

# Econometric Methods for Program Evaluation<sup>\*</sup>

Alberto Abadie<sup>†</sup>      Matias D. Cattaneo<sup>‡</sup>

March 19, 2018

## Abstract

Program evaluation methods are widely applied in economics to assess the effects of policy interventions and other treatments of interest. In this article, we describe the main methodological frameworks of the econometrics of program evaluation. In the process, we delineate some of the directions along which this literature is expanding, discuss recent developments and highlight specific areas where new research may be particularly fruitful.

**Keywords:** causal inference, policy evaluation, treatment effects.

---

<sup>\*</sup>We thank Josh Angrist, Max Farrell, Guido Imbens, Michael Jansson, J  r  my L’Hour, Judith Lok, Xinwei Ma, and Rocio Titiunik for thoughtful comments. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1459931.

<sup>†</sup>Department of Economics, Massachusetts Institute of Technology. Cambridge, Massachusetts 02142; email: [abadie@mit.edu](mailto:abadie@mit.edu)

<sup>‡</sup>Department of Economics and Department of Statistics, University of Michigan. Ann Arbor, Michigan 48109; email: [cattaneo@umich.edu](mailto:cattaneo@umich.edu)

## 1. Introduction

The nature of empirical research in economics has profoundly changed since the emergence in the 1990's of a new way to understand identification in econometric models. This approach emphasizes the importance of heterogeneity across units in the parameters of interest, and of the choice of the sources of variation in the data that are used to estimate those parameters. Initial impetus came from empirical research by Angrist, Ashenfelter, Card, Krueger and others, while Angrist, Heckman, Imbens, Manski, Rubin, and others provided many of the early methodological innovations.<sup>1</sup>

While this greater emphasis on identification has permeated all brands of econometrics and quantitative social science, nowhere have the effects been felt more profoundly than in the field of program evaluation—a domain expanding the social, biomedical, and behavioral sciences that studies the effects of policy interventions. The policies of interest are often governmental programs, like active labor market interventions or anti-poverty programs. In other instances, the term “policy” is more generally understood to include any intervention of interest by public or private agents, or by nature.

In this article, we offer an overview of classical tools and recent developments in the econometrics of program evaluation and discuss potential avenues for future research. Our focus is on ex-post evaluation exercises, where the effects of policy interventions are evaluated after some form of this intervention (such as a regular implementation of the policy, an experimental study, or a pilot study) is deployed and the relevant outcomes under the intervention are measured. Many studies in economics and the social sciences aim to evaluate existing or experimentally deployed interventions, and the methods surveyed in this article have been shown to have broad applicability in that setting. In many other cases, however, economists and social scientists are interested in ex-ante evaluation exercises, that is, in estimating the effect of a policy before such policy is implemented on the population of interest (e.g., Todd and Wolpin, 2010). In the absence of measures of the outcomes under

---

<sup>1</sup>See, in particular, Angrist (1990); Angrist and Krueger (1991); Angrist et al. (1996); Card and Krueger (1994); Card (1990); Heckman (1990); Manski (1990) and earlier contributions in Ashenfelter (1978); Ashenfelter and Card (1985); Heckman and Robb (1985).

the intervention of interest, ex-ante evaluations aim to extrapolate outside the support of the data. Often, economic models that precisely describe the behavior of economic agents serve as useful extrapolation devices for ex-ante evaluations. Methods that are suitable for ex-ante evaluation have been previously covered in this *Review* (Arcidiacono and Ellickson, 2011; Berry and Haile, 2016).

This distinction between ex-ante and ex-post program evaluations is closely related to often-discussed differences between the so-called “structural” and “causal inference” schools in econometrics. The use of economic models as devices to extrapolate outside the support of the data in ex-ante program evaluation is typically associated with the structural school, while ex-post evaluations that do not explicitly extrapolate using economic models of behavior are often termed causal or reduced form.<sup>2</sup> Given that they serve related but quite distinct purposes, we find the perceived conflict between these two schools rather artificial, and the debates about the purported superiority of one of the approaches over the other largely unproductive.

Our goal in this article is to provide a summary overview of the literature on the econometrics of program evaluation for ex-post analysis and, in the process, to delineate some of the directions along which it is expanding, discussing recent developments and the areas where research may be particularly fruitful in the near future.

Other recent surveys on the estimation of causal treatment effects and the econometrics of program evaluation from different perspectives and disciplines include Abadie (2005a), Angrist and Pischke (2008, 2014), Athey and Imbens (2017c), Blundell and Costa Dias (2009), DiNardo and Lee (2011), Heckman and Vytlacil (2007), Hernán and Robins (2018), Imbens and Rubin (2015), Imbens and Wooldridge (2009), Lee (2016), Manski (2008), Pearl (2009), Rosenbaum (2002, 2010), Van der Laan and Robins (2003), and VanderWeele (2015), among many others.

---

<sup>2</sup>As argued in Imbens (2010), however, “reduced form” is a misnomer relative to the literature on simultaneous equation models, where the terms “structural form” and “reduced form” have precise meanings. Moreover, terms like “structural model” or “causal inference” are nomenclature, and not exclusive attributes of the literatures and methods they refer to.

## 2. Causal Inference and Program Evaluation

Program evaluation is concerned with the estimation of the causal effects of policy interventions. These policy interventions can be of very different nature, depending on the context of the investigation, and they are often generically referred to as treatments. Examples include conditional transfer programs (Behrman et al., 2011), health care interventions (Newhouse, 1996; Finkelstein et al., 2012), and large-scale online A/B studies in which IP addresses visiting a particular web page are randomly assigned to different designs or contents (see, e.g., Bakshy et al., 2014).

### 2.1. Causality and Potential Outcomes

We represent the value of the treatment by the random variable  $W$ . We aim to learn the effect of changes in treatment status on some observed outcome variable, denoted by  $Y$ . Following Neyman (1923), Rubin (1974) and many others, we use potential outcomes to define causal parameters:  $Y_w$  represents the potential value of the outcome when the value of the treatment variable,  $W$ , is set to  $w$ . For each value of  $w$  in the support of  $W$ , the potential outcome  $Y_w$  is a random variable with a distribution over the population. The realized outcome,  $Y$ , is such that, if the value of the treatment is equal to  $w$  for a unit in the population, then for that unit  $Y = Y_w$ , while other potential outcomes  $Y_{w'}$  with  $w' \neq w$  remain counterfactual.

A strong assumption lurks implicit in the last statement. Namely, the realized outcome for each particular unit depends only on the value of the treatment of that unit, and not on the treatment or on outcome values of other units. This assumption is often referred to as “Stable Unit Treatment Value Assumption” or SUTVA, and rules out interference between units (Rubin, 1980). SUTVA is a strong assumption in many practical settings; for example, it may be violated in an educational setting with peer effects. However, concerns about interference between units can often be mitigated through careful study design (see, e.g., Imbens and Rubin, 2015).

The concepts of potential and realized outcomes are deeply ingrained in economics. A

demand function, for example, represents the potential quantity demanded as a function of price. Quantity demanded is realized for the market price and is counterfactual for other prices.

While, in practice, researchers may be interested in a multiplicity of treatments and outcomes, we will abstract from that in our notation, where  $Y$  and  $W$  are scalar random variables. In addition to treatments and potential outcomes, the population is characterized by covariates,  $X$ , a  $(k \times 1)$  vector of variables that are predetermined relative to the treatment. That is, while  $X$  and  $W$  may not be independent (perhaps because  $X$  causes  $W$ , or perhaps because they share common causes) the value of  $X$  cannot be changed by active manipulation of  $W$ . Oftentimes,  $X$  contains characteristics of the units measured before  $W$  is known.

Although the notation allows the treatment to take on an arbitrary number of values, we will introduce additional concepts and notation within the context of a binary treatment, that is,  $W \in \{0, 1\}$ . In this case,  $W = 1$  often denotes exposure to an active intervention (e.g., participation in an anti-poverty program), while  $W = 0$  denotes remaining at the status quo. For simplicity of exposition, our discussion will mostly focus on the binary treatment case. The causal effect of the treatment (or treatment effect) can be represented by the difference in potential outcomes,  $Y_1 - Y_0$ . Potential outcomes and the value of the treatment determine the observed outcome

$$Y = WY_1 + (1 - W)Y_0. \tag{1}$$

Equation (1) represents what is often termed “the fundamental problem of causal inference” (Holland, 1986). The realized outcome,  $Y$ , reveals  $Y_1$  if  $W = 1$  and  $Y_0$  if  $W = 0$ . However, the unit-level treatment effect,  $Y_1 - Y_0$ , depends on both quantities. As a result, the value of  $Y_1 - Y_0$  is unidentified from observing  $(Y, W, X)$ .

Beyond the individual treatment effects,  $Y_1 - Y_0$ , which are identified only under assumptions that are not plausible in most empirical settings, the objects of interest, or estimands, in program evaluation are characteristics of the joint distribution of  $(Y_1, Y_0, W, X)$  in the sample or in the population. For most of this survey, we will focus on estimands defined for a certain population of interest from which we have extracted a random sample. Like

in Abadie et al. (2017a), we will say that an estimand is causal when it depends on the distribution of the potential outcomes,  $(Y_1, Y_0)$ , beyond its dependence on the distribution of  $(Y, W, X)$ . This is in contrast to descriptive estimands, which are characteristics of the distribution of  $(Y, W, X)$ .<sup>3</sup>

The average treatment effect (ATE),

$$\tau_{\text{ATE}} = E[Y_1 - Y_0],$$

and the average treatment effect on the treated (ATET),

$$\tau_{\text{ATET}} = E[Y_1 - Y_0 | W = 1],$$

are causal estimands often of interest in the program evaluation literature. Under SUTVA, ATE represents the difference in average outcomes induced by shifting the entire population from the inactive to the active treatment. ATET represents the same object for the treated. To improve the targeting of a program, researchers often aim to estimate ATE or ATET after accounting for a set of units' characteristics,  $X$ . Notice that ATE and ATET depend on the distribution of  $(Y_1, Y_0)$  beyond their dependence on the distribution of  $(Y, W, X)$ . Therefore, they are causal estimands under the definition in Abadie et al. (2017a).

## 2.2. Confounding

Consider now the descriptive estimand,

$$\tau = E[Y | W = 1] - E[Y | W = 0],$$

which is the difference in average outcomes for the two different treatment values. ATE and ATET are measures of causation, while the difference in means,  $\tau$ , is a measure of association. The differences between them are

$$\begin{aligned} \tau &= \tau_{\text{ATE}} + b_{\text{ATE}} \\ &= \tau_{\text{ATET}} + b_{\text{ATET}}, \end{aligned} \tag{2}$$

---

<sup>3</sup>See Abadie et al. (2017a) for a discussion of descriptive and causal estimands in the context of regression analysis.

where  $b_{\text{ATE}}$  and  $b_{\text{ATET}}$  are bias terms given by

$$\begin{aligned} b_{\text{ATE}} = & \left( E[Y_1|W = 1] - E[Y_1|W = 0] \right) \Pr(W = 0) \\ & + \left( E[Y_0|W = 1] - E[Y_0|W = 0] \right) \Pr(W = 1), \end{aligned} \quad (3)$$

and

$$b_{\text{ATET}} = E[Y_0|W = 1] - E[Y_0|W = 0]. \quad (4)$$

Equations (2) to (4) combine into a mathematical representation of the often-invoked statement “association does not imply causation,” enriched with “lack of association does not imply lack of causation”. If average potential outcomes are identical for treated and non-treated units, an untestable condition, then the bias terms  $b_{\text{ATE}}$  and  $b_{\text{ATET}}$  disappear. However, if potential outcomes are not independent of the treatment, then in general the difference in mean outcomes between the treated and the untreated is not equal to ATE or ATET. Lack of independence between the treatment and the potential outcomes is often referred to as confounding. Confounding may arise, for example, when information that is correlated with potential outcomes is used for treatment assignment, or when agents actively self-select into treatment based on their potential gains.

Confounding is a powerful notion, as it explains departures of descriptive estimands from causal ones. Directed Acyclic Graphs (or DAGs, see Pearl, 2009) provide graphical representations of causal relationships and confounding.

A DAG is a collection of nodes and directed edges among nodes, with no directed cycles. Nodes represent random variables, while directed edges represent causal effects not mediated by other variables in the DAG. Moreover, a causal DAG must contain all causal effects among the variables in the DAG, as well as all variables that are common causes of any pair of variables in the DAG (even if common causes are unobserved).

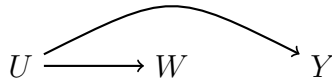


Figure 1: Confounding in a DAG

Consider the DAG in Figure 1. This DAG includes a treatment,  $W$ , and an outcome of interest,  $Y$ . In this DAG,  $W$  is not a cause of  $Y$ , as there is no directed edge from  $W$  to  $Y$ . The DAG also includes a confounder,  $U$ . This is a common cause of  $W$  and  $Y$ . Even when  $W$  does not cause  $Y$ , these two variables may be correlated because of the confounding effect of  $U$ . A familiar example of this structure in economics is Spence’s job-market signalling model (Spence, 1973), where  $W$  is worker education,  $Y$  is worker productivity, and there are other worker characteristics,  $U$ , that cause education and productivity (e.g., worker attributes that increase productivity and reduce the cost of acquiring education). In this example, even when education does not cause worker productivity, the two variables may be positively correlated.

The DAG in Figure 1 implies that  $Y_w$  does not vary with  $w$ , because  $W$  does not cause  $Y$ . However,  $Y$  and  $W$  are not independent, as they share a common cause,  $U$ . Confounding may arise in more complicated scenarios. In Section 4 we will cover other forms of confounding that may arise in DAGs. Confounding, of course, often coexists with true causal effects. That is, the DAG in Figure 1 implies the presence of confounding whether or not there is a directed edge from  $W$  to  $Y$ .

Beyond average treatment effects, other treatment parameters of interest depend on the distribution of  $(Y_1, Y_0)$ . These include treatment effects on characteristics of the distribution of the outcome other than the average (e.g., quantiles) as well as characteristic of the distribution of the treatment effect,  $Y_1 - Y_0$  (e.g., variance) other than its average. In the next section, we elaborate further on the distinction between treatment effects on the distribution of the outcome vs. characteristics of the distribution of the treatment effect.

### 3. Randomized Experiments

Active randomization of the treatment provides the basis of what arguably is the most successful and extended scientific research design for program evaluation. Often referred to as the “gold standard” of scientific evidence, randomized clinical trials play a central role in the natural sciences, as well as in the drug approval process in the U.S., Europe, and



elsewhere (Bothwell et al., 2016). While randomized assignment is often viewed as the basis of a high standard of scientific evidence, this view is by no means universal. Cartwright (2007), Heckman and Vytlacil (2007), and Deaton (2010), among others, challenge the notion that randomized experiments occupy a preeminent position in the hierarchy of scientific evidence, while Angrist and Pischke (2010), Imbens (2010) and others emphasize the advantages of experimental designs.

### 3.1. Identification through Randomized Assignment

Randomization assigns treatment using the outcome of a procedure (natural, mechanical, or electronic) that is unrelated to the characteristics of the units and, in particular, unrelated to potential outcomes. Whether randomization is based on coin tossing, random number generation in a computer, or on the radioactive decay of materials, the only requirement of randomization is that the generated treatment variable is statistically independent of potential outcomes. Complete (or unconditional) randomization implies

$$(Y_1, Y_0) \perp\!\!\!\perp W, \quad (5)$$

where the symbol  $\perp\!\!\!\perp$  denotes statistical independence. If Equation (5) holds, we will say that the treatment assignment is unconfounded. Later in this article, we will consider conditional versions of unconfoundedness, where the value of the treatment  $W$  for each particular unit may depend on certain characteristics of the units, and Equation (5) holds conditional on those characteristics.

