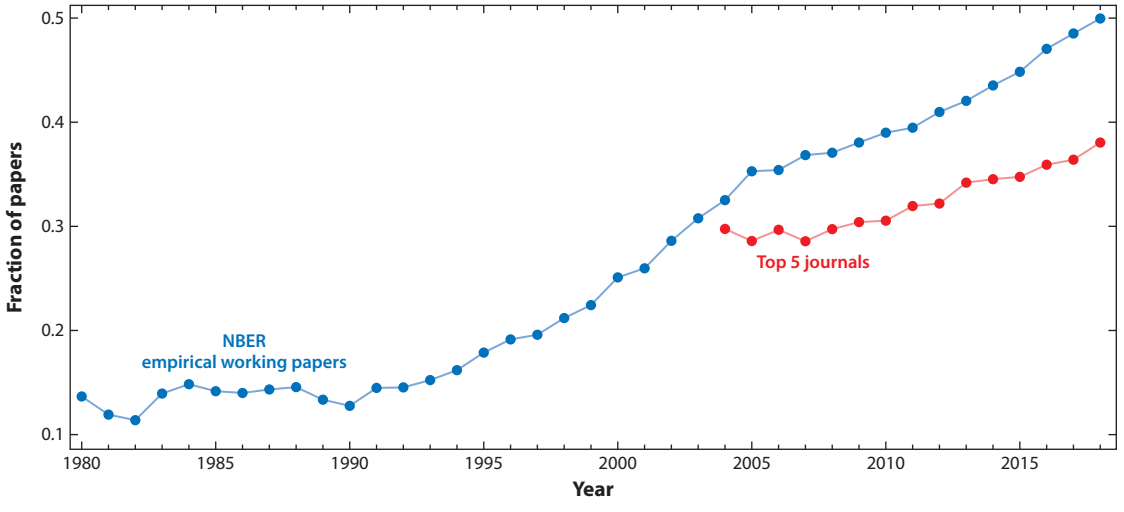

Figure 1

Fraction of papers using the term "causal" or "causality," motivated by similar figures in Currie et al. (2020). Abbreviation: NBER, National Bureau of Economic Research.

recently, researchers in statistics and social and computer sciences have developed innovative new experimental designs, partly motivated by the dramatic increase in online experiments by tech companies that now run hundreds of thousands of experiments annually (Gupta et al. 2019). In fact, there are now multiple companies dedicated to running online randomized experiments. These new experimental designs include sophisticated adaptive designs, as well as designs taking into account complex interactions between units. The second area discussed is the analysis of observational studies under unconfoundedness (Section 4). This is the most common setting for observational studies. Following Rosenbaum & Rubin (1983b), researchers often make assumptions that justify adjusting for observed confounders through regression methods, matching, inverse-propensity-score weighting, and variations thereon. The recent literature in this area has focused on analyzing heterogeneous effects in this setting, as well as allowing for the presence of high-dimensional confounders. The third area I discuss is methods for analyzing observational studies in settings where unconfoundedness is not a plausible assumption, and in fact not even a reasonable starting point (Section 5). This is an area that has been studied in econometrics since the 1930s (Tinbergen 1930, Haavelmo 1943). In this literature a number of methods have been developed that allow for credible estimation of causal effects in specific settings without unconfoundedness assumptions. The most popular of these methods include instrumental variables, difference-in-differences (DID) methods, synthetic control (SC), and regression discontinuity designs. Finally, I discuss methods for combining observational and experimental data leveraging the strengths of both in order to address the shortcomings in either of them (Section 6).

This review is not comprehensive. One important area that I do not discuss in detail concerns dynamic models, which have received much attention in epidemiology (Robins 1989, 1997; Robins et al. 2000) and in the older econometric literature on panel data (Chamberlain 1984), but less in the recent econometric literature, with an exception in Han (2021).

There have been a number of general books on causality and causal inference in statistics and social sciences in the past two decades, including those of Rubin (2006), Imbens & Rubin (2015), Cunningham (2018), Pearl (2000), Rosenbaum (2002, 2010), Morgan & Winship (2015), and Huntington-Klein (2021), but the pace of the research means it is difficult for these to be up-to-date. There are also a number of reviews in journals. Surveys with a focus on social sciences include those by Imbens & Wooldridge (2009), Abadie & Cattaneo (2018), and Keele (2015).

2. BASIC SETUP AND QUESTIONS

Following much of the empirical literature, in this review, I use the potential outcome framework, initially used in experimental settings by Splawa-Neyman [1990 (1923)] and proposed as a general framework for causal inference in observational studies by Rubin (1977). We start with a population of N units, indexed by $i = 1, \ldots, N$. In the case with binary treatment, we postulate the existence for each unit of two potential outcomes, $Y_i(C)$ for the outcome for unit i if the unit is exposed to the control treatment and $Y_i(T)$ if the unit is exposed to the active or new treatment. The notation can be extended to allow for multi-valued treatments (e.g., Imbens 2000). This notation already subsumes the stable unit treatment value assumption (SUTVA; Rubin 1978) so that the outcome for one unit is not affected by the treatment exposure for a different unit. We discuss the implications arising from spillovers where this assumption is violated in Section 3.3. Given the two potential outcomes, the causal effect is some comparison of $Y_i(C)$ and $Y_i(T)$, often the difference $Y_i(T) - Y_i(C)$, or some average thereof. The challenge in what Holland (1986) famously called "the fundamental problem of causal inference" is that we do not directly observe any causal effects: We can only see $Y_i(C)$ if unit i is exposed to the control treatment (which we denoted by $W_i = C$), and we can only observe $Y_i(T)$ if unit i is exposed to the new treatment (which

we denote by $W_i = T$). If Y_i is the realized and observed outcome, we have

$$Y_i = \begin{cases} Y_i(C) & \text{if } W_i = C, \\ Y_i(T) & \text{if } W_i = T. \end{cases}$$

In addition to the treatment W_i and the outcome Y_i, we may observe additional variables for each unit. Some of these include pretreatment variables, known to the researcher not to be affected by the treatment. We denote such variables by X_i for unit i.

Often the interest is in some average effect of the treatment, such as the sample average treatment effect,

$$\tau^{\text{sample}} \equiv \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i(T) - Y_i(C) \Big).$$

Alternatively we may be interested in the average effect in the population, $\tau^{\text{pop}} \equiv \mathrm{E}[Y_i(T) - Y_i(C)]$, if the sample can be viewed as a random sample from some population of interest, or the average effect in some subpopulation, for example, the average effect for the treated, $\tau^{\text{treated}} = \mathrm{E}[Y_i(T) - Y_i(C) | W_i = T]$. The difference between the sample average treatment effect and the population average treatment effect is subtle. Typically there are no implications for estimation: The best estimator for τ^{sample} is typically also the best estimator for τ^{pop} if the sole information is in the form of a random sample from the population. But there are implications for inference: We cannot estimate τ^{pop} as precisely as τ^{sample} because there is an additional layer of uncertainty. These issues have received some attention in the recent literature in the discussions of design-based versus model- or sampling-based uncertainty. For more information, readers are directed to Imbens (2004), Rosenbaum (2010), Imbens & Rubin (2015), and Abadie et al. (2020).

In the recent literature, there is additional emphasis on heterogeneity in treatment effects by characteristics of the population. Although there was always interest in differences in treatment effects by prespecified groups, the use of large data sets, both from online experiments and from observational studies based on administrative data, in combination with modern machine learning methods, has led researchers to develop effective methods for estimating either the conditional average treatment effect (CATE),

$$\tau(x) = \mathrm{E}\left[Y_i(T) - Y_i(C) \middle| X_i = x \right],$$

or summary statistics thereof (Chernozhukov et al. 2018, Wager & Athey 2018). Beyond estimating the CATE, researchers have also focused on estimating policy functions that capture the optimal assignment as a function of pretreatment variables (e.g., Athey & Wager 2021).

3. RANDOMIZED CONTROLLED TRIALS

Since the seminal work in the 1920s by Neyman and Fisher [Splawa-Neyman 1990 (1923), Fisher 1937], the use of RCTs has become the most trusted method for estimating causal effects. The insistence of regulatory agencies on experimental evidence in the drug approval process since the 1960s has led many researchers to refer to this as the gold standard for causal inference.

Experimental evaluations have also become prominent in social sciences in the past twenty years, as recognized in the Nobel Prize in Economic Sciences awarded to Abhijit Banerjee, Esther Duflo, and Michael Kremer in 2019 (see Banerjee 2020, Duflo 2020, Kremer 2020). Whereas previously, experiments were only occasionally conducted, notably the large scale negative income tax experiments in the 1970s (see, for example, Robins 1985) and various labor market experiments in the 1980s (see, for example, LaLonde 1986, Hotz et al. 2006), nowadays political scientists,

economists, and other social scientists routinely do targeted experiments (Harrison & List 2004, Kalla & Broockman 2018).

Although the original experimental designs going back to Fisher and Neyman continue to be widely used, interesting new designs have been developed in recent years. This is partly the result of the recent interest in conducting randomized trials in academic settings, and partly because of the interest from private sector organizations in experimental evaluations, often in online settings.

3.1. Simple Randomized Controlled Trials

The canonical RCT, going back to Neyman and Fisher [Splawa-Neyman 1990 (1923), Fisher 1937], focuses on the case with a population of N units, each characterized by a pair of potential outcomes $(Y_i(C), Y_i(T))$, as mentioned previously. This notation rules out the presence of spillover effects, where exposing one unit to the treatment affects outcomes for other units. The absence of such spillover effects is plausible in many of the traditional biomedical settings for experimentation, such as drug trials or agricultural experiments. However, even in biomedical settings, there are exceptions, such as experiments involving infectious diseases. Concern about the presence of spillovers in modern social science settings is widespread and has motivated new experimental designs and analysis methods that we discuss in Section 3.3. The interest in traditional RCTs is in the causal effects $\tau_i = Y_i(T) - Y_i(C)$, with the focus often on the average effect for the N units in the sample or study population:

$$\tau^{\text{sample}} = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(T) - Y_i(C)\Big).$$

Out of this population with N units, N_T units are drawn at random and assigned to the treatment group, and the remaining $N_C = N - N_T$ units are assigned to the control group, with $W_i \in \{C, T\}$ denoting the treatment.

In this setting, Fisher (1937) focused on testing sharp null hypotheses regarding the causal effects. The most common hypothesis is that the treatment had no effect on the outcomes whatsoever:

$$H_0: \;\; Y_i(C) = Y_i(T)\;\forall i, \qquad \text{against the alternative, } H_a: \exists\, i \text{ s.t. } Y_i(C) \neq Y_i(T).$$

Although approximations to such exact testing procedures are still widely used, in social sciences, it is rarely the case that tests of null hypotheses of no effects, either on average or for all units, are of primary interest. While such questions may be of substantial interest in the development of new drugs, in many settings decision makers are most interested in the magnitudes of effects, and whether these effects are substantially meaningful. In the experimental setting, that makes the results of Splawa-Neyman [1990 (1923)] more relevant. Neyman focused on estimating the overall average effect using the difference in means:

$$\hat{\tau} = \overline{Y}_T - \overline{Y}_C, \qquad \text{where } \overline{Y}_w = \frac{1}{N_w}\sum_{i:W_i=w} Y_i,\; w \in \{C, T\}.$$

He showed that this estimator is unbiased for τ^{sample}, the average effect in the sample. Neyman also derived the exact variance under randomization,

$$\mathbb{V} = \frac{1}{(N-1)N_C}\sum_{i=1}^{N}\Big(Y_i(C) - \overline{Y}(C)\Big)^2 + \frac{1}{(N-1)N_T}\sum_{i=1}^{N}\Big(Y_i(T) - \overline{Y}(T)\Big)^2$$

$$-\frac{1}{(N-1)N}\sum_{i=1}^{N}\Big(Y_i(T) - Y_i(C) - (\overline{Y}(T) - \overline{Y}(C))\Big)^2,$$

for $w \in \{C, T\}$. He proposed the following conservative estimator for this variance:

$$\hat{\mathbb{V}} = \frac{1}{N_C(N_C-1)} \sum_{i:W_i=C} \left(Y_i(C) - \overline{Y}_C\right)^2 + \frac{1}{N_T(N_T-1)} \sum_{i:W_i=T} \left(Y_i(T) - \overline{Y}_T\right)^2.$$

These basic results continue to be the basis of recent experimentation in biomedical, social science, and industry settings. Common modifications include the exploitation of unit-level covariates or pretreatment variables to increase the precision of the estimators. The presence of the covariates can be used to improve the design of the experiments through stratification or, in the limit, pairing of similar units (Athey & Imbens 2017). Although their incorporation in the design stage of an experiment is to be preferred (Rubin 2008), their presence can also be used in the analysis stage of the experiment through ex post adjustment, through regression or other methods. For most estimators (e.g., regression estimators), the accuracy or validity of the model does not affect the bias of the estimator (although this result is asymptotic and does not necessarily hold in finite samples) but can lead to substantial improvements of the asymptotic precision of the estimators (Lin 2013).

