Journal of Economic Perspectives—Volume 31, Number 2—Spring 2017—Pages 3–32

The State of Applied Econometrics: Causality and Policy Evaluation

Susan Athey and Guido W. Imbens

The gold standard for drawing inferences about the effect of a policy is a randomized controlled experiment. However, in many cases, experiments remain difficult or impossible to implement, for financial, political, or ethical reasons, or because the population of interest is too small. For example, it would be unethical to prevent potential students from attending college in order to study the causal effect of college attendance on labor market experiences, and politically infeasible to study the effect of the minimum wage by randomly assigning minimum wage policies to states. Thus, a large share of the empirical work in economics about policy questions relies on observational data—that is, data where policies were determined in a way other than through random assignment. Drawing inferences about the causal effect of a policy from observational data is quite challenging. To understand the challenges, consider the example of the minimum wage. A naive analysis of the observational data might compare the average employment level of states with a high minimum wage to that of states with a low minimum wage. This difference is surely *not* a credible estimate of the causal effect of a higher minimum wage, defined as the change in employment that would occur if the low-wage states raised their minimum wage. For example, it might be the case that states with higher costs of living, as well as more price-insensitive consumers, choose higher levels of the minimum wage

■ *Susan Athey is Economics of Technology Professor and Guido W. Imbens is Applied Econometrics Professor and Professor of Economics, both at the Graduate School of Business, Stanford University, Stanford, California. Both authors are also Research Associates, National Bureau of Economic Research, Cambridge, Massachusetts. Their email addresses are athey@stanford.edu and imbens@stanford.edu.*

† For supplementary materials such as appendices, datasets, and author disclosure statements, see the article page at
https://doi.org/10.1257/jep.31.2.3 doi=10.1257/jep.31.2.3

compared to states with lower costs of living and more price-sensitive consumers. These factors, which may be unobserved, are said to be "confounders," meaning that they induce correlation between minimum wage policies and employment that is not indicative of what would happen if the minimum wage policy changed.

In economics, researchers use a wide variety of strategies for attempting to draw causal inference from observational data. These strategies are often referred to as *identification strategies* or *empirical strategies* (Angrist and Krueger 1999), because they are strategies for identifying the causal effect. We say, somewhat loosely, that a causal effect is identified if it can be learned when the dataset is sufficiently large. In the first main section of the paper, we review developments corresponding to several of these identification strategies: regression discontinuity, synthetic control and differences-in-differences methods, methods designed for networks settings, and methods that combine experimental and observational data. In the next main section, we discuss *supplementary analyses*, by which we mean analyses where the results are intended to convince the reader of the credibility of the primary analyses. These supplementary analyses have not always been systematically applied in the empirical literature, but we believe they will be of growing importance. We then briefly discuss some new developments in the machine learning literature, which focus on the combination of predictive methods and causal questions. We argue that machine learning methods hold great promise for improving the credibility of policy evaluation, and they can also be used to approach supplementary analyses more systematically.

Overall, this article focuses on recent developments in econometrics that may be useful for researchers interested in estimating the effect of policies on outcomes. Our choice of topics and examples does not seek to be an overall review. Instead it is selective and subjective, based on our reading and assessment of recent research.

New Developments in Program Evaluation

The econometric literature on estimating causal effects has been very active for over three decades now. Since the early 1990s, the *potential outcome* approach, sometimes referred to as the Rubin Causal Model, has gained substantial acceptance as a framework for analyzing causal problems.[1] In the potential outcome approach, there is for each unit i and each level of the treatment w, a potential outcome $Y_i(w)$, which describes the value of the outcome under treatment level w for that unit. Researchers observe which treatment a given unit received and the corresponding outcome for each unit, but because we do not observe the outcomes for other levels of the treatment that a given unit did not receive, we can never directly observe the causal effects, which is what Holland (1986) calls the "fundamental problem of causal inference." Estimates of causal effects are ultimately based on comparisons of different units with different levels of the treatment.

[1] There is a complementary approach based on graphical models (for example, Pearl 2000) that is widely used in other disciplines.

In some settings, the goal is to analyze the effect of a binary treatment, and the *unconfoundedness assumption* can be justified. This assumption requires that all "confounding factors" (that is, factors correlated with both potential outcomes and with the assignment to the treatment) are observed, which in turn implies that conditional on observed confounders, the treatment is as good as randomly assigned. Rosenbaum and Rubin (1983a) show that under this assumption, the average difference between treated and untreated groups with the same values for the confounders can be given a causal interpretation. The literature on estimating average treatment effects under unconfoundedness is very mature, with a number of competing estimators and many applications. Some estimators use matching methods (where each treated unit is compared to control units with similar covariates), some rely on reweighting observations so that the observable characteristics of the treatment and control group are similar after weighting, and some involve the propensity score (that is, the conditional probability of receiving the treatment given the covariates) (for reviews, see Imbens 2004; Abadie and Imbens 2006; Imbens and Rubin 2015; Heckman and Vytlacil 2007). Because this setting has been so well studied, we do not cover it in this article; neither do we cover the voluminous (and very influential) literature on instrumental variables.[2] Instead, we discuss issues related to a number of other identification strategies and settings.