The immediate consequence of randomization is that the bias terms in Equations (3) and (4) are equal to zero:

$$b_{ATE} = b_{ATET} = 0,$$

which, in turn, implies

$$\tau_{ATE} = \tau_{ATET} = \tau = E[Y|W = 1] - E[Y|W = 0]. \quad (6)$$

Beyond average treatment effects, randomization identifies any characteristic of the marginal distributions of  $Y_1$  and  $Y_0$ . In particular, the cumulative distributions of potential outcomes

$F_{Y_1}(y) = \Pr(Y_1 \leq y)$  and  $F_{Y_0}(y) = \Pr(Y_0 \leq y)$  are identified by

$$F_{Y_1}(y) = \Pr(Y \leq y | W = 1)$$

and

$$F_{Y_0}(y) = \Pr(Y \leq y | W = 0).$$

While the average effect of a treatment can be easily described using  $\tau_{\text{ATE}}$ , it is more complicated to establish a comparison between the entire distributions of  $Y_1$  and  $Y_0$ . To be concrete, suppose that the outcome variable of interest,  $Y$ , is income. Then, an analyst may want to rank the income distributions under the active intervention and in the absence of the active intervention. For that purpose, it is useful to consider the following distributional relationships:

(i) Equality of distributions

$$F_{Y_1}(y) = F_{Y_0}(y) \quad \text{for all } y \geq 0.$$

(ii) First-order stochastic dominance (with  $F_{Y_1}$  dominating  $F_{Y_0}$ )

$$F_{Y_1}(y) - F_{Y_0}(y) \leq 0 \quad \text{for all } y \geq 0.$$

(iii) Second-order stochastic dominance (with  $F_{Y_1}$  dominating  $F_{Y_0}$ )

$$\int_0^y (F_{Y_1}(z) - F_{Y_0}(z)) dz \leq 0 \quad \text{for all } y \geq 0.$$

Equality of distributions implies that the treatment has no effect on the distribution of the outcome variable, in this case income. It implies  $\tau_{\text{ATE}} = 0$ , but represents a stronger notion of null effect. Under mild assumptions, first- and second-order stochastic dominance imply that the income distribution under the active treatment is preferred to the distribution of income in the absence of the active treatment (see, e.g., Foster and Shorrocks, 1988; Abadie, 2002).

Equivalently, the distributions of  $Y_1$  and  $Y_0$  can be described by their quantiles. Quantiles of  $Y_1$  and  $Y_0$  are identified by inverting  $F_{Y_1}$  and  $F_{Y_0}$  or, more directly, by

$$Q_{Y_1}(\theta) = Q_{Y|W=1}(\theta)$$

and

$$Q_{Y_0}(\theta) = Q_{Y|W=0}(\theta),$$

where  $\theta \in (0, 1)$ , and  $Q_Y(\theta)$  denotes the  $\theta$ -quantile of the distribution of  $Y$ . The effect of the treatment on the  $\theta$ -quantile of the outcome distribution is given by the quantile treatment effect,

$$Q_{Y_1}(\theta) - Q_{Y_0}(\theta). \tag{7}$$

Notice that, while quantile treatment effects are identified from Equation (5), quantiles of the distribution of the treatment effect,  $Q_{Y_1 - Y_0}(\theta)$ , are not identified. Unlike expectations, quantiles are non-linear operators, so in general  $Q_{Y_1 - Y_0}(\theta) \neq Q_{Y_1}(\theta) - Q_{Y_0}(\theta)$ . Moreover, while quantile treatment effects depend only on the marginal distributions of  $Y_1$  and  $Y_0$ , which are identified from Equation (5), quantiles of the treatment effect distribution are functionals of the joint distribution of  $(Y_1, Y_0)$ , which involve information beyond the marginals. Because, even in a randomized experiment, the two potential outcomes are never both realized for the same unit, the joint distribution of  $(Y_1, Y_0)$  is not identified. As a result, quantiles of  $Y_1 - Y_0$  are not identified even in the context of a randomized experiment.

To gain intuitive understanding of this lack of identification result, consider an example where the marginal distributions of  $Y_1$  and  $Y_0$  are identical, symmetric around zero, and non-degenerate. Identical marginals are consistent with a null treatment effects, that is,  $Y_1 - Y_0 = 0$  with probability one. In this scenario, all treatment effects are zero. However, identical marginals and symmetry around zero are also consistent with  $Y_1 = -Y_0$  or, equivalently,  $Y_1 - Y_0 = -2Y_0$  with probability one, which leads to positive treatment effects for half of the population and negative for the other half. In the first scenario, the treatment does not change the value of the outcome for any unit in the population. In contrast, in the second scenario, the locations of units within the distribution of the outcome variable are reshuffled,

but the shape of the distribution does not change. While these two different scenarios imply different distributions of the treatment effect,  $Y_1 - Y_0$ , they are both consistent with the same marginal distributions of the potential outcomes,  $Y_1$  and  $Y_0$ , so the distribution of  $Y_1 - Y_0$  is not identified.

Although the distribution of treatment effects is not identified from randomized assignment of the treatment, knowledge of the marginal distributions of  $Y_1$  and  $Y_0$  can be used to bound functionals of the distribution of  $Y_1 - Y_0$  (see, e.g., Firpo and Ridder, 2008; Fan and Park, 2010). In particular, if the distribution of  $Y$  is continuous, then for each  $t \in \mathbb{R}$ ,

$$\max \left\{ \sup_{y \in \mathbb{R}} \left( F_{Y_1}(y) - F_{Y_0}(y-t) \right), 0 \right\} \leq F_{Y_1 - Y_0}(t) \leq 1 + \min \left\{ \inf_{y \in \mathbb{R}} \left( F_{Y_1}(y) - F_{Y_0}(y-t) \right), 0 \right\}, \quad (8)$$

provides sharp bounds on the distribution of  $Y_1 - Y_0$ . This expression can be applied to bound the fraction of units with positive (or negative) treatment effects, or to bound quantile treatment effects in Equation (7). Bounds on  $Q_{Y_1 - Y_0}(\theta)$  can be obtained by inversion.

### 3.2. Estimation and Inference in Randomized Studies

So far, we have concentrated on identification issues. We now turn to estimation and inference. Multiple modes of statistical inference are available for treatment effects (see Rubin, 1990), and our discussion will mainly focus on two of them: (i) sampling-based frequentist inference, and (ii) randomization-based inference on a sharp null. We will also briefly mention permutation-based inference. However, because of space limitations, we omit Bayesian methods (see Rubin, 1978, 2005) and finite-sample frequentist methods based on randomization (see Neyman, 1923; Abadie et al., 2017a). Imbens and Rubin (2015) give an overview of many of these methods.

### 3.3. Sampling-Based Frequentist Inference

In the context of sampling-based frequentist inference, we assume that data consist of a random sample of treated and nontreated units, and that the treatment has been assigned randomly. For each unit  $i$  in the sample, we observe the value of the outcome,  $Y_i$ , the value of the treatment,  $W_i$ , and possibly the values of a set of covariates,  $X_i$ . Consider

the common scenario where randomization is carried out with a fixed number of treated units,  $n_1$ , and untreated units,  $n_0$ , such that  $n = n_1 + n_0$  is the total sample size. We will construct estimators using the analogy principle (Manski, 1988). That is, estimators are sample counterparts of population objects that identify parameters of interest. In particular, Equation (6) motivates estimating  $\tau = \tau_{\text{ATE}} = \tau_{\text{ATE}^*}$  using the difference in sample means of the outcome between treated and nontreated,

$$\hat{\tau} = \frac{1}{n_1} \sum_{i=1}^n W_i Y_i - \frac{1}{n_0} \sum_{i=1}^n (1 - W_i) Y_i.$$

Under conventional regularity conditions, a large sample approximation to the sampling distribution of  $\hat{\tau}$  is given by,

$$\frac{\hat{\tau} - \tau}{\widehat{\text{se}}(\hat{\tau})} \xrightarrow{d} \mathcal{N}(0, 1), \quad (9)$$

where

$$\widehat{\text{se}}(\hat{\tau}) = \left( \frac{\hat{\sigma}_{Y|W=1}^2}{n_1} + \frac{\hat{\sigma}_{Y|W=0}^2}{n_0} \right)^{1/2}, \quad (10)$$

and  $\hat{\sigma}_{Y|W=1}^2$  and  $\hat{\sigma}_{Y|W=0}^2$  are the sample counterparts of the conditional variance of  $Y$  given  $W = 1$  and  $W = 0$ , respectively.

The estimator  $\hat{\tau}$  coincides with the coefficient on  $W$  from a regression of  $Y$  on  $W$  and a constant, and  $\widehat{\text{se}}(\hat{\tau})$  is the corresponding heteroskedasticity-robust standard error. Inference based on (9) and (10) has an asymptotic justification but can perform poorly in finite samples, especially if  $n_1$  and  $n_0$  differ considerably. Imbens and Kolesár (2016) discuss small sample adjustments to the  $t$ -statistic and its distribution and demonstrate that such adjustments substantially improve coverage rates of confidence intervals in finite samples. In addition, more general sampling processes (e.g., clustered sampling) and/or randomization schemes (e.g., stratified randomization) are possible. They affect the variability of the estimator and, therefore, the standard error formulas (see Abadie et al., 2017b; Imbens and Rubin, 2015, and references therein for details).

Estimators of other quantities of interest discussed in Section 2 can be similarly constructed using sample analogs. In particular, for any  $\theta \in (0, 1)$ , the quantile treatment

effects estimator is given by

$$\widehat{Q}_{Y|W=1}(\theta) - \widehat{Q}_{Y|W=0}(\theta),$$

where  $\widehat{Q}_{Y|W=1}(\theta)$  and  $\widehat{Q}_{Y|W=0}(\theta)$  are the sample analogs of  $Q_{Y|W=1}(\theta)$  and  $Q_{Y|W=0}(\theta)$ . Sampling-based frequentist inference on quantile treatment effects follows from the results of Koenker and Bassett (1978) on quantile inference for quantile regression estimators. Bounds on the distribution of  $F_{Y_1-Y_0}$  can be computed by replacing the cumulative functions  $F_{Y|W=1}$  and  $F_{Y|W=0}$  in Equation (8) with their sample analogs,  $\widehat{F}_{Y|W=1}$  and  $\widehat{F}_{Y|W=0}$ . The estimators of the cumulative distribution functions,  $F_{Y_1}$  and  $F_{Y_0}$  (or, equivalently, estimators of the quantile functions  $Q_{Y_1}$  and  $Q_{Y_0}$ ) can also be applied to test for equality of distributions, and first- and second-order stochastic dominance (see, e.g., Abadie, 2002).

### 3.4. Randomization Inference on a Sharp Null

The testing exercise in the previous section is based on sampling inference. That is, it is based on the comparison of the value of a test statistic for the sample at hand to its sampling distribution under the null hypothesis. In contrast, randomization inference takes the sample as fixed, and concentrates on the null distribution of the test statistic induced by the randomization of the treatment. Randomization inference originated with Fisher’s proposal to use the “physical act of randomization” as a “reasoned basis for inference” (Fisher, 1935).

Consider the example in Panel A of Table 1. This panel shows the result of a hypothetical experiment where a researcher randomly selects four individuals out of a sample of eight individuals to receive the treatment, and excludes the remaining four individuals from the treatment. For each of the eight sample individuals,  $i = 1, \dots, 8$ , the researcher observes treatment status,  $W_i$ , and the value of the outcome variable,  $Y_i$ . For concreteness, we will assume that the researcher adopts the difference in mean outcomes between the treated and the non-treated as a test statistic for randomization inference. The value of this statistic in the experiment is  $\widehat{\tau} = 6$ .

The sample observed by the researcher is informative only about the value of  $\widehat{\tau}$  obtained

Panel A: Sample and sample statistic

$Y_i$	12	4	6	10	6	0	1	1	
$W_i$	1	1	1	1	0	0	0	0	$\hat{\tau} = 6$

Panel B: Randomization distribution

									$\hat{\tau}(\omega)$
$\omega = 1$	1	1	1	1	0	0	0	0	6
$\omega = 2$	1	1	1	0	1	0	0	0	4
$\omega = 3$	1	1	1	0	0	1	0	0	1
$\omega = 4$	1	1	1	0	0	0	1	0	1.5
				...					
$\omega = 70$	0	0	0	0	1	1	1	1	-6

Table 1: Randomization Distribution of a Difference in Means

under one particular realization of the randomized treatment assignment: the one observed in practice. There are, however, 70 possible ways to assign four individuals out of eight to the treatment group. Let  $\Omega$  be the set of all possible randomization realizations. Each element  $\omega$  of  $\Omega$  has probability  $1/70$ . Consider “Fisher’s null hypothesis” that the treatment does not have any effect on the outcomes of any unit, that is,  $Y_{1i} = Y_{0i}$  for each experimental unit. Notice that Fisher’s null pertains to the sample and not necessarily to the population. Under Fisher’s null it is possible to calculate the value  $\hat{\tau}(\omega)$  that would have been observed for each possible realization of the randomized assignment,  $\omega \in \Omega$ . The distribution of  $\hat{\tau}(\omega)$  in Panel B of Table 1 is called the randomization distribution of  $\hat{\tau}$ . A histogram of the randomization distribution of  $\hat{\tau}$  is depicted in Figure 1, along with a vertical line representing the sample value of  $\hat{\tau}$ .  $p$ -values for one-sided and two-sided tests are calculated in the usual fashion:  $\Pr(\hat{\tau}(\omega) \geq \hat{\tau}) = 0.0429$  and  $\Pr(|\hat{\tau}(\omega)| \geq |\hat{\tau}|) = 0.0857$ , respectively. These probabilities are calculated over the randomization distribution. Because the randomization distribution under the null can be computed without error, these  $p$ -values are exact.<sup>4</sup>

<sup>4</sup>In a setting with many experimental units, it may be computationally costly to calculate the value of the test statistic under all possible assignments. In those cases, the randomization distribution can be approximated from a randomly selected subset of assignments.