3.2. Adaptive Experiments

The standard experimental designs are powerful in the settings they were designed for. However, in many settings where researchers are currently interested in conducting randomized evaluations, they are not the most effective designs. One such setting that is common in online experimentation has four distinguishing features (see Gupta et al. 2019). First, units arrive sequentially into the study; second, outcomes are measured fast; third, there may be multiple, possibly many, treatment arms that the researcher may be interested in; and fourth, the researcher is interested in finding a good treatment and is not interested in finding precise estimates of the efficacy of inferior arms. These four features combined imply that the researcher can effectively adapt the experimental design during the experiment to substantially improve its information content for decision making for a given number of units.

Let us consider a specific example. Suppose an online marketer wants to choose one from ten different advertisements to use in an online advertising campaign, or a political campaign wants to choose one out of ten potential solicitations for campaign contributions. To make this choice, the decision maker wants to conduct a randomized experiment. A traditional, static experiment would involve assignment of equal amounts of incoming traffic to each of the potential treatments. To achieve reasonable precision for all treatment arms, such an experiment would require a large number of experimental units. An alternative is to use an adaptive experiment. Initially, units are assigned to each of the treatment arms with equal probability, and outcomes are measured. After some outcomes are observed, the assignment probabilities are updated, taking into account the information about the efficacy of the various treatment arms based on the observed outcomes for the initial units. If the goal is to find the best treatment, it is obvious that once one has seen some initial results, some treatment arms can likely be essentially dismissed as candidates for the best treatment. Sending a substantial amount of additional incoming traffic to those treatments would be a waste of experimental units. Multi-armed bandit algorithms adapt the experimental design in order to more efficiently explore which treatment arms are candidates for being the best, and at the same time exploit the information already acquired to assign new observations to better-performing arms. The balancing of exploration and exploitation is a key feature of these algorithms.

There are two general approaches to this type of adaptive experimentation (Lattimore & Szepesvári 2020). One is Thompson sampling (Thompson 1933, Scott 2010, Russo et al. 2018), where the probability of the next unit being assigned to any particular treatment is proportional

to the posterior probability that that treatment arm is the best one. Suppose the outcome is binary, and the joint prior distribution for the K success probabilities p_k is flat, that is, the product of independent Beta distributions with parameters $\alpha_k = \beta_k = 1$. Then, if, after N_k units are assigned to treatment k initially, we see M_k successes and $N_k - M_k$ failures, the posterior distribution for the success probability p_k is a Beta distribution with parameters $\alpha_k = M_k + 1$ and $\beta_k = N_k - M_k + 1$:

$$p_k|\text{data} \sim \mathcal{B}(M_k + 1, N_k - M_k + 1),$$

independent across treatment arms. Given the joint posterior distribution for the success probabilities, we can infer the posterior probability that treatment arm k is the best one, $\text{pr}(p_k = \max_{m=1}^{K} p_m)$, and we assign the next unit to treatment arm k with that probability. As a result, we assign few units to treatment arms that initially perform poorly and that are therefore judged unlikely to be the optimal arm. At the end of the experiment, we therefore may not be able to infer the precise success probabilities for the inferior arms, but that is not the goal here: We want to learn which arm is optimal, and our losses are related to differences in efficacy between the chosen arm and the optimal arm.

The second approach to updating the assignment probabilities is the upper confidence bounds (UCB) approach (Lai & Robbins 1985, Lattimore & Szepesvári 2020). Here we calculate, after some initial assignments for each treatment arm, a confidence interval for each of the success probabilities, with confidence level α—say, the empirical success rate plus and minus 1.96 times the standard error for a 95% confidence interval. The next unit is assigned to the treatment arm that has the highest value for the UCB, the highest value for the empirical success rate plus 1.96 times the standard error. With each treatment assignment, we slowly increase the level of the confidence intervals toward one so that every treatment arm still receives some traffic.

In both the Thompson sampling and UCB approaches, we increasingly de-emphasize treatment arms once we are confident that they are not the best in the set. The simple bandit algorithms lead to substantial improvements over standard experiments, but there are many subtle issues regarding their use as well as modifications geared toward more complex settings that are important in practice. The first issue concerns inference. Simply using the average outcomes as an estimator for the expected outcome and its standard deviation scaled by the square root of the number of units as the standard error introduces biases. This is easy to see in a simple example: Suppose there are two stages, where in the first stage, 100 units are assigned to one of two treatment arms, and in the second stage, the next 100 units are all assigned to the treatment arm with the highest empirical success rate. If the true expected outcomes are equal, the empirical success rate for the arm with the lowest initial success rate is biased downward. Second, interesting complications arise in settings with covariates, where the assignment probabilities depend both on earlier outcomes and on the characteristics of the incoming observations, in what are known as contextual bandits (Dimakopoulou et al. 2018). Third, concerns arise in settings where the expected outcomes may change over time, so-called nonstationary bandits. Changes can be in the form of stochastic trends or seasonal (day of the week or month of the year) effects. In that case, one needs to slow down the exploitation part of the algorithm in order to ensure that there is a sufficient number of units for each of the treatment arms so that we do not erroneously discard the good arms (Besbes et al. 2014, Liu et al. 2023).

3.3. Experiments in the Presence of Spillovers

One key assumption underlying standard RCTs is the no-interference assumption, or SUTVA, requiring that treatment assignments for one unit do not affect the outcome for any other unit. This is plausible in many biomedical drug trials, where one individual taking a drug typically does not affect any other individual. Even in such cases, it is not always plausible: Vaccinating

some individuals for infectious diseases affects outcomes for individuals not vaccinated. However, the problem of spillovers or interference between units is pervasive in social sciences, which fundamentally are concerned with settings involving individuals interacting in systematic ways. A recent example is from Crépon et al. (2013), who study the effect of a labor market program for unemployed individuals in multiple labor markets. The randomized assignment to the program was carried out in two stages. First, each labor market was randomly assigned to a treatment probability. Second, each unemployed individual was assigned to the treatment according to the probability assigned to the labor market the individual belonged to. The authors find that the estimated average treatment effect in each market declines with the fraction of treated individuals in that labor market. This is not surprising if part of the effect of the program comes through making the treated unemployed more attractive hires compared with control individuals, in a setting where the number of open positions in each labor market is approximately fixed.

The appropriate experimental design in the presence of spillovers, and the analysis of data from such experiments, depends on the precise nature of the spillovers. This has led to an extensive literature studying cases relevant in particular contexts, in both experimental and observational settings. A common theme is to limit the spillovers through exposure mappings (Aronow & Samii 2017) that measure what components of the full treatment vector matter for a particular unit.

One leading case is the setting where the population is partitioned into subsets, referred to as strata or clusters, such that the spillovers are limited to units within the cluster (Hudgens & Halloran 2008). Examples include the labor market setting in Crépon et al. (2013), but also educational settings where treatments applied to one student may affect all students in the same classroom (but not students in other classrooms), or rideshare companies where treatments applied to one customer affect all customers in the same market at that time (but not customers in other markets).

Another important setting is that of networks where spillovers or interactions arise through network links (Athey et al. 2018a, Basse et al. 2019). Here challenges are more substantial than in the stratified case because, depending on the nature of the spillovers, treating one unit may affect units it is not connected to. Bond et al. (2012) present a well-known example, where treating some individuals in a way that makes them more likely to vote affects the voting behavior of their friends as well as individuals beyond their direct friends.

An alternative setup that allows for general spillovers is based on bipartite graphs (Pouget-Abadie et al. 2019, Zigler & Papadogeorgou 2021). In contrast to much of the experimental design literature, the starting point is a single set of N units, to which the treatment could be applied and for which potential outcomes are defined. There is, in this approach, no longer a simple one-to-one correspondence between the units on which the treatments are defined and the units on which the outcomes are measured. The bipartite graph framework starts with a set of treatment units $\mathcal{T}$, with binary treatment indicators, $\{W_i, i \in \mathcal{T}\}$, and a set of outcome units $\mathcal{Q}$ with observed responses $\{Y_j, j \in \mathcal{Q}\}$, together with a bipartite graph with vertex sets $\mathcal{P}$ and $\mathcal{O}$ that describes which treatments affect which outcomes.

Bajari et al. (2021, 2023) and Johari et al. (2022) consider a setting with two or more populations where the treatments are assigned to, and outcomes are measured on, pairs or tuples of units. This setting is a natural one in marketplaces, for example, rideshare companies such as Uber and Lyft, or rental markets such as Airbnb. Treatments—say, changes in the interaction between drivers and riders such as default tipping policies—are assigned to pairs of drivers and riders, in contrast to traditional experiments where treatments are assigned to all drivers (for all riders) or to riders (for all drivers). By creating variation in the share of treated drivers for each rider, and the other way around, the researcher has the ability to learn about the interaction between the two sides of the market and the spillovers that result from those interactions.

4. OBSERVATIONAL STUDIES WITH UNCONFOUNDEDNESS

Although experiments have become more prominent in social sciences over the past thirty years, observational studies continue to be the mainstay of social sciences. To balance the superior internal validity of experimental studies, there are three aspects of causal studies where observational studies often have the advantage over randomized experiments: (*a*) more detailed information on units, (*b*) larger sample sizes, and (*c*) improved representativeness or external validity.

The most important approach to observational studies is the one in which, within homogenous subpopulations, the treatment assignment is assumed to be as good as random, or unconfounded (Rosenbaum & Rubin 1983b), so that within such subpopulations, we can analyze the data as if they arose from a randomized experiment. Formally, with X_i denoting a vector of pretreatment variables or covariates, the key assumption is

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \,\Big|\, X_i. \qquad 1.$$

In addition, there is typically an overlap assumption that guarantees that the assignment probability is bounded away from zero and one,

$$e(x) \equiv \text{pr}(W_i = T | X_i = x) \in (c, 1-c), \qquad \text{for some } c > 0, \qquad 2.$$

where $e(\cdot)$ is the propensity score that plays a key role in this setting. Combined, these two assumptions are referred to as strong ignorability (Rosenbaum & Rubin 1983b). There is a large theoretical literature developing methods for estimation and inference in this setting (reviews include Rosenbaum 1984, Rubin 2006, Stuart 2010, Imbens 2015, Zubizarreta et al. 2023), as well as a huge empirical literature that relies on some form of this assumption, variously referred to as ignorable treatment assignment (Rosenbaum & Rubin 1983b), exogeneity (Imbens 2004), and unconfoundedness (Rubin 1978).