Regression Discontinuity Designs

A regression discontinuity design enables the estimation of causal effects by exploiting discontinuities in incentives or ability to receive a discrete treatment.[3] For example, school district boundaries may imply that two children whose houses are on the same street will attend different schools, or birthdate cutoffs may limit eligibility to start kindergarten between two children born only a few days apart. Many government programs are means-tested, meaning that eligibility depends on income falling below a threshold. In these settings, it is possible to estimate the causal effect of attending a particular school or receiving a government program by comparing outcomes for children who live on either side of the boundary, or by comparing individuals on either side of an eligibility threshold.

[2]There are two recent strands of the instrumental variables literature. One focuses on heterogenous treatment effects, with a key development being the notion of the local average treatment effect (Imbens and Angrist 1994; Angrist, Imbens, and Rubin 1996). This literature has been reviewed in Imbens (2014). There is also a literature on weak instruments, focusing on settings with a possibly large number of instruments and weak correlation between the instruments and the endogenous regressor. On this topic, see Bekker (1994), Staiger and Stock (1997), and Chamberlain and Imbens (2004) for specific contributions, and Andrews and Stock (2006) for a survey. Also, we also do not discuss in detail bounds and partial identification analyses. Starting with the work by Manski (for instance, Manski 1990), these topics have received a lot of interest, with an excellent recent review in Tamer (2010).

[3]This approach has a long history, dating back to work in psychology in the 1950s by Thistlewaite and Campbell (1960), but did not become part of the mainstream economics literature until the early 2000s (with an exception being Goldberger 1972, 2008). Fairly recent reviews include Imbens and Lemieux (2008), Lee and Lemieux (2010), van der Klaauw (2008), and Skovron and Titiunik (2015).

In general, the key feature of the design is the presence of an exogenous variable, referred to as the *forcing variable*, like the student's birthday or address, where the probability of participating in the program changes discontinuously at a threshold value of the forcing variable. This design can be used to estimate causal effects under the assumption that the individuals close to the threshold but on different sides are otherwise comparable, so any difference in average outcomes between individuals just to one side or the other can be attributed to the treatment. If the jump in the conditional probability of treatment at the threshold value is from zero to one, we refer to the design as a "sharp" regression discontinuity design. In this case, a researcher can focus on the discontinuity of the conditional expectation of the outcome given the forcing variable at the threshold, interpreted as the average effect of the treatment for individuals close to the threshold. If the magnitude of the jump in probability of receiving the treatment at the threshold value is less than one, it is a "fuzzy" regression discontinuity design. For example, some means-tested government programs are also rationed, so that not all eligible people gain access. In this case, the focus is again on the discontinuity in the conditional expectation of the outcome at the threshold, but now it must be scaled by the discontinuity in the probability of receiving the treatment. The interpretation of the estimand is the average effect for "compliers" at the threshold, that is, individuals at the threshold whose treatment status would have been different had they been on the other side of the threshold (Hahn, Todd, and van der Klaauw 2001).

Let us illustrate a regression discontinuity design with data from Jacob and Lefgren (2004). They study the causal effect of attending summer school using administrative data from the Chicago Public Schools, which in 1996 instituted an accountability policy that tied summer school attendance and promotional decisions to performance on standardized tests. We use the data for 70,831 third-graders in years 1997–99. The rule was that individuals who scored below a threshold (2.75 in this case) on either reading or mathematics were required to attend summer school. Out of the 70,831 third graders, 15,846 scored below the threshold on the mathematics test, 26,833 scored below the threshold on the reading test, 12,779 scored below the threshold on both tests, and 29,900 scored below the threshold on at least one test. The outcome variable Y_i^{obs} is the math score after the summer school, normalized to have variance one. Table 1 presents some of the results. The first row presents an estimate of the effect of summer school attendance on the mathematics test, using for the forcing variable the minimum of the initial mathematics score and the initial reading score. We find that the summer school program has a substantial effect, raising the math test outcome score by 0.18 standard deviations.