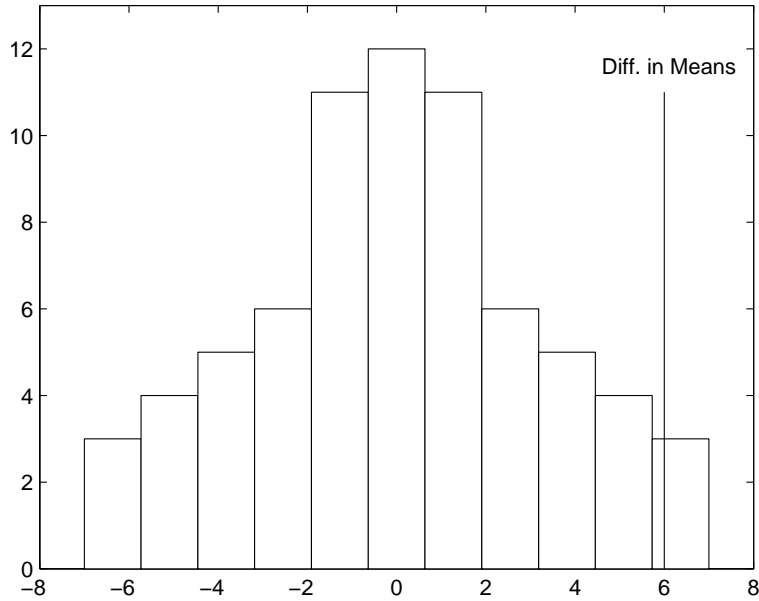


Figure 2: Randomization Distribution of the Difference in Means

Fisher’s null hypothesis “ $Y_{1i} - Y_{0i} = 0$  for every experimental unit” is an instance of a “sharp null”, that is, a null hypothesis under which the values of the treatment effect for each unit in the experimental sample are fixed. Notice that this sharp null, if extended to “every population unit”, implies but is not implied by “Neyman’s null”,  $E[Y_1] - E[Y_0] = 0$ . However, a rejection of Neyman’s null, using the t-statistic discussed above in the context of sampling-based frequentist inference, does not imply that Fisher’s test will reject. This is explained in Ding (2017) using the power properties of the two types of tests, and illustrated in Young (2017).

### 3.5. Permutation Methods

Some permutation methods are related to randomization tests but rely on a sampling interpretation. They can be employed to evaluate the null hypothesis that  $Y_1$  and  $Y_0$  have the same distribution. Notice that this is not a sharp hypothesis, and that it pertains to the population rather than to the sample at hand. These tests are based on the observation that, under the null hypothesis, the distribution of the vector of observed outcomes is



invariant under permutations. For more discussion and comparisons between Fisher’s randomization inference and permutation methods see Ernst (2004), and for a recent application of permutation-based inference in the context of randomized experiments see Bugni et al. (2018), who also include further references on this topic.

### 3.6. Recent Developments

Looking ahead, the literature on the analysis and interpretation of randomized experiments remains active with several interesting developments as well as challenges still present beyond what was discussed in this section. Athey and Imbens (2017a) provide a detailed introduction to the econometrics of randomized experiments. Some recent topics in this literature include (i) decision theoretic approaches to randomized experiments and related settings (Banerjee et al., 2017), (ii) design of complex/high-dimensional experiments accounting for new (big) data environments such as social media or networks settings (Bakshy et al., 2014), (iii) the role of multiple hypothesis testing and alternative inference methods, such as model selection, shrinkage and empirical Bayes approaches (see, e.g., Abadie and Kasy, 2017), and (iv) subgroup, dynamic and optimal treatment effect analysis as well as related issues of endogenous stratification (see, e.g., Athey and Imbens, 2017b; Murphy, 2003; Abadie et al., 2017c).

## 4. Conditioning on Observables

### 4.1. Identification by Conditional Independence

In the absence of randomization, the independence condition in Equation (5) is rarely justified, and the estimation of treatment effects parameters requires additional identifying assumptions. In this section, we discuss methods based on a conditional version of the unconfoundedness assumption in Equation (5). We will assume that potential outcomes are independent of the treatment after conditioning on a set of observed covariates. The key underlying idea is that confounding, if present, is fully accounted for by observed covariates.

Consider an example where newly-admitted college students are awarded a scholarship

on the basis of the result of a college entry exam,  $X$ , and other student characteristics,  $V$ . We will use the binary variable  $W$  to code the receipt of a scholarship award. To investigate the effect of the award on college grades,  $Y$ , one may be concerned about the potential confounding effect of pre-college academic ability,  $U$ , which may not be precisely measured by  $X$ . Confounding may also arise if  $V$  has a direct effect on  $Y$ .



Figure 3: Causal Effect and Confounding

This setting is represented in the DAG in Figure 3(a). Pre-college academic ability,  $U$ , is a cause of  $W$ , the scholarship award, through its effect on the college entry exam grade,  $X$ . Pre-college ability,  $U$ , is also a direct cause of college academic grades,  $Y$ . The confounding effect of  $U$  induces statistical dependence between  $W$  and  $Y$  that is not reflective of the causal effect of  $W$  on  $Y$ . In a DAG, a path is a collection of consecutive edges (e.g.,  $X \rightarrow W \rightarrow Y$ ), and confounding generates “backdoor paths”, that is, paths from the treatment,  $W$ , to the outcome,  $Y$ , that start with incoming arrows. The path  $W \leftarrow X \leftarrow U \rightarrow Y$  is a backdoor path in Figure 3(a). An additional backdoor path,  $W \leftarrow V \rightarrow Y$ , emerges if other causes of the award, aside from the college entry exam, also have a direct causal effect on college grades. In general (that is, except for special cases), the presence of confounding invalidates Equation (5).

Consider a conditional version of unconfoundedness,

$$(Y_1, Y_0) \perp\!\!\!\perp W | X. \quad (11)$$

Equation (11) states that  $(Y_1, Y_0)$  is independent of  $W$  given  $X$ .<sup>5</sup> This assumption allows identification of treatment effect parameters by conditioning on  $X$ . That is, controlling for

---

<sup>5</sup>For the identification results stated below, it is enough to have unconfoundedness in terms of the marginal distributions of potential outcomes, that is,  $Y_w \perp\!\!\!\perp W | X$  for  $w = \{0, 1\}$ .

$X$  makes the treatment unconfounded.<sup>6</sup> Along with the common support condition

$$0 < \Pr(W = 1|X) < 1, \quad (12)$$

Equation (11) allows identification of treatment effect parameters. Equation (11) implies

$$E[Y_1 - Y_0|X] = E[Y|X, W = 1] - E[Y|X, W = 0].$$

Now, the common support condition in Equation (12) implies

$$\tau_{ATE} = E\left[E[Y|X, W = 1] - E[Y|X, W = 0]\right] \quad (13)$$

and

$$\tau_{ATE|T} = E\left[E[Y|X, W = 1] - E[Y|X, W = 0] \mid W = 1\right]. \quad (14)$$

Other causal parameters of interest, like the quantile treatment effects in Equation (7), are identified by the combination of unconfoundedness and common support (Firpo, 2007).

The common support condition can be directly assessed in the data, as it depends on the conditional distribution of  $X$  given  $W$ , which is identified by the joint distribution of  $(Y, W, X)$ . Unconfoundedness, however, is harder to assess, as it depends on potential outcomes which are not always observed. Pearl (2009) and others have investigated graphical causal structures that allow identification by conditioning on covariates. In particular, Pearl's Backdoor Criterion (Pearl, 1993) provides sufficient conditions under which treatment effect parameters are identified via conditioning.<sup>7</sup>

In our college financial aid example, suppose that other causes of the award, aside from the college entry exam, do not have a direct effect on college grades (that is, all the effect of  $V$  on  $Y$  is through  $W$ ). This is the setting in Figure 3(b).<sup>8</sup> Using Pearl's Backdoor Criterion and provided that a common support condition holds, it can be shown that the causal

---

<sup>6</sup>The unconfoundedness assumption in Equation (11) is often referred to as conditional independence, exogeneity, selection on observables, ignorability, or missing at random.

<sup>7</sup>See Pearl (2009), 2009, Morgan and Winship (2015) for detailed introductions to identification in DAGs, and Richardson and Robins (2013) for the relationship between identification in DAGs and statements about conditional independence between the treatment and potential outcomes similar to that in Equation (11).

<sup>8</sup>Notice that the node representing  $V$  disappears from the DAG in Figure 3(b). Because  $V$  causes only one variable in the DAG, it cannot create confounding. As a result, it can be safely excluded from the DAG.

structure in Figure 3(b) implies that treatment effect parameters are identified by adjusting for  $X$ , as in Equations (13) and (14). The intuition for this result is rather immediate. Pre-college ability,  $U$ , is a common cause of  $W$  and  $Y$ , and the only confounder in this DAG. However, conditioning on  $X$  “blocks” the path from  $U$  to  $X$ . Once we condition on  $X$ , the variable  $U$  is no longer a confounder, as it is not a cause of  $W$ . In contrast, if other causes of the award, aside from the college entry exam, have a direct effect on college grades as in Figure 3(a), Pearl’s Backdoor Criterion does not imply Equations (13) or (14). Additional confounding, created by  $V$ , is not controlled for by conditioning on  $X$  only.

## 4.2. Regression Adjustments

There exists a wide array of methods for estimation and inference under unconfoundedness. However, it is important to notice that, in the absence of additional assumptions, least squares regression coefficients do not identify ATE or ATET. Consider a simple setting where  $X$  takes on a finite number of values  $x_1, \dots, x_m$ . Under Assumptions (11) and (12), ATE and ATET are given by

$$\tau_{\text{ATE}} = \sum_{k=1}^m \left( E[Y|X = x_k, W = 1] - E[Y|X = x_k, W = 0] \right) \Pr(X = x_k)$$

and

$$\tau_{\text{ATET}} = \sum_{k=1}^m \left( E[Y|X = x_k, W = 1] - E[Y|X = x_k, W = 0] \right) \Pr(X = x_k | W = 1).$$

These identification results form the basis for the subclassification estimators of ATE and ATET in Cochran (1968) and Rubin (1977). One could, however, attempt to estimate ATE or ATET using least squares. In particular, one could consider estimating  $\tau_{\text{OLS}}$  by least squares in the regression equation

$$Y_i = \tau_{\text{OLS}} W_i + \sum_{k=1}^m \beta_{\text{OLS},k} D_{ki} + \varepsilon_i,$$

under the usual restrictions on  $\varepsilon_i$ , where  $D_{ki}$  is a binary variable that takes value one if  $X_i = x_k$  and zero otherwise. It can be shown that, in the absence of additional restrictive

assumptions,  $\tau_{\text{OLS}}$  differs from  $\tau_{\text{ATE}}$  and  $\tau_{\text{ATET}}$ . In particular,

$$\tau_{\text{OLS}} = \sum_{k=1}^m \left( E[Y|X = x_k, W = 1] - E[Y|X = x_k, W = 0] \right) w_k$$

where

$$w_k = \frac{\text{var}(W|X = x_k) \Pr(X = x_k)}{\sum_{r=1}^m \text{var}(W|X = x_r) \Pr(X = x_r)}.$$

That is,  $\tau_{\text{OLS}}$  identifies a variance-weighted average treatment effect (see, e.g., Angrist and Pischke, 2008).<sup>9</sup> Even in this simple setting with a regression specification that is fully saturated in all values of  $X$  (including a dummy variable for each possible value),  $\tau_{\text{OLS}}$  differs from  $\tau_{\text{ATE}}$  and  $\tau_{\text{ATET}}$  except for special cases (e.g., when  $E[Y|X = x_k, W = 1] - E[Y|X = x_k, W = 0]$  is the same for all  $k$ ).

### 4.3. Matching Estimators

Partly because of the failure of linear regression to estimate conventional treatment effects parameters like ATE and ATET, researchers have resorted to flexible/nonparametric methods like matching.

#### 4.3.1. Matching on Covariates

Matching estimators of ATE and ATET can be constructed in the following manner. First, for each sample unit  $i$  select a unit  $j(i)$  in the opposite treatment group with similar covariate values. That is,  $W_{j(i)} = 1 - W_i$  and  $X_{j(i)} \simeq X_i$ . Then, a one-to-one nearest-neighbor matching estimator of ATET can be obtained as the average of the differences in outcomes between the treated units and their matches,

$$\hat{\tau}_{\text{ATET}} = \frac{1}{n_1} \sum_{i=1}^n W_i (Y_i - Y_{j(i)}). \quad (15)$$

---

<sup>9</sup>It can be shown that  $\text{var}(W|X = x_r)$  is maximal when  $\Pr(W = 1|X = x_r) = 1/2$ , and is decreasing as this probability moves towards zero or one. That is,  $\tau_{\text{OLS}}$  weights up groups with  $X = x_k$  where the size of the treated and untreated populations are roughly equal, and weights down groups with large imbalances in the sizes of these two groups.

An estimator of ATE can be obtained in a similar manner, but using the matches for both treated and non-treated,

$$\hat{\tau}_{\text{ATE}} = \frac{1}{n} \sum_{i=1}^n W_i(Y_i - Y_{j(i)}) - \frac{1}{n} \sum_{i=1}^n (1 - W_i)(Y_i - Y_{j(i)}). \quad (16)$$

Nearest-neighbor matching comes in many varieties. It can be done with replacement (that is, potentially using each unit in the control group as a match more than once) or without, with only one or several matches per unit (in which case, the outcomes of the matches for each unit are usually averaged before inserting them in Equations (15) and (16)), and there exist several potential choices for the distance that measures the discrepancies in the values of the covariates between units (like the normalized Euclidean distance and the Mahalanobis distance, see Imbens (2004) and Abadie and Imbens (2011)).

#### 4.3.2. Matching on the Propensity Score

Matching on the propensity score is another common variety of matching methods. Rosenbaum and Rubin (1983) define the propensity score as the conditional probability of receiving the treatment given covariate values,

$$p(X) = \Pr(W = 1|X).$$

Rosenbaum and Rubin (1983) prove that if the conditions in Equations (11) and (12) hold, then the same conditions hold after replacing  $X$  with the propensity score  $p(X)$ . An implication of this result is that if controlling for  $X$  makes the treatment unconfounded, then controlling for  $p(X)$  makes the treatment unconfounded too. Figure 4 provides intuition for this result. All the effect of  $X$  on  $W$  is mediated by the propensity score. Therefore, after conditioning on  $p(X)$ , the covariate  $X$  is no longer a confounder. In other, words conditioning on the propensity score blocks the backdoor path between  $W$  and  $Y$ . Therefore, any remaining statistical dependence between  $W$  and  $Y$  is reflective of the causal effect of  $W$  on  $Y$ . The propensity score result in Rosenbaum and Rubin (1983) motivates matching estimators that match on the propensity score,  $p(X)$ , instead on matching on the entire vector of covariates,  $X$ . Matching on the propensity score reduces the dimensionality of

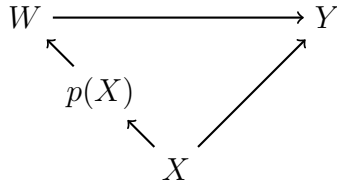


Figure 4: Identification by Conditioning on the Propensity Score

the matching variable, avoiding biases generated by curse of dimensionality issues (see, e.g., Abadie and Imbens, 2006). In most practical settings, however, the propensity score is unknown. Nonparametric estimation of the propensity score brings back dimensionality issues. In some settings, estimation of the propensity score can be guided by institutional knowledge about the process that produces treatment assignment (elicited from experts in the subject matter). In empirical practice, the propensity score is typically estimated using parametric methods like probit or logit.

Abadie and Imbens (2006, 2011, 2012) derive large sample results for estimators that match directly on the covariates,  $X$ . Abadie and Imbens (2016) provide large sample results for the case where matching is done on an estimated propensity score. For a review and further references on propensity score matching estimators, see Rosenbaum (2010), Imbens and Rubin (2015) and references therein.