In the graphical tradition (Pearl 1995, 2000; Peters et al. 2017; Pearl & Mackenzie 2018), the key unconfoundedness assumption can be expressed as in **Figure 2**, with n common ancestors for both the treatment and outcome. In this setting, there is no need to specify the causal links, absent or present, between these ancestors. Most common estimators are not affected by the presence or absence of those links. What is important, though, is that none of these variables are affected by the treatment or outcome—in the directed acyclic graph (DAG) terminology, none of them are descendants of the outcome or the treatment. Typically, the credibility of that assumption is based on the notion that these variables precede the treatment (Rosenbaum 1984, 2010). In practice, using variables causally affected by the treatment or outcome is the most common mistake in

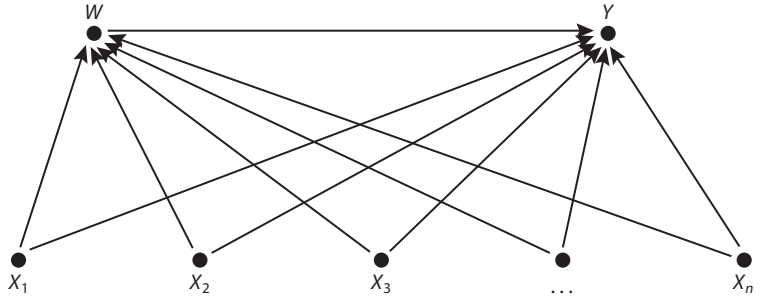

Figure 2

Unconfounded treatment assignment: $X_1, \ldots, X_n$ are exogenous (pretreatment variables). Adjusting for them in an unconfoundedness-based analysis removes all biases in comparisons between treated and control units.

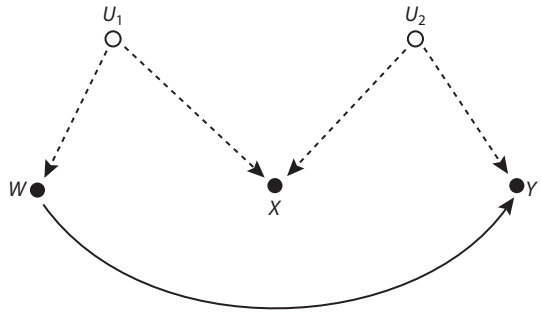

Figure 3

M-bias: Although X may be a pretreatment, using it as a conditioning variable in an unconfoundedness analysis introduces bias.

choosing variables to condition on in estimating average treatment effects using unconfoundedness approaches.

However, formally, the assumption that variables are not descendants of the outcomes and treatments is not sufficient to include them in the conditioning set. Cinelli et al. (2022) show how to derive the appropriate set of conditioning variables given a fully specified graphical model. A leading case where a variable that is not a descendant of the treatment or the outcome should still not be adjusted for is illustrated in **Figure 3**. In what is labeled M-bias, one unobserved variable, U_1, has a causal effect on the treatment W but not on the outcome Y. A second unobserved variable, U_2, has a causal effect on the outcome but not on the treatment. Both unobserved variables have a causal effect on an observed variable X that itself does not directly affect either treatment or outcome. Conditioning on the observed X_3, a so-called collider, is not necessary because there is no backdoor path from the treatment on the outcome through this variable that needs to be blocked. Moreover, conditioning on this collider is harmful because it could create biases. However, there have not been many empirical studies in economics where this collider concern has led to clear mistakes.

4.1. Estimating Average Treatment Effects Under Unconfoundedness

A major part of the unconfoundedness literature focuses on estimating the average effect of the treatment. Suppose that the sample itself is drawn randomly from a large population, and define

$$\tau^{\text{pop}} = \text{E}\left[Y_i(T) - Y_i(C)\right].$$

As discussed in the introduction to Section 2, there is a minor distinction between this estimand and the average over the sample of the CATE, $\tau^{\text{sample}} = \frac{1}{N}\sum_{i=1}^{N} \text{E}[Y_i(T) - Y_i(C)|X_i]$, but I ignore this for most of this discussion because it only affects the variance, not estimation. There are also closely related settings where the estimand is the average effect for subpopulations, defined either by pretreatment variables or for the subpopulation of treated units. Finally, there is a literature focusing on estimating differences in quantiles of the marginal potential outcome distributions (Firpo 2007). Note that these differences in quantiles are in general not equal to quantiles of the unit-level treatment effects in settings with heterogeneous treatment effects unless the potential outcomes have perfect rank correlation.

If the number of distinct values of X_i is small, one may simply divide the sample by these covariate values, estimate the average effects within these subsamples, and then average over the subsamples. In practice, a major concern is that one may wish to include many covariates or pretreatment variables into the conditioning set in order to make the unconfoundedness assumption

in Equation 1 more plausible. At the same time, including more covariates into this conditioning set makes the overlap assumption in Equation 2 more controversial and thus makes the practical challenges in adjusting effectively for all the covariates more severe. Finding methods that are effective in settings with a substantial number of covariates has been one of the main goals of this literature.

Define the conditional expectation of the outcome given treatment and covariates,

$$\mu(w,x) \equiv \mathrm{E}[Y_i|W_i = w, X_i = x],$$

and the conditional expectation of the potential outcomes given covariates,

$$\mu_w(x) \equiv \mathrm{E}[Y_i(w)|X_i = x] \quad \text{so that} \quad \tau^{\text{pop}} = \mathrm{E}\left[\mu_T(X_i) - \mu_C(X_i)\right].$$

Under the unconfoundedness assumption, these two conditional expectations are equal:

$$\mu_w(x) = \mu(w,x),\ \forall w,x, \quad \text{implying} \quad \tau^{\text{pop}} = \mathrm{E}\left[\mu(T,X_i) - \mu(C,X_i)\right].$$

Given unconfoundedness and overlap, in combination with some smoothness assumptions on the propensity score and the conditional outcome expectations, one can estimate the population average treatment effect at the parametric rate. The semiparametric efficiency bound (Newey 1990, Bickel et al. 1993, Hahn 1998) is

$$\mathbb{V} = \mathrm{E}\left[\frac{(Y_i(T) - \mu_T(X_i))^2}{e(X_i)} + \frac{(Y_i(C) - \mu_C(X_i))^2}{1 - e(X_i)} + (\mu_T(X_i) - \mu_C(X_i) - \tau^{\text{pop}})^2\right].$$

This setting is remarkable for two reasons. First, there is a huge theoretical literature with many proposed estimators for this setting, ostensibly quite different, yet many (though not all) of them are semiparametrically efficient. Second, at the same time, there is a vast empirical literature where many different estimators are frequently used in practice. This setting is one of the leading examples where semiparametric efficiency bounds and corresponding estimators have been studied in the econometric literature, and as a result, many insights that carry over to other semiparametric problems have been obtained. I want to divide this literature into four subliteratures corresponding to specific classes of estimators, described in detail in the following subsections.

4.1.1. Matching estimators. The first set of estimators uses one-to-one nearest neighbor matching, and extensions thereof. For each treated (control) unit, one or more control units are selected that are similar in terms of pretreatment variables, by minimizing some metric. Formally, for a treated unit i (a unit with $W_i = T$), a match is found in the unit $j(i)$ that minimizes

$$\min_{j:W_j=C} \left\| X_i - X_j \right\|,$$

for some metric $\|\cdot\|$, often the Mahalanobis metric based on the inverse of the full-sample covariance matrix. The difference between the outcome for the treated (control) unit and its match, $Y_i - Y_{j(i)}$ (or the average outcome for its matches if there are multiple matches), is then used as an estimate of the treatment effect for that unit. These unit-level estimates are then averaged to get an estimate of the overall average effect, or the average effect for the treated. There is a large literature discussing large sample properties of matching estimators, computational concerns, and variations on the basic versions with additional bias adjustment (Abadie & Imbens 2006, Rubin 2006, Rosenbaum 2020, Zubizarreta et al. 2023).

An advantage of matching estimators is that they are intuitive and easy to explain. However, there are two drawbacks associated with simple matching estimators. First, with a fixed number of matches, these matching estimators are never fully efficient. In order to improve the precision, and in fact to reach the efficiency bound, one needs to let the number of matches increase

with the sample size (Lin et al. 2021). Second, and more important, is that with many covariates, finding good matches is challenging, and the formal bias of matching estimators does not vanish asymptotically, as shown by Abadie & Imbens (2006). It is possible to combine matching with additional bias-reduction methods based on local or linear or nonparametric regression to make the resulting estimators more attractive (Abadie & Imbens 2011). Formally, Lin et al. (2021) show that such bias-adjusted matching estimators with an increasing number of matches can reach the semiparametric efficiency bound.

4.1.2. Regression estimators. The second class of estimators first estimates the conditional expectation $\mu(w, x)$, followed by averaging the differences over the sample of the estimated conditional expectations:

$$\hat{\tau}^{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{\mu}(T, X_i) - \hat{\mu}(C, X_i)\Big).$$

The estimator for conditional expectation itself can be based on a parametric specification—say, a simple linear model, arguably still the most common estimator for estimating average treatment effects under unconfoundedness—or a more flexible approach such as kernel regression or sieve methods. Modern implementations have used machine learning methods such as deep neural nets or random forests for this component. Hahn (1998) shows that regression estimators based on sufficiently flexible specifications of the regression function are semiparametrically efficient.

4.1.3. Propensity score estimators. In applications where the number of conditioning variables, that is, the dimension of the pretreatment variables X_i, is substantial, estimating the conditional expectation $\mu(w, x)$ can be a challenge. A celebrated result from Rosenbaum & Rubin (1983b) shows there are alternatives to estimating this conditional expectation if the goal is to estimate the population average treatment effect τ^{pop}. They show that under unconfoundedness, as in Equation 1, it is also true that

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \,\Big|\, e(X_i). \qquad 3.$$

Here, one only needs to condition on a scalar function of the covariates, known as the propensity score. More generally, one can condition on any balancing score such that conditioning on this balancing score makes X_i and W_i independent, with the propensity score and one-to-one functions thereof the lowest dimensional balancing scores.

The Rosenbaum–Rubin propensity score result can be exploited in a number of ways. Two of these treat the propensity score simply as a scalar pretreatment variable that needs to be adjusted for in the two methods described in the previous two subsections. That is, one can use the matching estimators from Section 4.1.1 by matching on the propensity score rather than by matching on all the pretreatment variables. Matching on a scalar avoids the bias concerns that arise with matching estimators where one matches on multiple variables. This method is fairly widely used. For a discussion of the formal properties, readers are directed to Abadie & Imbens (2016). Alternatively, one can use the regression estimators from Section 4.1.2, where instead of using the basic covariates, one estimates the conditional expectation of the outcomes given the treatment and the propensity score, $\tilde{\mu}(w, e) = \text{E}[Y_i(w)|e(X_i) = e]$. This method is not widely used, partly because there is no natural functional form for this conditional expectation—e.g., there is no reason to expect this conditional expectation to be linear in the propensity score.

A third, more direct, method for using the propensity score result directly exploits the interpretation of the propensity score as the probability of being exposed to the treatment, rather than

viewing it simply as a balancing score. Specifically, it exploits the results that

$$\mathrm{E}\left[\frac{\mathbf{1}_{W_i=T}Y_i}{e(X_i)}\right] = \mathrm{E}[Y_i(T)], \qquad \text{and } \mathrm{E}\left[\frac{\mathbf{1}_{W_i=C}Y_i}{1-e(X_i)}\right] = \mathrm{E}[Y_i(C)].$$

This result is then used by reweighting the units by the inverse of the probability of the treatment received:

$$\hat{\tau} = \sum_{i=1}^{N} \frac{Y_i \mathbf{1}_{W_i=T}}{e(X_i)} \Bigg/ \sum_{i=1}^{N} \frac{\mathbf{1}_{W_i=T}}{e(X_i)} - \sum_{i=1}^{N} \frac{Y_i \mathbf{1}_{W_i=C}}{1-e(X_i)} \Bigg/ \sum_{i=1}^{N} \frac{\mathbf{1}_{W_i=C}}{1-e(X_i)},$$

with the weights scaled to sum to zero by treatment group. The formal properties of this Horvitz–Thompson type estimator (Horvitz & Thompson 1952) are studied by Hirano et al. (2003), who show that when using suitably flexible estimators of the propensity score, the resulting estimator reaches the semiparametric efficiency bound. The perhaps surprising insight is that estimating the propensity score here is critical: Weighting by the inverse of the true propensity score does not lead to an efficient estimator. A simple example makes this clear. Suppose the propensity score is constant, equal to p for all units. Then, weighting by the inverse of the true propensity score leads to $\hat{\tau} = (1/N)\sum_i \mathbf{1}_{W_i=T}Y_i/p - (1/N)\sum_i \mathbf{1}_{W_i=C}Y_i/(1-p)$, whereas weighting by the estimated propensity score $\hat{p} = (1/N)\sum_i \mathbf{1}_{W_i=T}$ ensures that we have a weighted average of treated and control outcomes with the weights summing to one.

4.1.4. Doubly robust estimators. Matching, regression, and inverse-propensity-score weighting estimators are all widely used in empirical work. In addition, most of them are, in principle, semiparametrically efficient. Nevertheless, the current state of the literature suggests that the most attractive estimators for the average treatment effect combine estimates of the propensity score and estimates of the conditional expectation. Formally, such estimators rely on less restrictive assumptions on the smoothness of the conditional outcome expectations and the propensity score. They do so by formally requiring lower rates of convergence of the corresponding estimators compared with regression and inverse propensity score estimators. They also have the double robustness property that when either the conditional outcome expectations or the propensity score is estimated consistently, the estimator for the average treatment effect is consistent. Estimators using such combinations of estimators for the conditional outcome expectations and the propensity score were first introduced in a series of papers by Robins and coauthors (e.g., Robins et al. 1994, 2000), who focused on the double robustness property. These estimators build on the semiparametric efficiency bound literature (Newey 1990, Bickel et al. 1993). They are also related to the literature on targeted maximum likelihood (Van der Laan & Rose 2011), where the focus is more on the efficiency properties than the robustness. More recently, various specific estimators have been proposed for average treatment effects in this setting (e.g., Chernozhukov et al. 2017, Athey et al. 2018b).

A systematic way to generate semiparametrically efficient estimators in many settings is to use the influence function. First define

$$\begin{aligned}\psi(y,w,x) &= \mu(T,x) - \mu(C,x) + \frac{(y-\mu(w,x))\mathbf{1}_{w=T}}{e(x)} - \frac{(y-\mu(w,x))\mathbf{1}_{w=C}}{1-e(x)} \\ &= \frac{y\mathbf{1}_{w=T}}{e(x)} - \frac{y\mathbf{1}_{w=C}}{1-e(x)} + \frac{e(x)-\mathbf{1}_{w=T}}{e(x)(1-e(x))}\left(\mu(T,x)(1-e(x)) + \mu(C,x)e(x)\right),\end{aligned}$$

so that $\psi(y,w,x) - \tau$ is the influence function. Then, we have

$$\hat{\tau}^{\mathrm{dr}} = \frac{1}{N}\sum_{i=1}^{N} \hat{\psi}(Y_i, W_i, X_i),$$

with estimators $\hat{\mu}(w,x)$ and $\hat{e}(x)$ for the two components. It is easy to see that this estimator is doubly robust in the sense that as long as either $\hat{\mu}(w,x)$ is consistent for $\mu(w,x)$ or $\hat{e}(x)$ is consistent for $e(x)$, the resulting estimator is consistent for the average treatment effect. To see that the estimator is consistent even if the estimator for the propensity score is inconsistent, it is useful to consider the first representation. The second pair of terms has expectation zero when evaluated at the true conditional means $\mu(w,x)$, even if the propensity score is misspecified. To see that the estimator is consistent even if the estimator for the conditional outcome expectations is inconsistent, it is useful to consider the second representation. The second pair of terms in that expression has expectation zero when evaluated at the true propensity score $e(x)$, even if the conditional outcome expectations are misspecified. Although the double robustness property may in itself not be a compelling reason for using the estimator, formal arguments show that the doubly robust estimators have good properties even when $\hat{\mu}(w,x)$ and $\hat{e}(x)$ converge relatively slowly to their population counterparts. Because the convergence rates depend on the number of covariates, this makes these doubly robust estimators particularly attractive in settings with many covariates. Readers are directed to Chernozhukov et al. (2017) and Athey et al. (2018b) for formal properties given general estimators for the conditional expectations and propensity score.

4.2. Overlap

In practice, a big concern with estimators for average treatment effects under unconfoundedness is possible violations of the overlap assumptions. This concern became clear after LaLonde (1986), where the treatment group and control group were very far apart in terms of covariate distributions. Researchers have proposed various methods for dealing with this that change the focus from the average effect in the population to some other weighted average. Crump et al. (2009) propose changing the estimand dropping units with a propensity score close to zero and one, with the threshold determined by minimizing the asymptotic variance. Formally, let α be the solution to

$$\frac{1}{\alpha(1-\alpha)} = \mathrm{E}\left[\frac{1}{e(X_i)(1-e(X_i))}\middle|\, \alpha < e(X_i) < 1-\alpha\right].$$

Crump et al. (2009) suggest changing the estimand to

$$\tau = \mathrm{E}\left[\tau(X_i)\middle|\, \alpha < e(X_i) < 1-\alpha\right].$$

Li et al. (2018), focusing on the same objective function, modify this by weighting the units optimally by a function of the pretreatment variables, which leads to weighting units by the product of the propensity score and one minus the propensity score, leading to the estimand

$$\tau = \mathrm{E}\left[\frac{e(X_i(1-e(X_i))}{\mathrm{E}[e(X_i)(1-e(X_i)]}\tau(x)\right].$$

4.3. Estimating Heterogeneous Treatment Effects and Policy Rules

Traditionally, the target of much of the literature on estimating causal effects under unconfoundedness was the population average effect $\tau^{\text{pop}} = \mathrm{E}[Y_i(T) - Y_i(C)]$. More recently, attention has been turned toward estimating the CATE, that is, the average effect conditional on covariates:

$$\tau(x) = \mathrm{E}[Y_i(T) - Y_i(C)|X_i = x].$$

Although the earlier literature already allowed for general heterogeneity in the treatment effects, there was little direct emphasis on estimating the conditional average effect, with the exception of the estimation of some average effects for prespecified subpopulations. The main reason was

that estimating the entire function $\tau(\cdot)$ requires more data than typically were available, especially in settings where even the main effects are hard to detect because of a combination of the effects being small and the sample sizes being modest. Motivated by the availability of large data sets with rich detail, interest grew in effective methods for uncovering heterogeneity in treatment effects.

The earlier literature included some attempts to estimate $\tau(x)$ through series methods (Crump et al. 2008), which were not effective in settings with a substantial number of covariates. The more recent literature instead used machine learning methods, adapted to estimating causal effects. In standard settings where supervised machine learning methods are used for estimating conditional expectations, one has observations on outcomes that are unbiased for the conditional expectations that are being estimated. That makes cross-validation methods based on leaving out some units very effective. That does not directly work in settings where the focus is on estimating average causal effects because there is no direct observation of the causal effects.

Athey & Imbens (2016) proposed constructing regression trees that were tailored to estimating CATEs in experimental settings. One proposal was based on the insight that a Horvitz–Thompson-type transformation of the outcome, $Y_i\mathbf{1}_{W_i=T}/e(X_i) - Y_i\mathbf{1}_{W_i=C}/(1-e(X_i))$, has conditional expectation given $X_i = x$ equal to $\tau(x)$. Thus, if we first transform the outcome, and note that this transformation is known in the experimental case, then we can directly use methods for supervised learning, such as regression trees. Wager & Athey (2018) generalized these methods to random forests, which, importantly, allowed for honest inference without prespecifying which covariates are used to create subpopulations. These causal random forests estimators are now widely used in practice.

Athey & Wager (2021), Dehejia (2005), Hirano & Porter (2009), Manski (2004), and Kitagawa & Tetenov (2018) change the focus away from estimating the CATE $\tau(x)$ to estimating policy rules that assign units to the treatment based on the values of their pretreatment variables. The goal in this literature is to find estimators for the optimal policy rules that perform well when the CATE function is unknown. Athey & Wager (2021) show that this problem reduces to one that is very similar to that of estimating average treatment effects. Critical is the complexity of the class of policy rules that the researcher optimizes over.

4.4. Sensitivity Analyses and Bounds

Although the setting with unconfoundedness is a central one, it is clear that in many cases researchers are not confident that this assumption holds exactly. One strand of the literature has focused on clarifying how misleading analyses assuming unconfoundedness can be when the assumption does not actually hold. Researchers have approached this problem from three directions. One is by abandoning the unconfoundedness assumption entirely. This implies giving up on point identification and focusing on partial identification, or the estimation of bounds for the average treatment effects. This approach is associated with the work by Manski (in particular, Manski 1990, 2003). A second approach starts with unconfoundedness but relaxes it by assuming it holds only conditional on an unobserved covariate that has limited association with the potential outcomes and assignment. This line of research was initiated by Rosenbaum & Rubin (1983a). A third line of research, initiated by Rosenbaum (2002), also starts with unconfoundedness but avoids any attempt to restrict the association between the potential outcomes and the unobserved covariates and only limits the assignment probabilities.

First, consider the bounds approach. Following Manski (1990), suppose the potential outcomes are binary, $Y_i(w) \in \{0, 1\}$. Using iterated expectations, the marginal expectation of $Y_i(T)$ can be written as

$$\mathrm{E}[Y_i(T)] = \mathrm{E}[Y_i(T)|W_i = T]\mathrm{pr}(W_i = T) + \mathrm{E}[Y_i(T)|W_i = C]\mathrm{pr}(W_i = C).$$

Of the four components in this decomposition, three can be estimated consistently. The data are entirely uninformative about the fourth, $\mathrm{E}[Y_i(T)|W_i = C]$, which is only known to lie between 0 and 1 because the outcome is assumed to be binary. Thus, the bounds for $\mathrm{E}[Y_i(T)]$ are

$$\mathrm{E}[Y_i(T)|W_i = T]\mathrm{pr}(W_i = T) \leq \mathrm{E}[Y_i(T)] \leq \mathrm{E}[Y_i(T)|W_i = T]\mathrm{pr}(W_i = T) + \mathrm{pr}(W_i = C).$$

A similar decomposition can be derived for the expectation of $Y_i(C)$, leading to bounds on the population average treatment effect,

$$\begin{aligned}\mathrm{E}[Y_i(T)|W_i = T]\mathrm{pr}(W_i = T) - \mathrm{pr}(W_i = T) - \mathrm{E}[Y_i(C)|W_i = C]\mathrm{pr}(W_i = C) \leq \tau^{\mathrm{pop}} \\ \leq \mathrm{E}[Y_i(T)|W_i = T]\mathrm{pr}(W_i = T) + \mathrm{pr}(W_i = C) - \mathrm{E}[Y_i(C)|W_i = C]\mathrm{pr}(W_i = C).\end{aligned}$$

These bounds are generically wide: Their width is, in this case, always equal to one, because the width for the bounds for $\mathrm{E}[Y_i(T)]$ is $\mathrm{pr}(W_i = C)$ and the width for the bounds for $\mathrm{E}[Y_i(C)]$ is $\mathrm{pr}(W_i = T)$. This, in turns, implies that the bounds for the average treatment effect will always include zero.