#### 4.4. Inverse Probability Weighting

Inverse probability weighting methods (IPW, see Robins et al., 1994; Hirano et al., 2003, and references therein) are also based on the propensity score and provide an alternative to matching estimators. This approach proceeds by first obtaining estimates of propensity score values,  $\hat{p}(X_i)$ , and then using those estimates to weight outcome values. For example, an IPW estimator of ATE is

$$\hat{\tau}_{\text{ATE}} = \frac{1}{n} \sum_{i=1}^n \frac{W_i Y_i}{\hat{p}(X_i)} - \frac{1}{n} \sum_{i=1}^n \frac{(1 - W_i) Y_i}{1 - \hat{p}(X_i)}.$$

This estimator uses  $1/\hat{p}(X_i)$  to weigh observations in the treatment group and  $1/(1 - \hat{p}(X_i))$  to weigh observations in the untreated group. Intuitively, observations with large

$\hat{p}(X_i)$  are over-represented in the treatment group and, hence, weighted down when treated. The same observations are weighted up when untreated. The opposite applies to observations with small  $\hat{p}(X_i)$ . This estimator can be modified so that the weights in each treatment arm sum to one, which produces improvements in finite sample performance (see, e.g., Busso et al., 2014).

#### 4.5. Imputation/Projection Methods

Imputation/projection methods (e.g., Heckman et al., 1998; Little and Rubin, 2002; Imbens et al., 2006; Cattaneo and Farrell, 2011) provide an alternative class of estimators that rely on the assumptions in Equations (11) and (12). Regression imputation estimators are based on preliminary estimates of the outcome process, that is, the conditional distribution of  $Y$  given  $(X, W)$ . Let  $\hat{\mu}_1(x)$  and  $\hat{\mu}_0(x)$  be parametric or nonparametric estimators of  $E[Y|X = x, W = 1]$  and  $E[Y|X = x, W = 0]$ , respectively. Then, an empirical analog of Equation (13) provides a projection/imputation estimator of ATE:

$$\hat{\tau}_{\text{ATE}} = \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)).$$

Similarly, a projection/imputation estimator of ATET is given by the empirical analog of Equation (14).

#### 4.6. Hybrid Methods

Several hybrid methods combine matching/propensity score weighting with projection/imputation techniques. Among them are the doubly robust estimators (or, more generally, locally robust estimators) in Van der Laan and Robins (2003), Bang and Robins (2005), Cattaneo (2010), Farrell (2015), Chernozhukov et al. (2016), and Słoczyński and Wooldridge (2017). In the case of ATE estimation, for example, they can take the form:

$$\begin{aligned} \hat{\tau}_{\text{ATE}} = & \frac{1}{n} \sum_{i=1}^n \left( \frac{W_i (Y_i - \hat{\mu}_1(X_i))}{\hat{p}(X_i)} + \hat{\mu}_1(X_i) \right) \\ & - \frac{1}{n} \sum_{i=1}^n \left( \frac{(1 - W_i) (Y_i - \hat{\mu}_0(X_i))}{1 - \hat{p}(X_i)} + \hat{\mu}_0(X_i) \right). \end{aligned}$$



This estimator is doubly robust in the sense that consistent estimation of  $p(\cdot)$  *or* consistent estimation of  $\mu_1(\cdot)$  and  $\mu_0(\cdot)$  is required for the validity of the estimator. Part of the appeal of doubly and locally robust methods come from the efficiency properties of the estimators. However, relative to other methods discussed in this section, doubly robust methods require estimating not only  $p(\cdot)$  or  $\mu_1(\cdot)$  and  $\mu_0(\cdot)$ , but all these functions. As a result, relative of other methods, doubly robust estimators involve additional tuning parameters and implementation choices. Closely related to doubly/locally robust methods are the bias-corrected matching estimators of Abadie and Imbens (2011). These estimators include a regression-based bias correction designed to purge matching estimators from the bias that arises from imperfect matches in the covariates,

$$\begin{aligned}\hat{\tau}_{\text{ATE}} = & \frac{1}{n} \sum_{i=1}^n W_i \left( (Y_i - Y_{j(i)}) - (\hat{\mu}_0(X_i) - \hat{\mu}_0(X_{j(i)})) \right) \\ & - \frac{1}{n} \sum_{i=1}^n (1 - W_i) \left( (Y_i - Y_{j(i)}) - (\hat{\mu}_1(X_i) - \hat{\mu}_1(X_{j(i)})) \right).\end{aligned}$$

#### 4.7. Comparisons of Estimators

Busso et al. (2014) study the finite sample performance of matching and IPW estimators, as well as some of their variants. Kang and Schafer (2007) and the accompanying discussions and rejoinder provide finite sample evidence on the performance of double robust estimators and other related methods.<sup>10</sup>

As these and other studies demonstrate, the relative performance of estimators based on conditioning on observables depends on the features of the data generating process. In finite samples, IPW estimators became unstable when the propensity score approaches zero or one, while regression imputation methods may suffer from extrapolation biases. Estimators that match directly on covariates do not require specification choices, but may incorporate non-trivial biases if the quality of the matches is poor. Hybrid methods, such as bias-corrected matching estimators, or doubly-robust estimators, include safeguards against bias caused by

---

<sup>10</sup>Dehejia and Wahba (1999), Smith and Todd (2005), and others evaluate the performance of some of the estimators in this section against an experimental benchmark.

imperfect matching and misspecification, but impose additional specification choices that may affect the resulting estimate.

Apart from purely statistical properties, estimators based on conditioning on covariates differ in the way they are related to research transparency. In particular, during the design phase of a study (i.e., sample, variable, and model selection), matching and IPW methods employ only information about treatment,  $W$ , and covariates,  $X$ . That is, the matches and the specification for the propensity score can be constructed without knowledge or use of the outcome variable. This feature of matching and IPW estimator provides a potential safeguard against specification searches and  $p$ -hacking (a point forcefully made in Rubin, 2007).

#### **4.8. Recent Developments and Additional Topics**

All the estimation and inference methods discussed so far take the covariates,  $X$ , as known and “small” relative to the sample size, though in practice researchers often use high-dimensional models with the aim of using a setting where the conditional independence assumption is deemed appropriate. In most applications,  $X$  includes both raw pre-intervention covariates as well as transformations thereof, such as dummy variable expansions of categorical variables, power or other series expansions of continuous variables, higher-order interactions, or other technical regressors generated from the original available variables. The natural tension between standard theory (where the dimension of  $X$  is taken to be small) and practice (where the dimension of  $X$  tends to be large), has led to a new literature that aims to develop estimation and inference methods in program evaluation that account/allow for high-dimensional  $X$ .

Belloni et al. (2014) and Farrell (2015) develop new program evaluation methods employing machinery from the high-dimensional literature in statistics. This methodology, which allows for very large  $X$  (much larger than the sample size), proceeds in two steps. First, a parsimonious model is selected from the large set of pre-intervention covariates employing model selection via LASSO. Then, treatment effect estimators are constructed using only

the small (usually much smaller than the sample size) set of selected covariates. More general methods and a review of this literature are given in Belloni et al. (2017) and Athey et al. (2016), among others. Program evaluation methods proposed in this literature employ modern model selection techniques, which require a careful analysis and interpretation, but ultimately often produce classical distributional approximations (in the sense that these asymptotic approximations do not change relative to results in low-dimensional settings).

Cattaneo et al. (2018c,d) and Cattaneo et al. (2017a) develop program evaluation methods where the distributional approximations do change relative to low-dimensional results, due to the inclusion of many covariates in the estimation. These results can be understood as giving new distributional approximations that are robust to (i.e., also valid in) cases where either the researcher cannot select out covariates (e.g., when many multi-way fixed effects are needed) or many covariates remain included in the model even after model selection. These high-dimensional methods are not only valid when many covariates are included, but also continue to be valid in cases where only a few covariates are used, thereby offering demonstrable improvements for estimation and inference in program evaluation. Related kernel-based methods are developed in Cattaneo and Jansson (2018).

All the ideas and methods discussed so far are mostly concerned with average treatment effects in the context of binary treatments. Many of these results have been extended to multi-valued treatments (Imbens, 2000; Lechner, 2001; Hirano and Imbens, 2004; Imai and van Dyk, 2004; Cattaneo, 2010; Farrell, 2015; Yang et al., 2016), to quantiles and related treatment effects (Firpo, 2007; Cattaneo, 2010), and to the analysis and interpretation of counterfactual distributional treatment effects (DiNardo et al., 1996; Chernozhukov et al., 2013; Firpo and Pinto, 2016). We do not discuss this work due to space limitations.

Space restrictions also prevent us from discussing in detail other strands of the literature on estimation of treatment effects based on conditioning on covariates. In particular, in the biostatistics literature, structural nested models, marginal structural models, and optimal treatment regimes estimators are employed to estimate treatment effects in contexts with time varying treatments and confounders (see, e.g., Robins, 2000; Lok et al., 2004; Murphy,

2003). Also, a large literature on sensitivity to unobserved confounders analyzes the impact of departures from the unconfoundedness assumption in Equation (11), (see, e.g., Rosenbaum, 2002; Imbens, 2003; Altonji et al., 2005).

## 5. Difference in Differences and Synthetic Controls

### 5.1. Difference in Differences

In the previous section, we considered settings where treatment assignment is confounded but there exists a set of observed covariates,  $X$ , such that treatment assignment becomes unconfounded after conditioning on  $X$ . In many applied settings, however, researchers confront the problem of the possible existence of unobserved confounders. In these settings, difference-in-differences models aim to attain identification by restricting the way in which unobserved confounders affect the outcome of interests over time. Consider the following panel data regression model,

$$Y_{it} = W_{it}\tau_{it} + \mu_i + \delta_t + \varepsilon_{it},$$

where only  $Y_{it}$  and  $W_{it}$  are observed. We will regard the mapping represented in this equation as structural (or causal). That is, the equation describes potential outcomes, which are now indexed by time period,  $t$ ,

$$\begin{aligned} Y_{0it} &= \mu_i + \delta_t + \varepsilon_{it} \\ Y_{1it} &= \tau_{it} + \mu_i + \delta_t + \varepsilon_{it}. \end{aligned} \tag{17}$$

Then,  $\tau_{it} = Y_{1it} - Y_{0it}$ . To simplify the exposition, and because it is a common setting in empirical practice, we will assume that there are only two time periods. Period  $t = 0$  is the pre-intervention period, before the treatment is available, so  $W_{i0} = 0$  for all  $i$ . Period  $t = 1$  is the post-treatment period, when a fraction of the population units are exposed to the treatment.  $\delta_t$  is a time effect, common across units. We will treat  $\mu_i$  as a time invariant confounder, so  $\mu_i$  and  $W_{i1}$  are not independent. In contrast, we will assume that  $\varepsilon_{it}$  are causes of the outcome that are unrelated to selection for treatment, so  $E[\varepsilon_{it}|W_{it}] = E[\varepsilon_{it}]$ .

This condition can be weakened to  $E[\Delta\varepsilon_{i1}|W_{it}] = E[\Delta\varepsilon_{i1}]$ , where  $\Delta$  is the first difference operator, so  $\Delta\varepsilon_{i1} = \varepsilon_{i1} - \varepsilon_{i0}$ . This type of structure is the same as the widespread linear fixed effect panel data model (see, e.g., Arellano, 2003). Then,

$$\begin{aligned} E[Y_{i1}|W_{i1} = 1] &= E[\tau_{i1}|W_{i1} = 1] + E[\mu_i|W_{i1} = 1] + \delta_1 + E[\varepsilon_{i1}] \\ E[Y_{i0}|W_{i1} = 1] &= E[\mu_i|W_{i1} = 1] + \delta_0 + E[\varepsilon_{i0}] \\ E[Y_{i1}|W_{i1} = 0] &= E[\mu_i|W_{i1} = 0] + \delta_1 + E[\varepsilon_{i1}] \\ E[Y_{i0}|W_{i1} = 0] &= E[\mu_i|W_{i1} = 0] + \delta_0 + E[\varepsilon_{i0}]. \end{aligned}$$

In this model, the effect of the unobserved confounders on the average of the outcome variable is additive and does not change in time. It follows that,

$$\begin{aligned} \tau_{\text{ATE}} &= E[\tau_{i1}|W_{i1} = 1] \\ &= \left[ E[Y_{i1}|W_{i1} = 1] - E[Y_{i1}|W_{i1} = 0] \right] - \left[ E[Y_{i0}|W_{i1} = 1] - E[Y_{i0}|W_{i1} = 0] \right] \\ &= \left[ E[\Delta Y_{i1}|W_{i1} = 1] - E[\Delta Y_{i1}|W_{i1} = 0] \right]. \end{aligned} \tag{18}$$

The reason for the name “difference in differences” is apparent from Equation (18). Notice that  $\tau_{\text{ATE}}$  is defined here for the post-treatment period,  $t = 1$ . Intuitively, identification in the difference in differences model comes from a common trends assumption implied by Equation (17). It can be easily seen that Equation (17), along with  $E[\Delta\varepsilon_{i1}|W_{it}] = E[\Delta\varepsilon_{i1}]$ , implies

$$E[\Delta Y_{0i1}|W_{it} = 1] = E[\Delta Y_{0i1}|W_{it} = 0]. \tag{19}$$

That is, in the absence of the treatment, the average outcome for the treated and the average outcome for the nontreated would have experienced the same variation over time.

Figure 5 illustrates how identification works in a difference in differences model. In this setting, a set of untreated units identifies the average change that the outcome for the treated units would have experienced in the absence of the treatment (the dashed line in Figure 5). The set of untreated units selected to reproduce the counterfactual trajectory of the outcome for the treated is often called the control group in this literature (borrowing from the literature on randomized control trials).

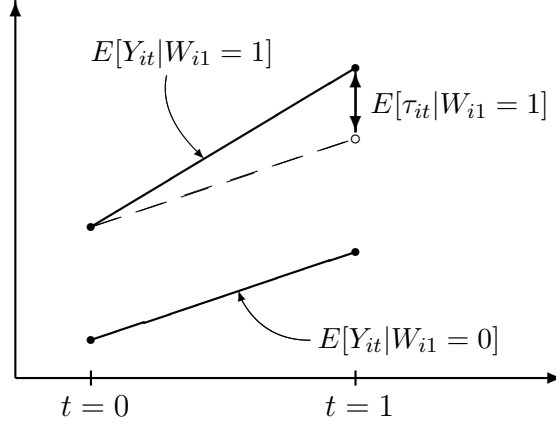


Figure 5: Identification in a Difference-in-Differences Model

Notice also that the common trend assumption in Equation (19) is not invariant to nonlinear transformations of the dependent variable. For example, if Equation (19) holds when the outcome is a wage rate measured in levels, then the same equation will not hold in general for wages measured in logs. In other words, identification in a difference in differences model is not invariant to nonlinear transformations in the dependent variable.

In a panel data regression (that is, in a setting with repeated pre- and post-treated observations for the same units), the right hand side of Equation (18) is equal to the regression coefficient on  $W_{i1}$  in a regression of  $\Delta Y_{i1}$  on  $W_{i1}$  and a constant. Consider instead a regression setting with pooled cross-sections for the outcome variable at  $t = 0$  and  $t = 1$ , and information on  $W_{i1}$  for all of the units in each cross-section. In this setting, one cross-section contains information on  $(Y_{i0}, W_{i1})$  and the other cross-section contains information on  $(Y_{i1}, W_{i1})$ . Let  $T_i$  be equal to zero if unit  $i$  belongs to the pre-treatment sample and equal to one if it belongs to the post-treatment sample. Then, the right hand side of Equation (18) is equal to the coefficient on the interaction  $W_{i1}T_i$  in a regression of the outcome on a constant,  $W_{i1}$ ,  $T_i$ , and the interaction,  $W_{i1}T_i$  for a sample that pools the pre-intervention and post-intervention cross-sections.