The Rosenbaum–Rubin sensitivity analysis starts with a modified unconfoundedness assumption:

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \,\Big|\, X_i, U_i.$$

The difference between this independence condition and the unconfoundedness condition in Equation 1 is that the second confounder, U_i, is not observed. Without loss of generality, we can take U_i to be binary. We cannot consistently estimate the average effect of the treatment under this assumption because U_i is not observed. We therefore augment it with additional assumptions on the relation between the potential outcomes, treatment assignment, and covariates given the unobserved confounder. In spirit, this is similar to the standard analysis of omitted variable bias in linear regression models. Suppose we are interested in the coefficient on W_i in a (long) regression,

$$Y_i = \beta_0 + \beta_W \mathbf{1}_{W_i=T} + \beta_X X_i + \beta_U U_i + \varepsilon_i,$$

but, because we do not observe U_i, we estimate the (short) regression,

$$Y_i = \alpha_0 + \alpha_W \mathbf{1}_{W_i=T} + \alpha_X X_i + \eta_i,$$

instead. The bias in estimating β_W by an estimator for α_W, $\beta_W - \alpha_W$, is equal to the coefficient on the omitted variable, β_U, times the coefficient δ_W on W_i in a regression of the omitted variable U_i on the included variables (an intercept, the treatment indicator, and X_i),

$$U_i = \delta_0 + \delta_W \mathbf{1}_{W_i=T} + \delta_X X_i + \nu_i.$$

Without information on U_i, we cannot estimate the magnitude of the bias as $\beta_W - \alpha_W = \beta_U \times \delta_W$, but we can speculate regarding the magnitudes of the components. In the Rosenbaum–Rubin sensitivity analyses, we do a similar exercise. We make assumptions on the relation between the omitted variable and the potential outcomes, β_U in the linear model, and we make assumptions on the relation between the omitted variable and the assignment, δ_W in the linear model, and explore the resulting range of values for τ^{pop}. In both cases, the assumptions are in the form of ranges of possible values.

The key to this approach to sensitivity analyses is the choice of the range of possible values. Without putting any limits on this range, one gets back to the Manski bounds analysis that drops the unconfoundedness assumption entirely. For the binary outcome case, Rosenbaum and Rubin focus on the log odds ratio within subpopulations. To see how this works, consider a case without

covariates and binary outcomes. Suppose the unobserved confounder U_i is binary. Then, we can model the probability of assignment in a parametric, logistic framework through the log odds ratio:

$$\ln\left(\frac{\text{pr}(W_i = T|U_i = u)}{\text{pr}(W_i = C|U_i = u)}\right) = \alpha_0 + \alpha_U \cdot u.$$

Similarly, we model the potential outcome distribution as

$$\ln\left(\frac{\text{pr}(Y_i(w) = 1|U_i = u)}{\text{pr}(Y_i(w)|U_i = u)}\right) = \beta_{w0} + \beta_{wU} \cdot u.$$

Now, given fixed values for the sensitivity parameters (α_U, β_{CU}, β_{TU}) and the data, we can first estimate the remaining parameters α_0 and β_{w0} and thus the average treatment effect. This defines a function

$$\hat{\tau}(\alpha_U, \beta_{CU}, \beta_{TU}, \text{data}).$$

Because the sensitivity parameters are defined in terms of log odds ratio, a willingness to put limits on the effect of the unobserved confounder on the log odds ratio then determines the function $\hat{\tau}(\alpha_U, \beta_{CU}, \beta_{TU}, \text{data})$. The key question is how to choose a reasonable range of values for the sensitivity parameters (α_U, β_{CU}, β_{TU}).

Imbens (2003) suggests limiting the range of plausible values for the sensitivity parameters by inspecting the association between observed confounders and assignment and potential outcomes. Specifically, the suggestion is to find the strongest association, in a logistic model, between the observed confounders and the assignment, and the strongest association between the observed confounders and the potential outcomes, and assume that the unobserved confounders do not have a stronger association with either assignment or potential outcomes than that. This work has been extended by Oster (2019), Cinelli & Hazlett (2020), Manski (1990), Masten et al. (2020), and Chernozhukov et al. (2022).

Rosenbaum (2002) develops a sensitivity analysis that does not require assumptions on the association between the unobserved confounder and the potential outcomes. In the example without covariates, his approach bounds the log odds ratio for the assignment probabilities:

$$\Gamma \leq \ln\left(\frac{\text{pr}(W_i = T)}{\text{pr}(W_i = C)}\right) \leq 1/\Gamma,$$

and then explores the range of possible estimates for $\hat{\tau}$ associated with the limit Γ. Implicitly, this allows the association between the unobserved confounder and the potential outcomes to be arbitrarily strong.

5. OBSERVATIONAL STUDIES WITHOUT UNCONFOUNDEDNESS

The setting with unconfoundedness is a natural extension of the setting with randomized assignment, and it remains the most widely used case in empirical work. Within subpopulations that are homogenous in a set of preselected covariates, the unconfoundedness assumption implies we can analyze the data as if the assignment was random. However, it is settings where unconfoundedness is not plausible that have traditionally been of great interest in econometrics. Early econometricians, such as Tinbergen (1930), studied cases where random assignment of treatments or causes, even within homogenous subpopulations, was not plausible. Active choices by optimizing agents, rather than chance assignments by researchers, were key drivers of the assignment mechanism in studies of the effect of prices on demand, and the effect of education on earnings. This has led to a

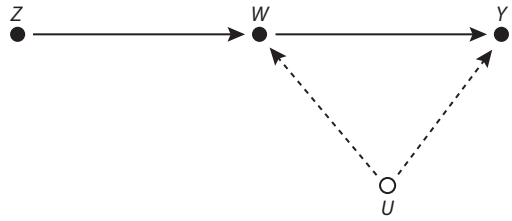

Figure 4

Directed acyclic graph for instrumental variables.

large literature in econometrics that focuses on methods that do not start from unconfoundedness assumptions.

There is no general solution for this case. As discussed before, and illustrated by the Manski bounds, simply dropping the unconfoundedness assumption implies average treatment effects are no longer identified, although informative bounds can be derived in some cases. Much of the econometrics literature has focused on special cases where additional information of some kind is available. This can be in the form of additional variables with specific causal structure, or in the form of additional assumptions placed on either the assignment mechanism or the potential outcome distributions, so that either the overall average treatment effect or some other estimand is identified. Much of the empirical work relies on a small set of what Angrist & Krueger (1999) call identification strategies. Occasionally, new strategies are proposed and make it into the toolkit of empirical researchers in social sciences. Here, we describe three of the leading strategies: instrumental variables, fixed effect and DID methods, and regression discontinuity designs.

5.1. Instrumental Variables

One of the earliest and conceptually most important approaches allowing for violations of unconfoundedness is instrumental variables. This method was first used in the 1930s in work by Wright (1934) and Tinbergen (1930) and has been a key identification strategy in economics ever since. The starting point is a concern about the presence of an unobserved confounder that is correlated with the treatment as well as with the outcome. This implies that comparing average outcomes for treated and control units will not estimate the average effect of the treatment but will instead estimate a combination of the effect of the treatment and the correlation between confounder and outcome. To address this endogeneity, the researcher uses the presence of a third observed variable, the instrument. This instrument is assumed to have a direct causal effect on the treatment but no direct effect on the outcome. **Figure 4** illustrates a DAG with this structure.

The DAG for an instrumental variables setting is closely related to that for a mediation setting (VanderWeele 2015), as in **Figure 5**.

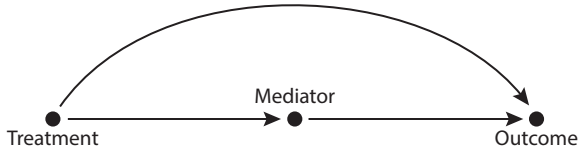

Figure 5

Directed acyclic graph with mediation.

The first difference is that what is the treatment in the instrumental variables case is, in the mediation setting, the mediator, and what is the instrument in the instrumental variables case is the treatment in the mediation setting. In addition, there is—and this is critical—no direct effect of the instrument on the outcome. Finally, there is an unobserved confounder that makes a direct comparison between treatment and outcome impossible, and that is the motivation to look for an instrument.

One of the most celebrated applications where this structure is plausible (and which is, in fact, referenced in Angrist's Nobel citation by the prize committee) is the draft lottery example of Angrist (1990), where the treatment is military service status, the outcome is earnings later in life, and the instrument is draft eligibility determined by the draft lottery number. In this case, it appears plausible that the effect of the lottery number on earnings is entirely, or at least largely, mediated by military service. A second classic example is Angrist & Krueger (1991), where the focus is on estimating the effect of years of education on earnings, using compulsory schooling laws as instruments. Another class of examples includes randomized experiments with imperfect compliance. Again, the key assumption is that the effect of the random assignment to treatment is entirely mediated by the receipt of the treatment.

One cannot simply compare outcomes by treatment status in an as-treated analysis because of the presence of an unobserved confounder U. We also cannot simply drop those who are observed not to comply with their treatment assignment in a per-protocol analysis. However, because there is no unobserved confounder for the relation between the additional variable, the instrument, and the treatment, we can estimate the average causal effect of the instrument on the treatment, the intention-to-treat effect. Similarly, there is no unobserved confounder for the relation between the instrument and the outcome as there is in **Figure 6** where the instrument assumptions do not hold. In addition, there is no direct effect of the instrument on the outcome, no arrow from Z to Y, as there is in **Figure 7**. This implies we can estimate the average causal effect of the instrument on the outcome. These two intention-to-treat effects, however, are not the primary estimand in these settings. Instead, we are interested in the causal effect of the treatment on the outcome.

In terms of potential outcomes, the instrumental variables analysis starts with two sets of potential outcomes: For the treatment, we have $W_i(z)$ for each value of the instrument, and for the outcomes, we have $Y_i(w, z)$ indexed by both the treatment and the instrument. The key assumptions are now that the potential outcomes are all independent of the instrument Z_i, which is an unconfoundedness type assumption. Second, the potential outcomes $Y_i(z, w)$ do not actually vary

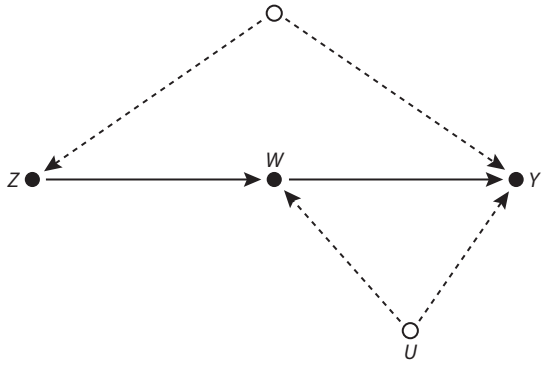

Figure 6

Directed acyclic graph with violation of the exogeneity assumption for the instrument.

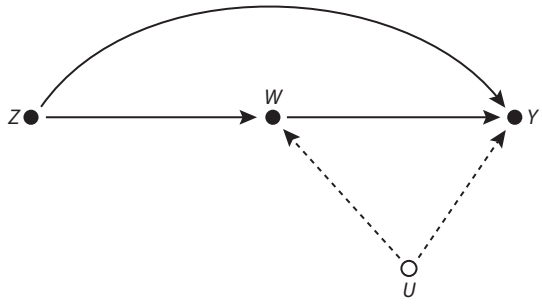

Figure 7

Directed acyclic graph with violation of the exclusion restriction because of a direct effect of the instrument on the outcome.

by the instrument, so we can drop the z argument and write $Y_i(w)$. This is the exclusion restriction captured in the graph by the absence of a direct link between the instrument and the outcome, with the treatment itself acting as a mediator.