Several variations of the basic difference in difference model have been proposed in the literature. In particular, the model naturally extends to a fixed-effects regression in set-

tings with more than two periods, and/or models with unit-specific linear trends (see, e.g., Bertrand et al., 2004). Logically, estimating models with many time periods and/or trends imposes greater demands on the data.

The common trends restriction in Equation (19) is clearly a strong assumption, and one that should be assessed in empirical practice. The plausibility of this assumption can sometimes be evaluated using (i) multiple pre-intervention periods (as in, e.g., Abadie and Dermisi, 2008), or (ii) population groups that are not at risk of being exposed to the treatment, as in Gruber (1994). In both cases, a test for common trends is based on the difference in differences estimate of the effect of a placebo intervention, that is, an intervention that did not happen. In case (i), the placebo estimate is computed using the preintervention data only and evaluates the effect of a non-existent intervention taking place before the actual intervention period. Rejection of a null effect of the placebo intervention provides direct evidence against common trends before the intervention. Case (ii) is similar, but uses an estimate obtained for a population group known not to be amenable to receive the treatment. For example, Gruber (1994) uses difference in differences to evaluate the effect of the passage of mandatory maternity benefits in some U.S. states on the wages of married women of childbearing age. In this context, Equation (19) means that, in the absence of the intervention, married women of childbearing age would have experienced the same increase in log wages in states that adopted mandated maternity benefits and states that did not. To evaluate the plausibility of this assumption, Gruber (1994) compares the changes in log wages in adopting and non-adopting states for single men and women over 40 year old.

The plausibility of Equation (19) may also be questioned if the treated and the control groups are different in the distribution of attributes that are known or suspected to affect the outcome trend. For example, if average earnings depend on age or on labor market experience in a nonlinear manner, then differences in the distributions of these two variables between treated and nontreated in the evaluation of an active labor market program pose a threat to the validity of the conventional difference in differences estimator. Abadie (2005b) proposes a generalization of the difference in differences model for the case when covariates explain

the differences in the trends of the outcome variable between treated and nontreated. The resulting estimators adjust the distribution of the covariates between treated and nontreated using propensity score weighting.

Finally, Athey and Imbens (2006) provide a generalization of the difference in differences model to the case when  $Y_{0it}$  is nonlinear in the unobserved confounder,  $\mu_i$ . Identification in this model comes from strict monotonicity of  $Y_{0it}$  with respect to  $\mu_i$  and from the assumption that the distribution of  $\mu_i$  is time invariant for the treated and the nontreated (although it might differ between the two groups). One of the advantages of their approach is that it provides an identification result that is robust to monotonic transformations of the outcome variable (e.g., levels,  $Y_{it}$ , vs. logs,  $\log(Y_{it})$ ).

## 5.2. Synthetic Controls

Difference in differences estimators are often used to evaluate the effects of events or interventions that affect entire aggregate units, such as states, school districts, or countries (see, e.g., Card, 1990; Card and Krueger, 1994). For example, Card (1990) estimates the effect of the Mariel Boatlift, a large and sudden influx of immigrants to Miami, on the labor market outcomes of native workers in Miami. In a difference in differences design, Card (1990) uses a combination of four other cities in the south of the U.S. (Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg) to approximate the change in native unemployment rates that would have been observed in Miami in the absence of the Mariel Boatlift. Studies of this type are sometimes referred to as “comparative case studies”, because a comparison group is selected to evaluate an instance or case of an event or policy of interest. In Card (1990), the case of interest is the arrival of Cuban immigrants to Miami during the Mariel Boatlift, and the comparison is provided by the combination of Atlanta, Los Angeles, Houston, and Tampa-St. Petersburg.

The synthetic control estimator of Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015) provides a data driven procedure to create a comparison unit in comparative case studies. A synthetic control is a weighted average of untreated units chosen to reproduce



characteristics of the treated unit before the intervention. The idea behind synthetic controls is that a combination of untreated units, as in Card (1990), may provide a better comparison to the treatment unit than any untreated unit alone.

Synthetic controls are constructed as follows. For notational simplicity, we will assume that there is only one treated unit. This is without loss of generality, because when there are more than one treated unit the procedure described next can be applied for each treated unit separately. Suppose that we observe  $J + 1$  units in periods  $1, 2, \dots, T$ . Unit “one” is exposed to the intervention of interest during periods  $T_0 + 1, \dots, T$ . The remaining  $J$  units are an untreated reservoir of potential controls (a “donor pool”). Let  $w = (w_2, \dots, w_{J+1})'$  with  $w_j \geq 0$  for  $j = 2, \dots, J + 1$  and  $w_2 + \dots + w_{J+1} = 1$ . Each value of  $w$  represents a potential synthetic control. Let  $X_1$  be a  $(k \times 1)$  vector of pre-intervention characteristics for the treated unit. Similarly, let  $X_0$  be a  $(k \times J)$  matrix which contains the same variables for the unaffected units. The vector  $w^* = (w_2^*, \dots, w_{J+1}^*)'$  is chosen to minimize  $\|X_1 - X_0 w\|$ , subject to the weight constraints. The synthetic control estimator of the effect of the treatment for the treated unit in a post-intervention period  $t$  ( $t \geq T_0$ ) is:

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt}.$$

A weighted Euclidean norm is commonly employed to measure the discrepancy between the characteristics of the treated unit and the characteristics of the synthetic control

$$\|X_1 - X_0 w\| = \sqrt{(X_1 - X_0 w)' V (X_1 - X_0 w)},$$

where  $V$  is a diagonal matrix with non-negative elements in the main diagonal that control the relative importance of obtaining a good match between each value in  $X_1$  and the corresponding value in  $X_0 w^*$ . Abadie and Gardeazabal (2003) and Abadie et al. (2010, 2015) propose data-driven selectors of  $V$ . Abadie et al. (2010) propose an inferential method for synthetic controls based on Fisher’s randomization inference.

It can be shown that, in general (that is, ruling out certain degenerate cases), if  $X_1$  does not belong to the convex hull of the columns of  $X_0$ , then  $w^*$  is unique and sparse. Sparsity

means that  $w^*$  has only a few non-zero elements. However, in some applications, especially in applications with a large number of treated units, there may be synthetic controls that are not be unique or sparse. Abadie and L’Hour (2017) propose a variant of the synthetic control estimator that addresses this issue. Their estimator minimizes

$$\|X_1 - X_0 w\|^2 + \lambda \sum_{j=2}^{J+1} w_j \|X_j - X_1\|^2,$$

with  $\lambda > 0$ , subject to the weight constraints. This estimator includes a penalization for pairwise matching discrepancies between the treated unit and each of the units that contribute to the synthetic control. Abadie and L’Hour (2017) show that aside from special cases, if  $\lambda > 0$  the synthetic control estimator is unique and sparse. Hsiao et al. (2012) and Doudchenko and Imbens (2016) propose variants of the synthetic control framework that extrapolate outside the support of the columns of  $X_0$ . Recent strands of the literature extend synthetic control estimators using matrix completion techniques (see Amjad et al., 2017; Athey et al., 2017) and propose sampling-base inferential methods (Hahn and Shi, 2017).

## 6. Instrumental Variables

### 6.1. The Framework

Instrumental variables (IV) methods are of widespread use in program evaluation. They are easiest to motivate in the context of a randomized experiment with imperfect compliance, that is, a setting where experimental units may fail to comply with the randomized assignment of the treatment. Other settings where IV methods are used include natural experiments in observation studies, measurement error, non-linear models with endogeneity, and dynamic/panel data/time series contexts (see, e.g., Wooldridge, 2010; Imbens, 2014, for an overview and further references, including material on weak and many instruments).

Imperfect compliance with a randomized assignment is commonplace in economics and other disciplines where assignment affects human subjects, so compliance may be difficult to enforce or not desirable (for example, in active labor market programs for the unemployed

where the employment status of the individuals participating in the experiment may change between randomization and treatment). In settings where the experimenter is the only possible provider of the treatment, randomized assignment can be enforced in the control group. Such a setting is often referred to as one-sided noncompliance.

In randomized experiments with noncompliance, randomization of treatment intake fails because of post-randomization selection into the treatment. Assignment itself remains randomized and the average effect of assignment on an observed outcome is identified. Assignment effects are often referred to as intention-to-treat effects. In some settings, the most relevant policy decision is whether or not to provide individuals with access to treatment, so intention-to-treat effects have direct policy relevance. In other instances, researchers and policy makers are interested in the effects of treatment intake, rather than on the effects of eligibility for treatment. In such cases, instrumental variable methods can be used to identify certain treatment effects parameters.

Instrumental variable methods rely on the availability of instruments, that is, variables that affect treatment intake but do not directly affect potential outcomes. In the context of randomized experiments with imperfect compliance, instruments are usually provided by randomized assignment to treatment. We will use  $Z$  to represent randomized assignment, so  $Z = 1$  indicates assignment to the treatment group and  $Z = 0$  indicates assignment to the control group. Like in previous sections, actual treatment intake will be represented by the binary variable,  $W$ .

Figure 6 offers a DAG representation of the basic instrumental variable setting. The treatment,  $W$ , is affected directly by an unobservable confounder,  $U$ , which also directly affects the outcome variable,  $Y$ . Because  $U$  is unobserved, conditioning on  $U$  is unfeasible. However, now there is an instrument,  $Z$ , which affects treatment,  $W$ , and only affects the outcome,  $Y$ , through  $W$ .

The key idea underlying instrumental variable methods is that the exogenous variation in the instrument,  $Z$ , which induces changes in the treatment variable,  $W$ , can be used to identify certain parameters of interest in program evaluation and causal inference. Variation

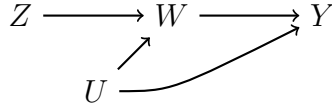


Figure 6: Instrumental Variable in a DAG

in  $Z$  is exogenous (or unconfounded) because  $Z$  affects  $Y$  only through  $W$ , and because there are no common causes of  $Z$  and  $Y$ . Chalak and White (2011) provide a detailed account of identification procedures via instrumental variables.

## 6.2. Local Average Treatment Effects

Employing a potential outcomes framework, Imbens and Angrist (1994) and Angrist et al. (1996) provide a formal setting to study identification and estimation of treatment effects under imperfect compliance. The observed treatment status is

$$W = ZW_1 + (1 - Z)W_0,$$

where now  $W_1$  and  $W_0$  denote the potential treatment status under treatment and control assignment, respectively. Perfect compliance corresponds to  $\Pr(W_1 = 1) = \Pr(W_0 = 0) = 1$ , so  $\Pr(W = Z) = 1$ . One-sided noncompliance corresponds to the case when  $\Pr(W_0 = 0) = 1$  but  $\Pr(W_1 = 1) < 1$ . In this case, units assigned to the control group never take the treatment, while units assigned to the treatment group may fail to take the treatment.

More generally, in the absence of perfect or one-sided compliance, the population can be partitioned into four groups defined in terms of the values of  $W_1$  and  $W_0$ . Angrist et al. (1996) define:

Compliers:	$W_1 > W_0$ (or $W_0 = 0$ and $W_1 = 1$ )
Always-takers:	$W_1 = W_0 = 1$
Never-takers:	$W_1 = W_0 = 0$
Defiers:	$W_1 < W_0$ ( $W_0 = 1$ and $W_1 = 0$ )

Experimental units that comply with treatment assignment in both treatment arms are called compliers. Always-takers and never takers are not affected by assignment. Defiers are affected by assignment in the opposite direction as expected: they take the treatment when

they are assigned to the control group but refuse to take it when they are assigned to the treatment group.

We will assume that treatment assignment is not trivial (in the sense that assignment is not always to the treatment group or always to the control group). In addition, we will assume that the assignment has an effect on treatment intake:

$$0 < \Pr(Z = 1) < 1 \quad \text{and} \quad \Pr(W_1 = 1) \neq \Pr(W_0 = 1). \quad (20)$$

A key assumption in the instrumental variable setting is that potential outcomes,  $Y_1$  and  $Y_0$ , do not depend on the value of the instrument. This is called the exclusion restriction. In Figure 6, the exclusion restriction is manifested in that the effect of  $Z$  on  $Y$  is completely mediated by  $W$ . The combination of the exclusion restriction with the assumption that the instrument is randomized (or as good as randomized) implies:

$$(Y_1, Y_0, W_1, W_0) \perp\!\!\!\perp Z. \quad (21)$$

In the language of instrumental variable estimation of linear regression models with constant coefficients, Equations (20) and (21) amount to the assumptions of instrument relevance and instrument exogeneity (see, e.g., Stock and Watson, 2003).

Consider the instrumental variable parameter

$$\tau_{\text{IV}} = \frac{\text{cov}(Y, Z)}{\text{cov}(D, Z)} = \frac{E[Y|Z = 1] - E[Y|Z = 0]}{E[W|Z = 1] - E[W|Z = 0]}. \quad (22)$$

Under the assumptions in Equations (20) and (21), it can be shown that

$$\tau_{\text{IV}} = \frac{E[(Y_1 - Y_0)(W_1 - W_0)]}{E[W_1 - W_0]}. \quad (23)$$

It follows that, if the treatment does not vary across units, that is, if  $Y_1 - Y_0$  is constant,  $\tau_{\text{IV}}$  is equal to the constant treatment effect.<sup>11</sup> The sample analog of  $\tau_{\text{IV}}$ ,

$$\hat{\tau}_{\text{IV}} = \frac{\left( \sum_{i=1}^n Z_i Y_i / \sum_{i=1}^n Z_i \right) - \left( \sum_{i=1}^n (1 - Z_i) Y_i / \sum_{i=1}^n (1 - Z_i) \right)}{\left( \sum_{i=1}^n Z_i W_i / \sum_{i=1}^n Z_i \right) - \left( \sum_{i=1}^n (1 - Z_i) W_i / \sum_{i=1}^n (1 - Z_i) \right)},$$

---

<sup>11</sup>More generally, if  $Y_1 - Y_0$  is mean independent of  $W_1 - W_0$ , then  $\tau_{\text{IV}}$  is equal to ATE.

provides an estimator of the constant average treatment effect. This is often called the Wald estimator.

In the presence of treatment effect heterogeneity, however, it can be easily shown that  $\tau_{IV}$  may fail to identify an average treatment effect for a well-defined population. In particular, it is easy to construct examples where  $Y_1 - Y_0 > 0$  with probability one, but  $\tau_{IV} = 0$ . The reason is that positive values of  $(Y_1 - Y_0)(W_1 - W_0)$  given by positive values of  $W_1 - W_0$  may cancel out with negative values in the expectation in the numerator of the expression for  $\tau_{IV}$  in Equation (23).

Imbens and Angrist (1994) define the Local Average Treatment Effect (LATE) as

$$\tau_{LATE} = E[Y_1 - Y_0 | W_1 > W_0].$$

That is, in the terminology of Angrist et al. (1996), LATE is the average effect of the treatment for compliers, or the average effect of the treatment for those units that would always be in compliance with the assignment no matter whether they are assigned to the treatment or to the control group.

Suppose that the assumptions in Equations (20) and (21) hold. Suppose also that,

$$\Pr(W_1 \geq W_0) = 1. \tag{24}$$

The assumption in Equation (24) is often referred to as monotonicity, and it rules out the existence of defiers. Then,

$$\tau_{IV} = \tau_{LATE}. \tag{25}$$

Monotonicity implies that  $W_1 - W_0$  is binary, so the result in Equation (25) follows easily from Equation (23).

LATE represents the average treatment effect for the units that change their treatment status according to the their treatment assignment. Notice that, in the absence of additional restrictions, compliers are not individually identified, because only  $W_1$  or  $W_0$  is observed in practice, but not both. However, the result in equation (25) implies that average treatment effects are identified for compliers.

A key identifying assumption for LATE is the monotonicity condition (24), which rules out defiers. Balke and Pearl (1997), Imbens and Rubin (1997), Heckman and Vytlačil (2005) and Kitagawa (2015) discuss testable implications of this assumption.

One-sided noncompliance,  $\Pr(W_0 = 1) = 0$ , is fairly common in settings where the experimenter is the only potential provider of the treatment. It has important practical implications. In particular, one-sided noncompliance implies monotonicity. It also implies that LATE is equal to ATET. That is, under one-sided noncompliance,

$$\tau_{\text{LATE}} = \tau_{\text{ATET}}.$$

More generally, however, LATE is not the same as ATE or ATET, and the interest in and interpretation of the LATE parameter has to be judged depending on the application at hand. Suppose, for example, that granting access to treatment represents a policy option of interest. Then, in a randomized experiment with imperfect compliance (and if monotonicity holds), the LATE parameter recovers the effect of the treatment for those units affected by being granted access to treatment, which is the policy decision under consideration. See, e.g., Imbens (2010) and references therein for more discussion.

Angrist et al. (2000) extend the methods described in this section to estimation in simultaneous equations models. This work connects with marginal treatment effects estimators discussed in Section 6.4. The local nature of LATE – that is, the fact that it applies to compliers only – raises the issue of under what conditions LATE estimates can be generalized to larger or different populations. For a discussion of extrapolation of LATE estimates and related problems, see Angrist and Fernandez-Val (2013).

### 6.3. General Identification and Estimation Results for Compliers (Kappa)

Abadie (2003) proposes a general class of estimators of LATE-type parameters (that is, parameters defined for the population of compliers). Like in Imbens and Angrist (1994) and Angrist et al. (1996), Abadie (2003) adopts a context akin to a randomized experiment with imperfect compliance, where the instrument and treatment are binary, extending it to the case where conditioning on covariates may be needed for instrument validity. In particular, in

the absence of unconditional randomization of the treatment, there may be common causes,  $X$ , that affect the instrument and the outcome. In this case, the instrument will not be valid unless validity is assessed after conditioning on  $X$ .

Consider versions of Equations (20), (21) and (24), where the same assumptions hold after conditioning on  $X$ . Let

$$\kappa_0 = (1 - W) \frac{(1 - Z) - \Pr(Z = 0|X)}{\Pr(Z = 0|X) \Pr(Z = 1|X)}, \quad \kappa_1 = W \frac{Z - \Pr(Z = 1|X)}{\Pr(Z = 0|X) \Pr(Z = 1|X)},$$

and

$$\kappa = 1 - \frac{W(1 - Z)}{\Pr(Z = 0|X)} - \frac{(1 - W)Z}{\Pr(Z = 1|X)}.$$

Abadie (2003) shows that,

$$E[h(Y_0, X)|W_1 > W_0] = \frac{1}{\Pr(W_1 > W_0)} E[\kappa_0 h(Y, X)], \quad (26)$$

$$E[h(Y_1, X)|W_1 > W_0] = \frac{1}{\Pr(W_1 > W_0)} E[\kappa_1 h(Y, X)], \quad (27)$$

and

$$E[g(Y, W, X)|W_1 > W_0] = \frac{1}{\Pr(W_1 > W_0)} E[\kappa g(Y, W, X)]. \quad (28)$$

where  $h$  and  $g$  are arbitrary functions (provided the expectations exist.) Moreover,

$$\Pr(W_1 > W_0) = E[\kappa_0] = E[\kappa_1] = E[\kappa].$$

Equations (26)-(28) provide a general identification results for compliers. In particular, Equations (26)-(28) identify the distribution of attributes,  $X$ , for compliers, e.g.,

$$E[X|W_1 > W_0] = E[\kappa X]/E[\kappa].$$

This result allows a more detailed interpretation of the LATE parameter in Imbens and Angrist (1994) and Angrist et al. (1996). Although, as indicated above, compliers are not identified individually, the distribution of any observed characteristic can be described for the population of compliers. More generally, (26)-(28) allow identification of regression models for compliers. Such models can be used to describe how the average effect of the treatment



for compliers changes with  $X$  (see Abadie, 2003, for details). Equations (26) and (27) can be applied to the identification of average complier responses to treatment or no treatment, as well as identification of the average treatment effect for compliers. In particular, making  $h(Y, X) = Y$  we obtain:

$$\begin{aligned} E[Y_1|W_1 > W_0] &= \frac{E[\kappa_1 Y]}{E[\kappa_1]}, \\ E[Y_0|W_1 > W_0] &= \frac{E[\kappa_0 Y]}{E[\kappa_0]}. \end{aligned}$$

Therefore,

$$\begin{aligned} \tau_{\text{LATE}} &= \frac{E[\kappa_1 Y]}{E[\kappa_1]} - \frac{E[\kappa_0 Y]}{E[\kappa_0]} \\ &= \frac{E \left[ Y \frac{Z - \Pr(Z = 1|X)}{\Pr(Z = 0|X) \Pr(Z = 1|X)} \right]}{E \left[ W \frac{Z - \Pr(Z = 1|X)}{\Pr(Z = 0|X) \Pr(Z = 1|X)} \right]}. \end{aligned}$$

Kappa-based estimators are sample analogs of the identification results in this section. They require first-step estimation of  $\Pr(Z = 1|X)$ . For the same setting, Frölich (2007) derives an alternative estimator of  $\tau_{\text{LATE}}$  based on the sample analog of the identification result

$$\tau_{\text{LATE}} = \frac{E \left[ E[Y|X, Z = 1] - E[Y|X, Z = 0] \right]}{E \left[ \Pr(W = 1|X, Z = 1) - \Pr(W = 1|X, Z = 0) \right]}.$$

An estimator of  $\tau_{\text{LATE}}$  based on this result requires first-step estimation of  $E[Y|X, Z = 1]$ ,  $E[Y|X, Z = 0]$ ,  $\Pr(W = 1|X, Z = 1)$ , and  $\Pr(W = 1|X, Z = 0)$ .

Hong and Nekipelov (2010) derive efficient instruments for kappa-based estimators. Ogburn et al. (2015) propose doubly-robust estimators of LATE-like parameters. Abadie (2002); Abadie et al. (2002); Frölich and Melly (2013) propose estimators of distributional effects for compliers. Chernozhukov and Hansen (2005, 2013) propose a rank similarity condition under which they derive an estimator of quantile treatment effects for the entire population (rather than for compliers only).

#### 6.4. Marginal Treatment Effects

A closely related causal framework for understanding instrumental variables (and other program evaluation) settings was introduced by Björklund and Moffitt (1987), and later developed and popularized by Heckman and Vytlačil (1999, 2005) and Heckman et al. (2006), and more recently studied by Kline and Walters (2016) and Brinch et al. (2017), among others. In this approach, the main parameters of interest are the marginal treatment effects (MTE) and certain functionals thereof.

To describe the model and MTE estimand, we continue to employ potential outcomes, but specialize the notation. In this setting, the instrument,  $Z$ , may be a vector of discrete or continuous random variables, and the observed data are assumed to be generated according to the following model. Let  $X$  be a vector of covariates. Conditional on  $X = x$ , we have

$$W = 1(p(Z) \geq V), \quad V|Z \sim \text{Uniform}[0, 1] \quad (29)$$

$$Y_1 = \mu_1(x) + U_1, \quad Y_0 = \mu_0(x) + U_0, \quad (30)$$

where  $\mu_1()$  and  $\mu_0()$  are unknown functions,  $1(A)$  is the indicator function that takes value one if  $A$  occurs and value zero otherwise, and  $(U_1, U_0, V)$  have mean zero and are independent of  $Z$  (given  $X = x$ ). The condition  $V|Z \sim \text{Uniform}[0, 1]$  can be obtained as a normalization, provided that  $V$  has an absolutely continuous distribution. Equation (29) implies  $p(Z) = \Pr(W = 1|Z)$ , the propensity score, which is in this setting a function of  $Z$ . For expositional simplicity, we will assume that  $Z$  includes all elements of  $X$ .

Conditional on  $X = x$ , the marginal treatment effect at level  $v$  is defined as

$$\tau_{\text{MTE}}(v|x) = E[Y_1 - Y_0|X = x, V = v].$$

The parameter  $\tau_{\text{MTE}}(v|x)$  is understood as the treatment effect for an individual who is at the margin of being treated when  $p(Z) = v$  and  $X = x$ . From a conceptual perspective, the MTE is useful because other treatment and policy effects can be recovered as weighted integrals of the MTE, where the weighting scheme depends on the parameter of interest (e.g.,  $\tau_{\text{ATE}}$ ,  $\tau_{\text{ATET}}$  or  $\tau_{\text{LATE}}$ ). From a practical perspective, the MTE can be used to conduct

certain counterfactual policy evaluations, sometimes going beyond what can be learned from the more standard treatment effect parameters in the IV literature, provided additional assumptions hold. See Heckman and Vytlacil (1999, 2005) for more details and discussion.

The key identifying assumptions for the MTE and related parameters are that, conditional on  $X = x$ ,  $Z$  is nondegenerate and  $(U_1, U_0, V) \perp\!\!\!\perp Z$ , together with the common support condition  $0 < p(Z) < 1$ . Define the local instrumental variable (LIV) parameter as

$$\tau_{\text{LIV}}(p|x) = \frac{\partial}{\partial p} E[Y|X = x, p(Z) = p].$$

The parameter  $\tau_{\text{LIV}}$  can be seen a limiting version of LATE. For a continuous  $p(Z)$ , let

$$\Delta(p, h, x) = \frac{E[Y|X = x, p(Z) = p + h] - E[Y|X = x, p(Z) = p - h]}{E[W|X = x, p(Z) = p + h] - E[W|X = x, p(Z) = p - h]}, \quad (31)$$

where  $E[W|X = x, p(Z) = p + h] - E[W|X = x, p(Z) = p - h] = 2h$ . By the LATE result, it follows that

$$\Delta(p, h, x) = E[Y_1 - Y_0|X = x, p - h < V \leq p + h]. \quad (32)$$

Consider any  $x$  in the support of  $X$ , and any limit point  $p$  of the support of  $p(Z)|X = x$ . Taking limits as  $h \rightarrow 0$ , on the right hand sides of Equations (31) and (32),

$$\tau_{\text{MTE}}(p|x) = \tau_{\text{LIV}}(p|x).$$

As a result, MTE is nonparametrically identified by the LIV parameter.

In practice, because of the difficulties associated with nonparametric estimations of functions, researchers often employ parametric forms  $E[Y|X = x, p(Z) = p] = m(x, p, \theta)$  for the estimation of MTE and functionals thereof. Then, the MTE is estimated as  $\hat{\tau}_{\text{MTE}}(p|x) = \partial m(x, p, \hat{\theta}) / \partial p$  where  $\hat{\theta}$  is an estimator of  $\theta$ , and  $p(Z)$  is estimated in a first-step.

An alternative approach to LIV estimation of MTE is given by a control function representation of the estimand (see Wooldridge, 2015, for a review of control function methods in econometrics). This produces a two-step estimator of MTE, but this time with two regression functions,  $E[Y|X = x, p(Z) = p, W = 1]$  and  $E[Y|X = x, p(Z) = p, W = 0]$ , estimated separately (both depending on the first-step estimate of  $p(Z)$ ). For recent examples and further developments, including the study of cases with discrete instruments see Kline and Walters (2016) and Brinch et al. (2017).

## 7. The Regression Discontinuity Design

The regression discontinuity (RD) design allows researchers to learn about causal effects in settings where the treatment is not explicitly randomized and unobservables cannot be ruled out. In its most basic form, this design is applicable when each unit receives a score, also called running variable or index, and only units whose score is above a known cutoff point are assigned to treatment status, while the rest are assigned to control status. Examples include anti-poverty social programs assigned on a need basis after ranking units based on a poverty index, and educational programs assigned on a performance basis after ranking units based on a test score. See Imbens and Lemieux (2008) and Lee and Lemieux (2010) for early reviews, Cattaneo and Escanciano (2017) for an edited volume with a recent overview of the literature, and Cattaneo et al. (2018a,b) for a practical introduction.

The RD treatment assignment mechanism is  $W = 1(X \geq c)$ , where  $X$  denotes the score and  $c$  denotes the cutoff point. The key idea underlying any RD design is that units near the cutoff are comparable. At the core of the design is the assumption that units are not able to precisely manipulate their score in order to systematically place themselves above or below the cutoff. The absence of such endogenous sorting into treatment and control status is a crucial identifying assumption in the RD framework.

There is a wide variety of RD designs and closely related frameworks. Sharp RD designs are settings where treatment assignment, as determined by the value of the running variable relative to the cutoff, and actual treatment status coincide. This is akin to perfect compliance in randomized experiments. Fuzzy RD designs are settings where compliance with treatment assignment, as determined by the value of the running variable, is imperfect. This is analogous to the instrumental variable setting but for units with values of the running variable close to the cutoff. Kink RD designs are sharp or fuzzy settings where the object of interest is the derivative rather than the level of the average outcome near or at the cutoff (see Card et al., 2015). Multi-Cutoff RD designs are settings where the treatment assignment rule depends on more than one cutoff point (see Cattaneo et al., 2016). Multi-Score and Geographic RD designs are settings where the treatment assignment rule depends on more

than one score variable (see Papay et al., 2011; Keele and Titiunik, 2015).

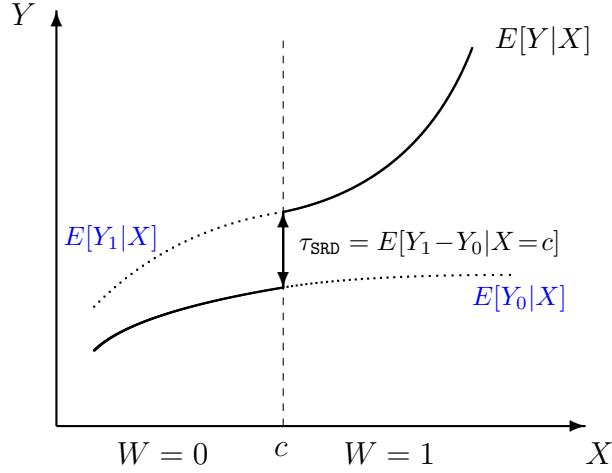


Figure 7: Regression Discontinuity Design

In the canonical sharp RD design with a continuously distributed score, the most common parameter of interest is

$$\tau_{\text{SRD}} = E[Y_1 - Y_0 | X = c] = \lim_{x \downarrow c} E[Y | X = x] - \lim_{x \uparrow c} E[Y | X = x], \quad (33)$$

which corresponds to the average causal treatment effect at the score level  $X = c$ . Figure 7 provides a graphical representation of the RD design. If the score variable satisfies  $X < c$  then units are assigned to control group ( $W = 0$ ) and  $E[Y_0 | X = x]$  is observed, while if  $X \geq c$  then units are assigned to treatment group ( $W = 1$ ) and  $E[Y_1 | X = x]$  is observed. The parameter of interest is the jump at the point  $X = c$ , where treatment assignment changes discontinuously.