These assumptions by themselves are not sufficient to identify the average effect of the treatment on the outcome, as pointed out by Heckman (1990) and Manski (1990). To make further progress, Imbens & Angrist (1994) added one more assumption, what they labeled monotonicity. This requires that the causal effect of the instrument on the treatment is in the same direction for all units. That is, changing the instrument from 0 to 1 can leave the treatment status for a unit unchanged, or it can move the unit from the untreated state to the treated state, but it cannot move the unit from the treated state to the untreated state. Formally, the assumption requires that for all i, $W_i(1) \geq W_i(0)$. Even adding this monotonicity assumption does not allow for point identification of the average effect of the treatment, but the combination of assumptions does allow for the identification of the average effect of the treatment for the subpopulation of units for whom changing the instrument from 0 to 1 moves them from untreated to treated. This subpopulation is generally referred to as the compliers, and the average effect for this group is the local average treatment effect (LATE) (Imbens & Angrist 1994, Angrist et al. 1996, Imbens 2014),

$$\tau^{\text{LATE}} = \text{E}[Y_i(T) - Y_i(C) | W_i(1) = T, W_i(0) = C].$$

This identification result is unusual because the resulting estimand, the LATE, is unconventional. There is no particular reason why the subpopulation it refers to, the compliers, is necessarily an interesting subpopulation. It may be, and in the Angrist draft lottery example it arguably is, but that is not the reason for focusing on it. The main reason is that it is the only subpopulation for which we can identify the average effect of the treatment. We may be more interested in the overall average effect, but we cannot identify that without substantially stronger assumptions, e.g., constant treatment effects.

Note that the monotonicity assumption is difficult to capture in the graphical representation. However, it is plausible in many applications, and similar shape restrictions (monotonicity of demand or supply functions, convexity of preferences or production functions, decreasing returns to scale) play an important role in econometric identification strategies (e.g., Matzkin 1994).

Concerns with the exclusion restriction have often driven researchers to use instruments for which that assumption may be plausible, but that have only limited effects on the treatment. (In the limit, a random number would by definition satisfy the exclusion restriction, but it would not have a causal effect on the treatment, so there would be no compliers.) This has led to concerns about the properties of instrumental variables estimators when the instruments are weak, in the sense of having only a weak correlation with the treatment. In that case, instrumental variables estimators

have poor properties, and the standard Normal-distribution-based confidence intervals may not be valid (Staiger & Stock 1997, Andrews et al. 2019).

Researchers have also studied quantile regression estimators in instrumental variables settings, including Abadie et al. (2002) and Chernozhukov & Hansen (2005).

5.2. Fixed Effect, Difference-in-Differences, and Synthetic Control Methods

One of the most common identification strategies in the empirical economics literature is referred to as difference-in-differences (DID). There are multiple settings where these methods apply, and also variations on the specific implementation, some referred to as two-way-fixed-effect (TWFE) methods. Somewhat more distantly related are the SC methods recently developed by Abadie & Gardeazabal (2003) and Abadie et al. (2010), and currently widely used.

I start with a discussion of a canonical case for DID with panel data. The researcher observes outcomes for units in two subpopulations, a treated and a control subpopulation. Units are observed in two periods, a pretreatment period and a posttreatment period, with Y_{it} denoting the observed outcome for unit i in period t. Only for units in the treatment group in the second period do we observe the treated outcome, $Y_{it}(T)$. For the other three group/time periods, we observe the control outcome $Y_{it}(C)$. In this case, the DID estimator is

$$\hat{\tau}^{\text{DID}} = \left(\overline{Y}_{T,\text{post}} - \overline{Y}_{C,\text{post}}\right) - \left(\overline{Y}_{T,\text{pre}} - \overline{Y}_{C,\text{pre}}\right).$$

Here, the four terms are all averages—e.g., $\overline{Y}_{T,\text{post}}$ is the average outcome for units in the treatment group who were observed in the posttreatment period. The DID estimator can be motivated by a TWFE model for the potential outcomes that has an additive fixed effect for the group (treatment versus control) and an additive fixed effect for the time period (post versus pre):

$$Y_{it}(C) = \beta_i + \gamma_t + \varepsilon_{it}. \tag{4.}$$

Combined with an additive treatment, $Y_{it}(T) = Y_{it}(C) + \tau$, this leads to a regression in terms of the realized outcome of the form

$$Y_{it} = \beta_i + \gamma_t + W_{it}\tau + \varepsilon_{it}.$$

This setup extends naturally to allow for multiple time periods where we can still use the specification in Equation 4. This includes settings where the treated units do not all receive the treatment in the same period. For example, a common setting is that with staggered adoption (Athey & Imbens 2021), where units enter the treatment group at different times, but once they are in the treatment group they remain there for all subsequent periods. This setting is also known as a stepped wedge design. In this case, the standard TWFE estimator may estimate a weighted average of treatment effects with some of the weights negative. This has led to a number of alternative estimators (Callaway & Sant'Anna 2020, Goodman-Bacon 2021, Sun & Abraham 2021). Recent extensions are discussed by de Chaisemartin & d'Haultfœuille (2020), Freyaldenhoven et al. (2019), Imai & Kim (2019), Liu et al. (2022), Sant'Anna & Zhao (2020), and Roth et al. (2022).

Abadie & Gardeazabal (2003) and Abadie et al. (2010) focused on a different special case of this setup, where a single unit was treated from a point in time onwards. In what they called the synthetic control (SC) approach, the central idea was to approximate the treated unit by a synthetic version consisting of a convex combination of the control units. If unit N is treated in period T, its counterfactual outcome $Y_{NT}(0)$ is estimated as

$$\hat{Y}_{N_T}(0) = \sum_{i=1}^{N-1} \omega_i Y_{iT},$$

with the weights ω_i nonnegative and summing to one. In the case without additional covariates, the weights are chosen to minimize the difference between the synthetic control and the pretreatment outcomes for the treated units:

$$\omega = \arg \min_{w: w_j \geq 0, \sum_{j=1}^{N-1} w_j = 1} \sum_{t=1}^{T-1} \left(Y_{Nt} - \sum_{i=1}^{N-1} w_i Y_{it} \right)^2 .$$

Various modifications and extensions of the basic SC method have been proposed. Two of the most important ones are, first, allowing for an intercept in the objective function to allow for permanent stable differences between the treated unit and the convex combination of the other units, and second, allowing for some negative weights (Doudchenko & Imbens 2016, Ferman & Pinto 2021). Both imply that the SC can be outside of the convex hull of the control units. A Bayesian approach was introduced by Brodersen et al. (2015). Chernozhukov et al. (2021) discuss applications of conformal inference to SC settings. Xu (2017) and Athey et al. (2021) discuss general factor models and their relation to SC methods.

Arkhangelsky et al. (2021) combine the DID and SC approaches in the synthetic DID estimator. Here, the SC weights are combined with a TWFE model for the outcomes:

$$(\hat{\tau}, \hat{\alpha}, \hat{\beta}) = \arg \min_{\alpha, \beta, \tau} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \alpha_i - \beta_t - \tau W_{it})^2 \, \omega_i \lambda_t,$$

where the SC weights ω_i are calculated as before. In addition, time weights λ_t are used to put more emphasis on time periods that are similar to the time periods where the treatment occurs. This combination of an outcome model with the SC weighting leads to more robust estimates.

5.3. Regression Discontinuity Designs

A third identification strategy for settings where unconfoundedness is not plausible is referred to as regression discontinuity designs. This is a set of methods that originated in the 1960s in the psychology literature (Thistlewaite & Campbell 1960). Cook (2008) provides a historical perspective. It became popular in the social science literature in the early 2000s, with some influential applications including those of Black (1999), Van Der Klaauw (2002), and Lee et al. (2004). Subsequently, it has become a very popular method in the empirical literature, and there have been substantial advances in terms of methodology.

The canonical setting is one where receipt of the treatment is a deterministic function of a pretreatment variable, the running variable, denoted by X_i, switching at some threshold c from control to treatment:

$$W_i = \begin{cases} C & \text{if } X_i \leq c, \\ T & \text{if } X_i > c. \end{cases}$$

Assuming that the conditional distribution, and in particular the conditional expectation, of the potential outcomes given X_i is smooth in the covariate, the average effect of the treatment for units with $X_i = c$ is

$$\tau = \lim_{x \to c} \mathrm{E}[Y_i(T) - Y_i(C) | X_i = x] = \lim_{x \downarrow c} \mathrm{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathrm{E}[Y_i | X_i = x].$$

We can then estimate the conditional expectation $\mathrm{E}[Y_i | X_i = x]$ at the two limits to get an estimator for the average causal effect τ at the threshold.

This setting arises very naturally in social science applications, and the identification strategy often has great credibility there. In education settings, administrators often use thresholds for

test scores to offer different menus of options for students, including access to selective schools (Abdulkadıroğlu et al. 2022) or required attendance in summer programs (Matsudaira 2008). Elections also create sharp thresholds that allow for the evaluation of the effect of incumbency (Lee et al. 2004).

Since the work by Hahn et al. (2001) and Porter (2003), the most common estimator for τ is based on local linear regression using observations close to, on the left and on the right of, the threshold c. The local linear regression has largely replaced the global polynomial methods, which are sensitive to the choice of degree of the polynomial as documented by Gelman & Imbens (2018). The key choice in implementation is the bandwidth for the local linear estimator. Since the rediscovery of regression discontinuity designs in the early 2000s, in the social science literature, various algorithms have been proposed for this choice (Imbens & Kalyanaraman 2012, Calonico et al. 2014).

A second case of interest is the fuzzy regression discontinuity design. Here, the probability of exposure to the treatment does not jump from zero to one at the threshold. Rather, there is a discontinuity at the threshold, but only a limited one. In this case, the estimand is the ratio of the jump in the average outcome, scaled by the magnitude of the jump in the probability of receipt of the treatment. The interpretation is now that this estimates, under some regularity conditions, the average effect of the treatment, for a subpopulation from the population of units with covariate values at the threshold. This subpopulation is like the compliers in the instrumental variables setting, consisting of units who are affected by being on the right versus the left of the threshold. The estimator for the fuzzy regression discontinuity setting is simply the ratio of two estimates of the differences in regression functions at the threshold. The numerator is the magnitude in the discontinuity of the expected value of the outcome, and the denominator is the estimated magnitude in the discontinuity of the expected value of the treatment. This setting is common, though perhaps understudied, in biomedical settings. Guidelines for treating patients often use somewhat arbitrary thresholds based on tests or age for recommending in favor of or against particular procedures.

Recently there has been important work on alternatives to local linear estimators. This work focuses on the characterization of the estimator as a weighted average of the outcomes to the right minus a weighted average of the outcomes to the left of the threshold. Choices of estimators correspond to choices of weight functions. This literature has attempted to characterize optimal choices for weights by minimizing the expected squared error under the worst-case scenario for the conditional expectation of the outcome given the covariate. These approaches deal very effectively with settings with discrete as well as continuous covariates and make the reliance on smoothness assumptions explicit (see Armstrong & Kolesár 2018, Imbens & Wager 2018).

To assess the plausibility of regression discontinuity designs, some specific procedures have been proposed. These are useful in cases with concerns that the value of the running variable might have been manipulated. Given such manipulation, one would expect to see that the distribution of the running variable is discontinuous around the threshold. The McCrary test (McCrary 2008) formalizes this. One can also test whether the expected value for other covariates is discontinuous at the threshold for the running variable.

A practical concern with regression discontinuity estimators is their limited external validity. They are only valid for units close to the threshold, and in the fuzzy case only for the complier subpopulation of those units. Angrist & Rokkanen (2015) and Bertanha & Imbens (2020) assess the plausibility of extrapolating to larger subpopulations, in particular away from the threshold. This involves inspecting discontinuities in the conditional expectation of the outcomes given treatment status as a function of the covariate at the threshold. Smoothness of this conditional expectation implies that control compliers and never-takers are not substantially different, and

treated compliers and always-takers are not substantially different, which makes it more likely we can extrapolate the effect away from the threshold.