Underlying the second equality in Equation (33) is the key notion of comparability between units with very similar values of the score but on opposite sides of the cutoff. This idea dates all the way back to the original landmark contribution of Thistlethwaite and Campbell (1960), but was formalized from a continuity-at-the-cutoff perspective more recently by Hahn et al. (2001). To be specific, under regularity conditions, if the regression functions  $E[Y_1 | X = x]$  and  $E[Y_0 | X = x]$  are continuous at  $x = c$ , then  $\tau_{\text{SRD}}$  is nonparametrically identified. The continuity conditions imply that the average response of units just below the

cutoff provides a close approximation to the average response that would have been observed for units just above the cutoff had they not been assigned to treatment.

### 7.1. Falsification and Validation

An important feature of all RD designs is that its key identifying assumption can be supported by a variety of empirical tests, which enhances the plausibility of the design in applications. The most popular approaches for validating RD designs in practice include: (i) testing for balance on pre-intervention covariates among units near the cutoff (Lee, 2008), (ii) testing for continuity of the score’s density function near the cutoff (McCrary, 2008; Cattaneo et al., 2017a), and (iii) graphing an estimate of  $E[Y|X = x]$  at the cutoff and away from the cutoff (Calonico et al., 2015). Other falsification and validation methods are also used in practice, though these are usually tailored to special cases. See Cattaneo et al. (2018a, Section 6) for a practical introduction to all these methods.

### 7.2. Estimation and Inference

The continuity assumptions on conditional expectations at the cutoff, which necessarily rely on the score variable being continuously distributed near the cutoff, imply that extrapolation is unavoidable in canonical RD designs. Observations near the cutoff are used to learn about the fundamentally unobserved features of the conditional expectations  $E[Y_1|X = c]$  and  $E[Y_0|X = c]$  at the cutoff. Local polynomial regression methods are the most common estimation approach in this context, as they provide a good compromise between flexibility and practicality.

The core idea underlying local polynomial estimation and inference involves three simple steps: (i) localize the sample near the cutoff by discarding observations with scores far from the cutoff, (ii) employ weighted least-squares to approximate the regression functions on either side of the cutoff separately, and (iii) predict/extrapolate the value of the regression functions at the cutoff. More formally, assuming a random sample of size  $n$ , the standard local-linear RD estimator  $\hat{\tau}_{\text{SRD}}$  is obtained by the local (to the cutoff  $c$ ) weighted linear least

squares regression:

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\tau}_{\text{SRD}} \\ \hat{\beta} \end{bmatrix} = \arg \min_{\beta_0, \tau, \beta_1, \beta_2} \sum_{i=1}^n (Y_i - \beta_0 - W_i \tau - (X_i - c)\beta_1 - (X_i - c)W_i \beta_2)^2 K\left(\frac{X_i - \bar{x}}{h}\right),$$

where  $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)'$ ,  $K(\cdot)$  denotes a compact-supported (kernel) weighting function, and  $h$  denotes the bandwidth around the cutoff  $c$ . Thus, this is a local regression that employs only observations  $X_i \in [c - h, c + h]$  with weights according to the kernel function; the two more common weighting schemes are equal weights (Uniform Kernel) and linear decreasing weights (Triangular Kernel). The coefficient  $\hat{\beta}_0$  is an estimator of  $E[Y_0|X = c]$  and can be used to assess the economic significance of the RD treatment effect estimate  $\hat{\tau}_{\text{SRD}}$ . The coefficients  $\hat{\beta}$  capture the non-constant, linear relationship between the outcome and score variables on either side of the cutoff.

Localization near the cutoff via the choice of bandwidth  $h$  is crucial and should be implemented in a systematic and objective way; ad hoc bandwidth selection methods can hamper empirical work using RD designs. Principled methods for bandwidth selection have been recently developed (see Cattaneo and Vazquez-Bare, 2016, for an overview). Imbens and Kalyanaraman (2012) develop mean-squared-error (MSE) optimal bandwidth selection and RD point estimation, while Calonico et al. (2014) propose a robust bias correction approach leading to valid inference methods based on the MSE-optimal bandwidth and RD point estimator. Using higher order distributional approximations, Calonico et al. (2018a,b) show that robust bias corrected inference is demonstrably better than other alternatives, and optimal in some cases. They also develop coverage optimal bandwidth selection methods to improve RD inference in applications. Calonico et al. (2018c) study inclusion of pre-intervention covariates using local polynomial methods in RD designs. A practical introduction to these methods is given by Cattaneo et al. (2018a,b), where additional references can be found.

While the continuity-based approach to RD designs is the most popular in practice, researchers often explicitly or implicitly rely on a local randomization interpretation of RD designs when it comes to the heuristic conceptualization of RD empirical methods and results (see, e.g., Lee, 2008; Lee and Lemieux, 2010). The two methodological frameworks are,

however, substantively distinct because they require different assumptions in general (Sekhon and Titiunik, 2016, 2017). At an intuitive level, the local randomization RD approach states that, for units with score sufficiently close to the cutoff, the assignment mechanism resembles that in a randomized controlled trial. However, because the RD treatment assignment rule  $W = 1(X \geq c)$  implies a fundamental lack of common support among control and treatment units, the local randomization RD framework requires additional assumptions for identification, estimation, inference, and falsification.

Cattaneo et al. (2015) introduce a Fisherian model for inference in RD designs using a local randomization assumption. The key assumption in this model is the existence of a neighborhood around the cutoff where two conditions hold: (i) the treatment assignment mechanism is known (e.g., complete randomization), and (ii) the non-random potential outcomes are a function of the score only through treatment status (i.e., an exclusion restriction). Under these assumptions, they develop finite-sample exact randomization inference methods for the sharp null hypothesis of no treatment effect, construct confidence intervals under a given treatment effect model, and discuss several falsification methods. They also propose data-driven methods for neighborhood selection, based on the local randomization assumptions.

Cattaneo et al. (2017c) extend the basic Fisherian RD randomization inference framework allowing for parametric adjustments of potential outcomes as functions of the score. This approach resembles the continuity-based approach and basic parametric regression adjustment methods. Cattaneo et al. (2017c) also discuss other approaches for RD analysis based on the analysis of experiments under a local randomization assumption, and compare these methods to the continuity-based approach with local polynomial fitting. Relevant to the comparison between RD approaches is that the parameter of interest differs depending on the approach. In the continuity-based approach,  $\tau_{\text{SRD}}$  is the only nonparametrically identified estimand. In the local randomization approach, other parameters are identifiable and may be of equal or greater interest. In particular, a parameter of interest in some settings is

$$\tau_{\text{LR}}(h) = E[Y_1 - Y_0 | c - h \leq X \leq c + h].$$



$\tau_{\text{LR}}$  is the average treatment effect for units with score falling within the local randomization window  $[c - h, c + h]$ , for some choice of window length  $2h > 0$ .

### 7.3. Extensions, Modifications and Recent Topics

Another situation where the distinction between parameters and methodological approaches is important is when the score variable is discrete. In this case,  $\tau_{\text{SRD}}$  is not identified nonparametrically, which renders the basic continuity-based approach for identification inapplicable. Local polynomial fits can still be used if the score variable takes on many discrete values because this estimation method takes the number of mass points as the effective sample size. The discrete nature of the score, however, changes the interpretation of the canonical RD parameter,  $\tau_{\text{SRD}}$ . In this setting, estimation and inference employs a fixed region/bandwidth around the cutoff, so identification relies on a local parametric extrapolation (see Lee and Card, 2008).

Alternatively, when the score variable  $X$  is discrete (and more so when it has only a few mass points), it may be more natural to consider an alternative RD parameter :

$$\tau_{\text{DS}} = E[Y_1|X = x_+] - E[Y_0|X = x_-],$$

where  $x_+$  and  $x_-$  denote the closest mass points above and below the cutoff  $c$ , respectively (with either  $x_+ = c$  or  $x_- = c$ ). Under a local randomization assumption,  $\tau_{\text{DS}}$  is nonparametrically identified, and estimation and inference can be conducted using any of the methods discussed previously. Whether or not  $\tau_{\text{DS}}$  is of interest will necessarily depend on each specific application, but one advantage of focusing on this parameter is that no parametric extrapolation is needed, because the object of interest is no longer the fundamentally (nonparametrically) unidentified parameter  $\tau_{\text{SRD}} = E[Y_1|X = c] - E[Y_0|X = c]$ . See Cattaneo et al. (2018b, Section 3) for further discussion, illustrations, and references.

While the RD design has recently become one of the leading program evaluation methods, there are still several important methodological challenges when it comes to its implementation and interpretation. To close this section, we briefly mention two of these outstanding research avenues. First, a crucial open question concerns the extrapolation of RD treatment

effects. This design often offers very credible and robust results for the local subpopulation of units whose scores are near or at the cutoff, but is not necessarily informative about the effects at score values far from the cutoff. In other words, while RD treatment effects are regarded as having high internal validity, their external validity is usually suspect. The lack of external validity is, of course, a major cause of concern from a program evaluation and policy prescription perspective. While there is some very recent work that addresses the issue of extrapolation of RD treatment effects (e.g., Angrist and Rokkanen, 2015; Dong and Lewbel, 2015; Bertanha and Imbens, 2017; Cattaneo et al., 2017b), more work is certainly needed, covering both methodological and empirical issues.

A second important open area of research concerns the formalization and methodological analysis of recently developed RD-like research designs. One prominent example is the recent literature on bunching and density discontinuities (Kleven, 2016; Jales and Yu, 2017), where the objects of interest are related to discontinuities and other sharp changes in a probability density function. Another example is the dynamic RD design (e.g., Cellini et al., 2010), where different units cross the cutoff point at different times and therefore other parameters of interest as well as identifying assumptions are considered. These and other RD-like designs have several common features, and can be analyzed employing ideas and tools from the classical RD literature. So far, most the work in this area has been primarily empirical but contains interesting methodological ideas that should now be formalized and analyzed in a principled way. Important issues remain unresolved, ranging from core identification questions to case-specific empirical implementation methods.

## **8. The Road Ahead**

We have provided an overview of some of the most common econometric methods for program evaluation, covering ideas and tools related to randomized experiments, selection-on-observables, difference-in-differences and synthetic controls, instrumental variables, and regression discontinuity designs. We also discussed on-going developments as well as some open questions/specific problems related to these policy evaluation methods.

Our review is, of course, far from exhaustive in terms of both depth within the topics covered and scope for program evaluation methodology. This large literature continues to evolve and adapt to new empirical challenges, and hence much more methodological work is needed to address the wide range of old and new open problems. We finish by offering a succinct account of some of these problems.

An important recent development that has had a profound impact in the program evaluation literature is the arrival of new data environments (Mullainathan and Spiess, 2017). The availability of big data has generated the need for methods able to cope with data sets that are either too large or too complex to be analyzed using standard econometric methods. Of particular importance is the role of administrative records and of large data sets collected by automated systems. These are, in some cases, relatively new source of information and pose challenges in terms of identification, estimation and inference. Model selection, shrinkage and empirical Bayes approaches are particular useful in this context (e.g., Efron, 2012; Abadie and Kasy, 2017), though these methods have not yet been fully incorporated into the program evaluation toolkit. Much of the current research in this area develops machine learning methods to estimate heterogeneous treatment effects in contexts with many covariates (see, e.g., Athey and Imbens, 2016; Wager and Athey, 2017; Taddy et al., 2016).

Also potentiated by the rise of new big, complex data, is the very recent work on networks, spillovers, social interactions, and interference (Graham, 2015; Graham and de Paula, 2018). While certainly of great importance for policy, these research areas are still evolving and not yet fully incorporated into the program evaluation literature. Bringing developments in these relatively new areas to the analysis and interpretation of policy and treatment effects is important to improve policy prescriptions.

Finally, because of space limitations, there are important topics not covered in this review. Among them are mediation analysis (VanderWeele, 2015; Lok, 2016), dynamic treatment effects and duration models (Abbring and Van den Berg, 2003; Abbring and Heckman, 2007), bounds and partial identification methods (Manski, 2008; Tamer, 2010), and optimal design of policies (Hirano and Porter, 2009; Kitagawa and Tetenov, 2018). These are also

important areas of active research in the econometrics of program evaluation.

## References

- Abadie, A. (2002). Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, 97(457):284–292.
- Abadie, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263.
- Abadie, A. (2005a). Causal inference. In Kempf-Leonard, K., editor, *Encyclopedia of Social Measurement*, vol. 1, pages 259–266. Academic.
- Abadie, A. (2005b). Semiparametric difference-in-differences estimators. *Review of Economic Studies*, 72(1):1–19.
- Abadie, A., Angrist, J., and 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. W., and Wooldridge, J. M. (2017a). Sampling-based *vs.* design-based uncertainty in regression analysis. Working Paper.
- Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2017b). When should you adjust standard errors for clustering? Working Paper.
- Abadie, A., Chingos, M. M., and West, M. R. (2017c). Endogenous stratification in randomized experiments. Working Paper.
- Abadie, A. and Dermisi, S. (2008). Is terrorism eroding agglomeration economies in central business districts? Lessons from the office real estate market in downtown Chicago. *Journal of Urban Economics*, 64(2):451–463.
- Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

- Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, 59(2):495–510.
- Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. *American Economic Review*, 93(1):113–132.
- Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267.
- Abadie, A. and Imbens, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11.
- Abadie, A. and Imbens, G. W. (2012). A martingale representation for matching estimators. *Journal of the American Statistical Association*, 107(498):833–843.
- Abadie, A. and Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, 84(2):781–807.
- Abadie, A. and Kasy, M. (2017). The risk of machine learning. Working Paper.
- Abadie, A. and L’Hour, J. (2017). A penalized synthetic control estimator for disaggregated data. Working Paper.
- Abbring, J. H. and Heckman, J. J. (2007). Econometric evaluation of social programs, part iii: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, vol. VI, pages 5145–5303. Elsevier Science B.V.
- Abbring, J. H. and Van den Berg, G. J. (2003). The nonparametric identification of treatment effects in duration models. *Econometrica*, 71(5):1491–1517.
- Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1):151–184.

- Amjad, M. J., Shah, D., and Shen, D. (2017). Robust synthetic control. Working Paper.
- Angrist, J. D. (1990). Lifetime earnings and the Vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review*, 80(3):313–336.
- Angrist, J. D. and Fernandez-Val, I. (2013). ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In Acemoglu, D., Arellano, M., and Dekel, E., editors, *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress, Volume III: Econometrics*. Econometric Society Monographs.
- Angrist, J. D., Graddy, K., and Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *Review of Economic Studies*, 67(3):499–527.
- Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.
- Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.
- Angrist, J. D. and Pischke, J. S. (2008). *Mostly Harmless Econometrics: An Empiricist’s Companion*. Princeton University Press.
- Angrist, J. D. and Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30.
- Angrist, J. D. and Pischke, J. S. (2014). *Mastering ’Metrics: The Path from Cause to Effect*. Princeton University Press.
- Angrist, J. D. and Rokkanen, M. (2015). Wanna get away? Regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344.