6. COMBINING EXPERIMENTAL AND OBSERVATIONAL DATA

One interesting recent area of research investigates systematic ways of combining experimental and observational data. Here I discuss two specific examples.

6.1. Surrogacy

Gupta et al. (2019) discuss as one of the main challenges in online experimentation the problem of estimating long-term causal effects from short experiments. Experimenters often want to act on results quickly but may wish to optimize for long-term outcomes. During the experiment they may be able to measure a number of short-term outcomes that are all related to the primary (long-term) outcome. The question arises of how to combine these multiple short-term outcomes into a single variable that can be used to decide on the efficacy of the intervention. Athey et al. (2020a) suggest combining the short-term variables into a single predictor of the long-term outcome. They consider a setting with two samples, one experimental, where we observe the treatment and the short-term outcomes, the surrogates, but not the primary outcome, as illustrated in **Figure 8*b***, and an observational sample, where we observe the surrogates and the primary outcome, but not the treatment, as illustrated in **Figure 8*a***. This can be motivated by the assumption that the short-term variables are valid surrogates in the sense of Prentice (1989). The key component of the assumption is that all causal paths from the treatment to the outcome go through at least one of the surrogates, so that there is no direct causal effect of the treatment on the outcome, only indirect effects through the surrogates, as illustrated in **Figure 8**. In the case where all the effects are linear, this leads to a Baron & Kenny (1986)–type approach where the causal effects on the surrogates are weighted by their coefficients in a predictive regression.

6.2. Using Experiments to Remove Selection Bias

The difference between the first and second setting we consider is that in the second case, in the observational data set, we also observe the treatment, as illustrated in **Figure 9*a***. This gives us additional options to model the process. Athey et al. (2020b) assume that in the observational study, there is an unobserved confounder that invalidates direct comparisons of primary and secondary outcomes by treatment status. However, they put some structure on this unobserved confounder, namely that it is the same for both the primary and secondary outcome, as illustrated in **Figure 9**. The specific structure they impose on this unobserved confounder allows them to extract the unobserved confounder from the observational and experimental data on the

Figure 8

(*a*) Data generation for observational data. We observe in the observational data the surrogates and the primary outcome, but not the treatment. (*b*) Data generation for experimental data. For the experimental data, we observe the treatment and the surrogates but not the primary outcome.

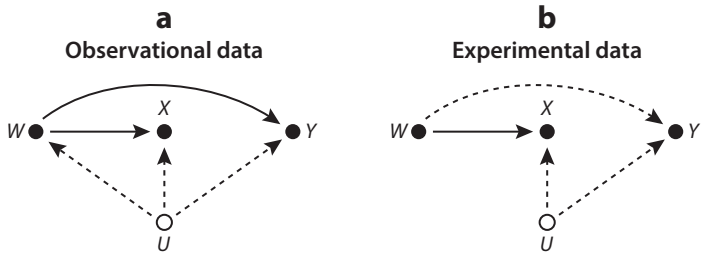

Figure 9

(*a*) Observational data. For the observational data, we observe all three variables, the treatment, the secondary outcome, and the primary outcome. (*b*) Experimental data. For the experimental sample, we observe only the treatment and the secondary outcome.

treatment and the secondary outcome. They then adjust for the estimated unobserved confounder in the observational data to estimate the average effect of the treatment in the observational study.

7. CONCLUSION

The literature on causal inference in statistics and social sciences has been a fast growing-one in the past twenty years. With close interactions between methodologists and empirical researchers, in a variety of disciplines including political science, economics, statistics, and computer science, the credibility of empirical work has been substantially improved. This trend shows no sign of slowing down, with important advances being made in the literature on spillovers in observational studies, and the estimation and inference of dynamic effects in panel data.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was partially supported by the Office of Naval Research under grant N00014-17-1-2131.

LITERATURE CITED

Abadie A, Angrist J, Imbens G. 2002. Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica* 70(1):91–117

Abadie A, Athey S, Imbens G, Wooldridge J. 2020. Sampling-based versus design-based uncertainty in regression analysis. *Econometrica* 88:265–96

Abadie A, Cattaneo M. 2018. Econometric methods for program evaluation. *Annu. Rev. Econ.* 10:465–503

Abadie A, Diamond A, Hainmueller J. 2010. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *J. Am. Stat. Assoc.* 105(490):493–505

Abadie A, Gardeazabal J. 2003. The economic costs of conflict: a case study of the Basque Country. *Am. Econ. Rev.* 93:113–32

Abadie A, Imbens G. 2006. Large sample properties of matching estimators for average treatment effects. *Econometrica* 74(1):235–67

Abadie A, Imbens G. 2011. Bias-corrected matching estimators for average treatment effects. *J. Bus. Econ. Stat.* 29(1):1–11

Abadie A, Imbens G. 2016. Matching on the estimated propensity score. *Econometrica* 84(2):781–807

Abdulkadıroğlu A, Angrist J, Narita Y, Pathak P. 2022. Breaking ties: regression discontinuity design meets market design. *Econometrica* 90(1):117–51

Andrews I, Stock J, Sun L. 2019. Weak instruments in instrumental variables regression: theory and practice. *Annu. Rev. Econ.* 11:727–53

Angrist J. 1990. Lifetime earnings and the Vietnam era draft lottery: evidence from Social Security administrative records. *Am. Econ. Rev.* 80(3):313–36

Angrist J, Imbens G, Rubin D. 1996. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* 91:444–72

Angrist J, Krueger A. 1991. Does compulsory schooling affect schooling and earnings? *Q. J. Econ.* 106(4):979–1014

Angrist J, Krueger A. 1999. Empirical strategies in labor economics. In *Handbook of Labor Economics*, Vol. 3, ed. OC Ashenfelter, D Card, pp. 1277–366. Amsterdam: Elsevier

Angrist J, Rokkanen M. 2015. Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *J. Am. Stat. Assoc.* 110(512):1331–44

Arkhangelsky D, Athey S, Hirshberg D, Imbens G, Wager S. 2021. Synthetic difference-in-differences. *Am. Econ. Rev.* 111(12):4088–118

Armstrong T, Kolesár M. 2018. Optimal inference in a class of regression models. *Econometrica* 86(2):655–83

Aronow P, Samii C. 2017. Estimating average causal effects under general interference, with application to a social network experiment. *Ann. Appl. Stat.* 11(4):1912–47

Athey S, Bahati M, Doudchenko N, Imbens G, Khosravi K. 2021. Matrix completion methods for causal panel data models. *J. Am. Stat. Assoc.* 116(536):1716–30

Athey S, Chetty R, Imbens G, Kang H. 2020a. Estimating treatment effects using multiple surrogates: the role of the surrogate score and the surrogate index. arXiv:1603.09326 [stat.ME]

Athey S, Chetty R, Imbens G. 2020b. Combining experimental and observational data to estimate treatment effects on long term outcomes. arXiv:2006.09676 [stat.ME]

Athey S, Eckles D, Imbens GW. 2018a. Exact *p*-values for network interference. *J. Am. Stat. Assoc.* 113(521):230–40

Athey S, Imbens G. 2016. Recursive partitioning for heterogeneous causal effects. *PNAS* 113(27):7353–60

Athey S, Imbens G. 2017. The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, Vol. 1, ed. AV Banerjee, E Duflo, pp. 73–140. Amsterdam: Elsevier

Athey S, Imbens G. 2021. Design-based analysis in difference-in-differences settings with staggered adoption. *J. Econom.* 226(1):62–79

Athey S, Imbens G, Wager S. 2018b. Approximate residual balancing. *J. R. Stat. Soc. Ser. B* 80(4):597–623

Athey S, Wager S. 2021. Policy learning with observational data. *Econometrica* 89(1):133–61

Bajari P, Burdick B, Imbens G, Masoero L, McQueen J, et al. 2021. Multiple randomization designs. arXiv:2112.13495 [stat.ME]

Bajari P, Burdick B, Imbens G, Masoero L, McQueen J, et al. 2023. Experimental design in marketplaces. *Stat. Sci.* 38(3):458–76

Banerjee A. 2020. Field experiments and the practice of economics. *Am. Econ. Rev.* 110(7):1937–51

Baron R, Kenny D. 1986. The moderator–mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51(6):1173–82

Basse GW, Feller A, Toulis P. 2019. Randomization tests of causal effects under interference. *Biometrika* 106(2):487–94

Bertanha M, Imbens G. 2020. External validity in fuzzy regression discontinuity designs. *J. Bus. Econ. Stat.* 38(3):593–612

Besbes O, Gur Y, Zeevi A. 2014. Stochastic multi-armed-bandit problem with non-stationary rewards. In *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, ed. Z Ghahramani, M Welling, C Cortes, ND Lawrence, KQ Weinberger, pp. 199–207. Cambridge, MA: MIT Press

Bickel P, Klaassen C, Ritov Y, Wellner J. 1993. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins Univ. Press

Brodersen K, Gallusser F, Koehler J, Remy N, Scott S. 2015. Inferring causal impact using Bayesian structural time-series models. *Ann. Appl. Stat.* 9(1):247–74

Black S. 1999. Do better schools matter? Parental valuation of elementary education. *Q. J. Econ.* 114(2):577–99
Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, et al. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–98
Callaway B, Sant'Anna P. 2020. Difference-in-differences with multiple time periods. *J. Econom.* 225(2):200–30
Calonico S, Cattaneo M, Titiunik R. 2014. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82(6):2295–326
Chamberlain G. 1984. Panel data. In *Handbook of Econometrics*, Vol. 2, ed. Z Griliches, MD Intriligator, pp. 1247–318. Amsterdam: Elsevier
Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W. 2017. Double/debiased/Neyman machine learning of treatment effects. *Am. Econ. Rev.* 7(5):261–65
Chernozhukov V, Cinelli C, Newey W, Sharma A, Syrgkanis V. 2022. Long story short: omitted variable bias in causal machine learning. arXiv:2112.13398 [econ.EM]
Chernozhukov V, Demirer M, Duflo E, Fernandez-Val I. 2018. *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India*. NBER Work. Pap. 24678
Chernozhukov V, Hansen C. 2005. An IV model of quantile treatment effects. *Econometrica* 73(1):245–61
Chernozhukov V, Wüthrich K, Zhu Y. 2021. An exact and robust conformal inference method for counterfactual and synthetic controls. *J. Am. Stat. Assoc.* 116(536):1849–64
Cinelli C, Forney A, Pearl J. 2022. A crash course in good and bad controls. *Sociol. Methods Res.* In press. **https://doi.org/10.1177/00491241221099552**
Cinelli C, Hazlett C. 2020. Making sense of sensitivity: extending omitted variable bias. *J. R. Stat. Soc. Ser. B* 82(1):39–67
Cook T. 2008. Waiting for life to arrive: a history of the regression-discontinuity design in psychology, statistics and economics. *J. Econom.* 142(2):636–54
Crépon B, Duflo E, Gurgand M, Rathelot R, Zamora P. 2013. Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *Q. J. Econ.* 128(2):531–80
Crump RK, Hotz VJ, Imbens GW, Mitnik OA. 2008. Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.* 90(3):389–405
Crump RK, Hotz VJ, Imbens GW, Mitnik OA. 2009. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 96(1):187–99
Cunningham S. 2018. *Causal Inference: The Mixtape*. New Haven, CT: Yale Univ. Press
Currie J, Kleven H, Zwiers E. 2020. Technology and big data are changing economics: mining text to track methods. *AEA Pap. Proc.* 110:42–48
de Chaisemartin C, d'Haultfœuille X. 2020. Two-way fixed effects estimators with heterogeneous treatment effects. *Am. Econ. Rev.* 110(9):2964–96
Dehejia R. 2005. Program evaluation as a decision problem. *J. Econom.* 125(1):141–73
Dimakopoulou M, Zhou Z, Athey S, Imbens G. 2018. Estimation considerations in contextual bandits. arXiv:1711.07077 [stat.ML]
Doudchenko N, Imbens G. 2016. *Balancing, regression, difference-in-differences and synthetic control methods: a synthesis*. NBER Work. Pap. 22791
Duflo E. 2020. Field experiments and the practice of policy. *Am. Econ. Rev.* 110(7):1952–73
Ferman B, Pinto C. 2021. Synthetic controls with imperfect pre-treatment fit. arXiv:1911.08521 [econ.EM]
Firpo S. 2007. Efficient semiparametric estimation of quantile treatment effects. *Econometrica* 75(1):259–76
Fisher R. 1937. *The Design of Experiments*. London: Oliver and Boyd
Freyaldenhoven S, Hansen C, Shapiro J. 2019. Pre-event trends in the panel event-study design. *Am. Econ. Rev.* 109(9):3307–38
Gelman A, Imbens G. 2018. Why high-order polynomials should not be used in regression discontinuity designs. *J. Bus. Econ. Stat.* 37(3):447–56
Goodman-Bacon A. 2021. Difference-in-differences with variation in treatment timing. *J. Econom.* 225(2):254–77
Gupta S, Kohavi R, Tang D, Xu Y, Andersen R, et al. 2019. Top challenges from the first Practical Online Controlled Experiments Summit. *ACM SIGKDD Explor. Newsl.* 21(1)

Haavelmo T. 1943. The statistical implications of a system of simultaneous equations. *Econometrica* 11(1):1–12
Hahn J. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2):315–31
Hahn J, Todd P, Van der Klaauw W. 2001. Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* 69(1):201–9
Han S. 2021. Identification in nonparametric models for dynamic treatment effects. *J. Econom.* 225(2):132–47
Harrison G, List J. 2004. Field experiments. *J. Econ. Lit.* 42(4):1009–55
Heckman J. 1990. Varieties of selection bias. *Am. Econ. Rev.* 80(2):313–18
Hirano K, Imbens G, Ridder G. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4):1161–89
Hirano K, Porter J. 2009. Asymptotics for statistical treatment rules. *Econometrica* 77(5):1683–701
Holland P. 1986. Statistics and causal inference. *J. Am. Stat. Assoc.* 81(396):945–60
Horvitz D, Thompson D. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47(260):663–85
Hotz V, Imbens G, Klerman J. 2006. Evaluating the differential effects of alternative welfare-to-work training components: a reanalysis of the California GAIN program. *J. Labor Econ.* 24(3):521–66
Hudgens M, Halloran M. 2008. Toward causal inference with interference. *J. Am. Stat. Assoc.* 103(482):832–42
Huntington-Klein N. 2021. *The Effect: An Introduction to Research Design and Causality*. Boca Raton, FL: CRC
Imai K, Kim I. 2019. When should we use unit fixed effects regression models for causal inference with longitudinal data? *Am. J. Political Sci.* 63(2):467–90
Imbens G. 2000. The role of the propensity score in estimating dose–response functions. *Biometrika* 87(3):706–10
Imbens G. 2003. Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev. Pap. Proc.* 93(2):126–32
Imbens G. 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev. Econ. Stat.* 86(1):4–29
Imbens G. 2014. Instrumental variables: an econometrician's perspective. *Stat. Sci.* (3):323–58
Imbens G. 2015. Matching methods in practice: three examples. *J. Hum. Resourc.* 50(2):373–419
Imbens G, Angrist J. 1994. Identification and estimation of local average treatment effects. *Econometrica* 61:467–76
Imbens G, Kalyanaraman K. 2012. Optimal bandwidth choice for the regression discontinuity estimator. *Rev. Econ. Stud.* 79(3):933–59
Imbens G, Rubin D. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge, UK: Cambridge Univ. Press
Imbens G, Wager S. 2018. Optimized regression discontinuity designs. *Rev. Econ. Stat.* 101(2):264–78
Imbens G, Wooldridge J. 2009. Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* 47(1):5–86
Johari R, Li H, Liskovich I, Weintraub G. 2022. Experimental design in two-sided platforms: an analysis of bias. *Manag. Sci.* 68(10):7069–89
Kalla J, Broockman D. 2018. The minimal persuasive effects of campaign contact in general elections: evidence from 49 field experiments. *Am. Political Sci. Rev.* 112(1):148–66
Keele L. 2015. The statistics of causal inference: a view from political methodology. *Political Anal.* 23(3):313–35
Kitagawa T, Tetenov A. 2018. Who should be treated? Empirical welfare maximization methods for treatment choice. *Econometrica* 86(2):591–616
Kremer M. 2020. Experimentation, innovation, and economics. *Am. Econ. Rev.* 110(7):1974–94
Lai T, Robbins H. 1985. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.* 6(1):4–22
LaLonde RJ. 1986. Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* 76(4):604–20
Lattimore T, Szepesvári C. 2020. *Bandit Algorithms*. Cambridge, UK: Cambridge Univ. Press
Lee D, Moretti E, Butler M. 2004. Do voters affect or elect policies? Evidence from the US House. *Q. J. Econ.* 119(3):807–59
Li F, Morgan K, Zaslavsky A. 2018. Balancing covariates via propensity score weighting. *J. Am. Stat. Assoc.* 113(521):390–400

Lin W. 2013. Agnostic notes on regression adjustments to experimental data: reexamining Freedman's critique. *Ann. Appl. Stat.* 7(1):295–318
Lin Z, Ding P, Han F. 2021. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. arXiv:2112.13506 [math.ST]
Liu L, Wang Y, Xu Y. 2022. A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *Am. J. Political Sci.* In press
Liu Y, Van Roy B, Xu K. 2023. A definition of non-stationary bandits. arXiv:2302.12202 [cs.LG]
Manski CF. 1990. Nonparametric bounds on treatment effects. *Am. Econ. Rev.* 80(2):319–23
Manski CF. 2003. *Partial Identification of Probability Distributions*. New York: Springer
Manski CF. 2004. Statistical treatment rules for heterogeneous populations. *Econometrica* 72(4):1221–46
Masten M, Poirier A, Zhang L. 2020. Assessing sensitivity to unconfoundedness: estimation and inference. arXiv:2012.15716 [econ.EM]
Matsudaira J. 2008. Mandatory summer school and student achievement. *J. Econom.* 142(2):829–50
Matzkin R. 1994. Restrictions of economic theory in nonparametric methods. In *Handbook of Econometrics*, Vol. 4, ed. RF Engle, DL McFadden, pp. 2523–58. Amsterdam: Elsevier
McCrary J. 2008. Testing for manipulation of the running variable in the regression discontinuity design. *J. Econom.* 142(2):698–714
Morgan S, Winship C. 2015. *Counterfactuals and Causal Inference*. Cambridge, UK: Cambridge Univ. Press
Newey W. 1990. Semiparametric efficiency bounds. *J. Appl. Econom.* 5(2):99–135
Oster E. 2019. Unobservable selection and coefficient stability: theory and evidence. *J. Bus. Econ. Stat.* 37(2):187–204
Pearl J. 1995. Causal diagrams for empirical research. *Biometrika* 82(4):669–88
Pearl J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge, UK: Cambridge Univ. Press
Pearl J, Mackenzie D. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books
Peters J, Janzing D, Schölkopf B. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: MIT Press
Porter J. 2003. *Estimation in the regression discontinuity model*. Work. Pap., Dep. Econ., Harvard Univ., Cambridge, MA. **https://users.ssc.wisc.edu/~jrporter/reg_discont_2003.pdf**
Pouget-Abadie J, Aydin K, Schudy W, Brodersen K, Mirrokni V. 2019. Variance reduction in bipartite experiments through correlation clustering. In *NIPS'19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*, ed. H Wallach, H Larochelle, A Beygelzimer, F d'Alché-Buc, E Fox, R Garnett, pp. 13309–19. Red Hook, NY: Curran
Prentice R. 1989. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat. Med.* 8(4):431–40
Robins J. 1989. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In *Health Service Research Methodology: A Focus on AIDS*, pp. 113–59. Washington, DC: US Public Health Serv.
Robins J. 1997. Causal inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality*, ed. M Berkane, pp. 69–117. New York: Springer
Robins J, Hernán MA, Brumback B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5):550–60
Robins J, Rotnitzky A, Zhao L. 1994. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* 89(427):846–66
Robins PK. 1985. A comparison of the labor supply findings from the four negative income tax experiments. *J. Hum. Resourc.* 20(4):567–82
Rosenbaum PR. 1984. The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. R. Stat. Soc. Ser. A* 147(5):656–66
Rosenbaum PR. 2002. *Observational Studies*. New York: Springer
Rosenbaum PR. 2010. *Design of Observational Studies*. New York: Springer
Rosenbaum PR. 2020. Modern algorithms for matching in observational studies. *Annu. Rev. Stat. Appl.* 7:143–76
Rosenbaum PR, Rubin DB. 1983a. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Stat. Soc. Ser. B* 45(2):212–18

Rosenbaum PR, Rubin DB. 1983b. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55
Roth J, Sant'Anna PHC, Bilinski A, Poe J. 2022. What's trending in difference-in-differences? A synthesis of the recent econometrics literature. arXiv:2201.01194 [econ.EM]
Rubin DB. 1977. Assignment to treatment group on the basis of a covariate. *J. Educ. Stat.* 2(1):1–26
Rubin DB. 1978. Bayesian inference for causal effects: the role of randomization. *Ann. Stat.* 6(1):34–58
Rubin DB. 2006. *Matched Sampling for Causal Effects*. Cambridge, UK: Cambridge Univ. Press
Rubin DB. 2008. For objective causal inference, design trumps analysis. *Ann. Appl. Stat.* 2(3):808–40
Russo D, Van Roy B, Kazerouni A, Osband I, Wen Z. 2018. *A Tutorial on Thompson Sampling*. Boston: Now
Sant'Anna P, Zhao J. 2020. Doubly robust difference-in-differences estimators. *J. Econom.* 219(1):101–22
Scott S. 2010. A modern Bayesian look at the multi-armed bandit. *Appl. Stoch. Models Bus. Industry* 26(6):639–58
Splawa-Neyman J. 1990 (1923). On the application of probability theory to agricultural experiments. Essay on principles. Section 9, transl. D Dabrowska, T Speed. *Stat. Sci.* 5(4):465–72
Staiger D, Stock J. 1997. Instrumental variables regression with weak instruments. *Econometrica* 65(3):557–86
Stuart E. 2010. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 25(1):1–21
Sun L, Abraham S. 2021. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *J. Econom.* 225(2):175–99
Thistlewaite D, Campbell D. 1960. Regression-discontinuity analysis: an alternative to the ex-post facto experiment. *J. Educ. Psychol.* 51(2):309–17
Thompson W. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3–4):285–94
Tinbergen J. 1930. Determination and interpretation of supply curves: an example. *Z. Nationalokonomie* 1(5):669–79
Van Der Klaauw W. 2002. Estimating the effect of financial aid offers on college enrollment: a regression-discontinuity approach. *Int. Econ. Rev.* 43(2):1249–87
Van der Laan M, Rose S. 2011. *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer
VanderWeele TJ. 2015. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford, UK: Oxford University Press
Wager S, Athey S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* 113.523(2018):1228–42
Wright S. 1934. The method of path coefficients. *Ann. Math. Stat.* 5(3):161–215
Xu Y. 2017. Generalized synthetic control method: causal inference with interactive fixed effects models. *Political Anal.* 25(1):57–76
Zigler C, Papadogeorgou G. 2021. Bipartite causal inference with interference. *Stat. Sci.* 36(1):109–23
Zubizarreta J, Stuart E, Small D, Rosenbaum P. 2023. *Handbook of Matching and Weighting Adjustments for Causal Inference*. Boca Raton, FL: CRC