- Arcidiacono, P. and Ellickson, P. B. (2011). Practical methods for estimation of dynamic discrete choice models. *Annual Review of Economics*, 3(1):363–394.
- Arellano, M. (2003). *Panel Data Econometrics*. Advanced texts in econometrics. Oxford University Press.
- Ashenfelter, O. (1978). Estimating the effect of training programs on earnings. *Review of Economics and Statistics*, 60(1):47–57.
- Ashenfelter, O. and Card, D. (1985). Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics*, pages 648–660.
- Athey, S., Bayatiz, M., Doudchenko, N., Imbens, G., and Khosravik, K. (2017). Matrix completion methods for causal panel data models. Working Paper.
- Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.
- Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.
- Athey, S. and Imbens, G. W. (2017a). Econometrics of randomized experiments. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Economic Field Experiments*, vol. 1, pages 73–140. North-Holland.
- Athey, S. and Imbens, G. W. (2017b). Machine learning methods for estimating heterogeneous causal effects. Working paper.
- Athey, S. and Imbens, G. W. (2017c). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, 31(2):3–32.
- Athey, S., Imbens, G. W., and Wager, S. (2016). Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*.



- Bakshy, E., Eckles, D., and Bernstein, M. S. (2014). Designing and deploying online field experiments. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 283–292, New York, NY, USA. ACM.
- Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.
- Banerjee, A. V., Chassang, S., and Snowberg, E. (2017). Decision theoretic approaches to experiment design and external validity. In Banerjee, A. V. and Duflo, E., editors, *Handbook of Economic Field Experiments*, vol. 1, pages 141–174. North-Holland.
- Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972.
- Behrman, J. R., Parker, S. W., and Todd, P. E. (2011). Do conditional cash transfers for schooling generate lasting benefits? A five-year followup of PROGRESA/Oportunidades. *Journal of Human Resources*, 46(1):93–122.
- Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2017). Program evaluation with high-dimensional data. *Econometrica*, 85(1):233–298.
- Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81(2):608–650.
- Berry, S. and Haile, P. (2016). Identification in differentiated products markets. *Annual Review of Economics*, 8(1):27–52.
- Bertanha, M. and Imbens, G. W. (2017). External validity in fuzzy regression discontinuity designs. National Bureau of Economic Research, working paper 20773.
- Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates?\*. *Quarterly Journal of Economics*, 119(1):249–275.

- Björklund, A. and Moffitt, R. (1987). The estimation of wage gains and welfare gains in self-selection models. *Review of Economics and Statistics*, 69(1):42–49.
- Blundell, R. and Costa Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 44(3):565–640.
- Bothwell, L. E., Greene, J. A., Podolsky, S. H., and Jones, D. S. (2016). Assessing the gold standard — Lessons from the history of RCTs. *New England Journal of Medicine*, 374(22):2175–2181.
- Brinch, C. N., Mogstad, M., and Wiswall, M. (2017). Beyond LATE with a discrete instrument. *Journal of Political Economic*, 125(4):985–1039.
- Bugni, F., Canay, I., and Shaikh, A. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, *forthcoming*.
- Busso, M., DiNardo, J., and McCrary, J. (2014). New evidence on the finite sample properties of propensity score reweighting and matching estimators. *Review of Economics and Statistics*, 96(5):885–897.
- Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018a). Coverage error optimal confidence intervals for regression discontinuity designs. Working paper.
- Calonico, S., Cattaneo, M. D., and Farrell, M. H. (2018b). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, *forthcoming*.
- Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2018c). Regression discontinuity designs using covariates. Working paper.
- Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326.

- Calonico, S., Cattaneo, M. D., and Titiunik, R. (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association*, 110(512):1753–1769.
- Card, D. (1990). The impact of the Mariel Boatlift on the Miami labor market. *ILR Review*, 43(2):245–257.
- Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4):772–793.
- Card, D., Lee, D. S., Pei, Z., and Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, 83(6):2453–2483.
- Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, 2:11–20.
- Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, 155(2):138–154.
- Cattaneo, M. D. and Escanciano, J. C. (2017). *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing.
- Cattaneo, M. D. and Farrell, M. H. (2011). Efficient estimation of the dose-response function under ignorability using subclassification on the covariates. In Drukker, D., editor, *Missing-Data Methods: Cross-sectional Methods and Applications (Advances in Econometrics, vol. 27)*, pages 93–127. Emerald Group Publishing.
- Cattaneo, M. D., Frandsen, B., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. Senate. *Journal of Causal Inference*, 3(1):1–24.
- Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2018a). *A Practical Introduction to Regression Discontinuity Designs: Part I*. Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press, *forthcoming*.

- Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2018b). *A Practical Introduction to Regression Discontinuity Designs: Part II*. Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press, *in preparation*.
- Cattaneo, M. D. and Jansson, M. (2018). Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency. *Econometrica*, *forthcoming*.
- Cattaneo, M. D., Jansson, M., and Ma, X. (2017a). Two-step estimation and inference with possibly many included covariates. Working paper.
- Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018c). Alternative asymptotics and the partially linear model with many regressors. *Econometric Theory*, *forthcoming*.
- Cattaneo, M. D., Jansson, M., and Newey, W. K. (2018d). Inference in linear regression models with many covariates and heteroskedasticity. *Journal of the American Statistical Association*, *forthcoming*.
- Cattaneo, M. D., Keele, L., Titiunik, R., and Vazquez-Bare, G. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *Journal of Politics*, 78(4):1229–1248.
- Cattaneo, M. D., Keele, L., Titiunik, R., and Vazquez-Bare, G. (2017b). Extrapolating treatment effects in multi-cutoff regression discontinuity designs. Working paper, University of Michigan.
- Cattaneo, M. D., Titiunik, R., and Vazquez-Bare, G. (2017c). Comparing inference approaches for RD designs: A reexamination of the effect of Head Start on child mortality. *Journal of Policy Analysis and Management*, 36(3):643–681.
- Cattaneo, M. D. and Vazquez-Bare, G. (2016). The choice of neighborhood in regression discontinuity designs. *Observational Studies*, 2:134–146.
- Cellini, S. R., Ferreira, F., and Rothstein, J. (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *Quarterly Journal of Economics*, 125(1):215–261.

- Chalakov, K. and White, H. (2011). An extended class of instrumental variables for the estimation of causal effects. *Canadian Journal of Economics/Revue canadienne d'économique*, 44(1):1–51.
- Chernozhukov, V., Escanciano, J. C., Ichimura, H., and Newey, W. K. (2016). Locally robust semiparametric estimation. Working paper.
- Chernozhukov, V., Fernandez-Val, I., and Melly, B. (2013). Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268.
- Chernozhukov, V. and Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261.
- Chernozhukov, V. and Hansen, C. (2013). Quantile models with endogeneity. *Annual Review of Economics*, 5(1):57–81.
- Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24(2):295–313.
- Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, 48:424–455.
- Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062.
- DiNardo, J., Fortin, N. M., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica*, 64(5):1001–1044.
- DiNardo, J. and Lee, D. S. (2011). Program evaluation and research designs. In Ashenfelter, A. and Card, D., editors, *Handbook of Labor Economics*, vol. 4A, chapter 5, pages 463–536. Elsevier Science B.V.

- Ding, P. (2017). A paradox from randomization-based causal inference. *Statistical Science*, 32(3):331–345.
- Dong, Y. and Lewbel, A. (2015). Identifying the effect of changing the policy threshold in regression discontinuity models. *Review of Economics and Statistics*, 97(5):1081–1092.
- Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Working Paper 22791, National Bureau of Economic Research.
- Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.
- Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, 19(4):676–685.
- Fan, Y. and Park, S. S. (2010). Sharp bounds on the distribution of treatment effects and their statistical inference. *Econometric Theory*, 26(3):931–951.
- Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23.
- Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J. P., Allen, H., Baicker, K., and Oregon Health Study Group (2012). The Oregon health insurance experiment: Evidence from the first year. *Quarterly Journal of Economics*, 127(3):1057–1106.
- Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276.
- Firpo, S. and Pinto, C. (2016). Identification and estimation of distributional impacts of interventions using changes in inequality measures. *Journal of Applied Econometrics*, 31(3):457–486.

- Firpo, S. and Ridder, G. (2008). Bounds on functionals of the distribution of treatment effects. Working Paper.
- Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd, Edinburgh.
- Foster, J. and Shorrocks, A. F. (1988). Poverty orderings. *Econometrica*, 56(1):173–177.
- Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, 139(1):35–75.
- Frölich, M. and Melly, B. (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics*, 31(3):346–357.
- Graham, B. S. (2015). Methods of identification in social networks. *Annual Review of Economics*, 7(1):465–485.
- Graham, B. S. and de Paula, A. (2018). *The Econometrics of Social and Economic Networks*. Elsevier, *forthcoming*.
- Gruber, J. (1994). The incidence of mandated maternity benefits. *American Economic Review*, 84(3):622–641.
- Hahn, J. and Shi, R. (2017). Synthetic control and inference. Working Paper.
- Hahn, J., Todd, P., and van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.
- Heckman, J. (1990). Varieties of selection bias. *American Economic Review*, 80(2):313–318.
- Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *Review of Economic Studies*, 65(2):261–294.
- Heckman, J. J. and Robb, R. (1985). *Alternative methods for evaluating the impact of interventions*, page 156–246. Econometric Society Monographs. Cambridge University Press.

- Heckman, J. J., Urzua, S., and Vytlačil, E. J. (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics*, 88(3):389–432.
- Heckman, J. J. and Vytlačil, E. J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceedings of the national Academy of Sciences*, 96(8):4730–4734.
- Heckman, J. J. and Vytlačil, E. J. (2005). Structural equations, treatment effects and econometric policy evaluation. *Econometrica*, 73(3):669–738.
- Heckman, J. J. and Vytlačil, E. J. (2007). Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In Heckman, J. and Leamer, E., editors, *Handbook of Econometrics*, vol. VI, pages 4780–4874. Elsevier Science B.V.
- Hernán, M. A. and Robins, J. M. (2018). *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.
- Hirano, K. and Imbens, G. W. (2004). The propensity score with continuous treatments. In A., G. and Meng, X.-L., editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 73–84. Wiley, New York.
- Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.
- Hirano, K. and Porter, J. R. (2009). Asymptotics for statistical treatment rules. *Econometrica*, 77(5):1683–1701.
- Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.
- Hong, H. and Nekipelov, D. (2010). Semiparametric efficiency in nonlinear LATE models. *Quantitative Economics*, 1(2):279–304.



- Hsiao, C., Steve Ching, H., and Ki Wan, S. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of Hong Kong with mainland China. *Journal of Applied Econometrics*, 27(5):705–740.
- Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.
- Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710.
- Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132.
- Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29.
- Imbens, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48:399–423.
- Imbens, G. W. (2014). Instrumental variables: An econometrician’s perspective. *Statistical Science*, 29(3):323–358.
- Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.
- Imbens, G. W. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, 79(3):933–959.
- Imbens, G. W. and Kolesár, M. (2016). Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4):701–712.
- Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2):615–635.

- Imbens, G. W., Newey, W. K., and Ridder, G. (2006). Mean-squared-error calculations for average treatment effects. Working Paper.
- Imbens, G. W. and Rubin, D. B. (1997). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies*, 64(4):555–574.
- Imbens, G. W. and Rubin, D. B. (2015). Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.
- Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.
- Jales, H. and Yu, Z. (2017). Identification and estimation using a density discontinuity approach. In Cattaneo, M. D. and Escanciano, J. C., editors, *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*, pages 29–72. Emerald Group Publishing.
- Kang, J. D. and Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, pages 523–539.
- Keele, L. J. and Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, 23(1):127–155.
- Kitagawa, T. (2015). A test for instrument validity. *Econometrica*, 83(5):2043–2063.
- Kitagawa, T. and Tetenov, A. (2018). Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, *forthcoming*.
- Kleven, H. J. (2016). Bunching. *Annual Review of Economics*, 8:435–464.
- Kline, P. and Walters, C. (2016). Evaluating public programs with close substitutes: The case of Head Start. *Quarterly Journal of Economics*, 131(4):1795–1848.
- Koenker, R. W. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):33–50.

- Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. *Econometric evaluation of labour market policies*, pages 43–58.
- Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, 142(2):675–697.
- Lee, D. S. and Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, 142(2):655–674.
- Lee, D. S. and Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2):281–355.
- Lee, M.-J. (2016). *Matching, Regression Discontinuity, Difference in Differences, and Beyond*. Oxford University Press.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc.
- Lok, J. J. (2016). Defining and estimating causal direct and indirect effects when setting the mediator to specific values is not feasible. *Statistics in Medicine*, 35(22):4008–4020.
- Lok, J. J., Gill, R., Van Der Vaart, A., and Robins, J. (2004). Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models. *Statistica Neerlandica*, 58(3):271–295.
- Manski, C. F. (1988). *Analog Estimation Methods in Econometrics*. Chapman & Hall, New York.
- Manski, C. F. (1990). Nonparametric bounds on treatment effects. *American Economic Review*, 80(2):319–323.
- Manski, C. F. (2008). *Identification for Prediction and Decision*. Harvard University Press.

- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, 142(2):698–714.
- Morgan, S. L. and Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2nd edition.
- Mullainathan, S. and Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.
- Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 65(2):331–366.
- Newhouse, J. P. (1996). *Free for All? Lessons from the RAND Health Insurance Experiment*. Harvard University Press, Cambridge, MA.
- Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. Section 9. *Statistical Science*, 5:465–480.
- Ogburn, E. L., Rotnitzky, A., and Robins, J. M. (2015). Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396.
- Papay, J. P., Willett, J. B., and Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161(2):203–207.
- Pearl, J. (1993). Comment: Graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.
- Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York.
- Richardson, T. S. and Robins, J. M. (2013). Single world intervention graphs: A primer. Working paper.

- Robins, J. M. (2000). *Marginal Structural Models versus Structural nested Models as Tools for Causal inference*, pages 95–133. Springer New York, New York, NY.
- Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.
- Rosenbaum, P. R. (2002). *Observational Studies*. Springer, New York.
- Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer, New York.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701.
- Rubin, D. B. (1977). Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1):1–26.
- Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1):34–58.
- Rubin, D. B. (1980). Discussion of “Randomization analysis of experimental data in the Fisher randomization test” by Basu. *Journal of the American Statistical Association*, 75(371):591–593.
- Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25:279–292.
- Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1).

- Sekhon, J. S. and Titiunik, R. (2016). Understanding regression discontinuity designs as observational studies. *Observational Studies*, 2:174–182.
- Sekhon, J. S. and Titiunik, R. (2017). On interpreting the regression discontinuity design as a local experiment. In Cattaneo, M. D. and Escanciano, J. C., editors, *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*, pages 1–28. Emerald Group Publishing.
- Sloczynski, T. and Wooldridge, J. M. (2017). A general double robustness result for estimating average treatment effects. *Econometric Theory*.
- Smith, J. A. and Todd, P. E. (2005). Does matching overcome LaLonde’s critique of nonexperimental estimators? *Journal of Econometrics*, 125(1):305–353.
- Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374.
- Stock, J. and Watson, M. W. (2003). *Introduction to Econometrics*. Prentice Hall, New York.
- Taddy, M., Gardner, M., Chen, L., and Draper, D. (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics*, 34(4):661–672.
- Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics*, 2(1):167–195.
- Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex-post facto experiment. *Journal of Educational Psychology*, 51(6):309–317.
- Todd, P. E. and Wolpin, K. I. (2010). Ex ante evaluations of social programs. *Annales d’Économie et de Statistique*, 91:259–286.

- Van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Science & Business Media.
- VanderWeele, T. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press, New York.
- Wager, S. and Athey, S. (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, forthcoming.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT press.
- Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, 50(2):420–445.
- Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4):1055–1065.
- Young, A. (2017). Channelling Fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. Working paper.