Researchers who are implementing a regression discontinuity approach might usefully bear four pointers in mind. First, we recommend using *local linear* methods for the estimation process, rather than *local constant* methods that simply attempt to estimate average outcomes on either side of the boundary using a standard kernel regression. A kernel regression predicts the average outcome at a point by taking a weighted average of outcomes for nearby observations, where closer observations are weighted more highly. The problem is that when applying such a method near a

Table 1
Regression Discontinuity Designs: The Jacob–Lefgren Data

Outcome	*Sample*	*Estimator*	*Estimate*	*Standard error*	*IK Bandwidth*
Math	All	Local Linear	0.18	(0.02)	0.57
Math	Reading > 3.32	Local Linear	0.15	(0.02)	0.57
Math	Math > 3.32	Local Linear	0.17	(0.03)	0.57
Math	Math and Reading < 3.32	Local Linear	0.19	(0.02)	0.57
Math	All	Local Constant	–0.15	(0.02)	0.57

Note and Source: This table illustrates a regression discontinuity design with data from Jacob and Lefgren (2004). They study the causal effect of attending summer school, using use administrative data from the Chicago Public Schools, which in 1996 instituted an accountability policy that tied summer school attendance and promotional decisions to performance on standardized tests. We use the data for 70,831 third-graders in years 1997–99. The rule was that individuals who scored below a threshold (2.75 in this case) on either a reading or mathematics were required to attend summer school. Out of the 70,831 third graders, 15,846 scored below the threshold on the mathematics test, 26,833 scored below the threshold on the reading test, 12,779 score below the threshold on both tests, and 29,900 scored below the threshold on at least one test. The outcome variable Y_i^{obs} is the math score after the summer school, normalized to have variance one. The first row presents an estimate of the effect of summer school attendance on the mathematics test, using for the forcing variable the minimum of the initial mathematics score and the initial reading score. We find that the summer school program has a substantial effect, raising the math test outcome score by 0.18 standard deviations. Rows 2–4 in Table 1 present estimates for separate subsamples. In this case, we find relatively little evidence of heterogeneity in the estimates.

boundary, all of the observations lie on one side of the boundary, creating a bias in the estimates (Porter 2003). As an alternative Porter suggested *local linear regression*, which involves estimating linear regressions of outcomes on the forcing variable separately on the left and the right of the threshold, and then taking the difference between the predicted values at the threshold. This approach works better if the outcomes change systematically near the boundary because the model accounts for this and corrects the bias that arises due to truncating data at the boundary. The local linear estimator has substantially better finite sample properties than nonparametric methods that do not account for threshold effects, and it has become the standard in the empirical literature. For details on implementation, see Hahn, Todd, and van der Klaauw (2001), Porter (2003), and Calonico, Cattaneo, and Titiunik (2014a).[4]

A second key element in carrying out regression discontinuity analysis, given a local linear estimation method, is the choice of the bandwidth—that is, how to weight nearby versus more distant observations. Conventional methods for choosing optimal bandwidths in nonparametric regressions look for bandwidths that are optimal for estimating an entire regression function, but here the interest is solely in the value of the regression function at a particular point. The current literature

[4]There are some suggestions that using local quadratic methods may work well given the current technology for choosing bandwidths. Some empirical studies use global high-order polynomial approximations to the regression function, but Gelman and Imbens (2014) argue that such methods have poor properties.

suggests choosing the bandwidth for the local linear regression using asymptotic expansions of the estimators around small values for the bandwidth (Imbens and Kalyanaraman 2012; Calonico, Cattaneo, and Titiunik 2014a).

This example of summer school attendance also illustrates a situation in which the discontinuity involves multiple exogenous variables: in this case, students who score below a threshold on either a language or a mathematics test are required to attend summer school. Although not all the students who are required to attend summer school do so (a fuzzy regression discontinuity design), the fact that the forcing variable is a known function of two observed exogenous variables makes it possible to estimate the effect of summer school at different margins. For example, one can estimate the effect of summer school for individuals who are required to attend because of failure to pass the language test, and compare this with the estimate for those who are required because of failure to pass the mathematics test. The dependence of the threshold on multiple exogenous variables improves the ability to detect and analyze heterogeneity in the causal effects. Rows 2–4 in Table 1 present estimates for separate subsamples. In this case, we find relatively little evidence of heterogeneity in the estimates.

A third concern for regression discontinuity analysis is how to assess the validity of the assumptions required for interpreting the estimates as causal effects. We recommend carrying out supplementary analyses to assess the credibility of the design, and in particular to test for evidence of manipulation of the forcing variable, as well as to test for discontinuities in average covariate values at the threshold. We will discuss examples later.

Fourth, we recommend that researchers investigate the external validity of the regression discontinuity estimates by assessing the credibility of extrapolations to other subpopulations (Bertanha and Imbens 2014; Angrist and Rokkanen 2015; Angrist and Fernandez-Val 2010; Dong and Lewbel 2015). Again, we return to this topic later in the paper.

An interesting recent development in the area of regression discontinuity designs involves the generalization to discontinuities in derivatives, rather than levels, of conditional expectations. The basic idea is that at a threshold for the forcing variable, the slope of the outcome function (as a function of the forcing variable) changes, and the goal is to estimate this change in slope. The first discussions of these regression kink designs appear in Nielsen, Sorensen, and Taber (2010), Card, Lee, Pei, and Weber (2015), and Dong (2014). For example, in Card, Lee, Pei, and Weber (2015), the goal of the analysis is to estimate the causal effect of an increase in the unemployment benefits on the duration of unemployment spells, where earnings are the forcing variable. The analysis exploits the fact that, at the threshold, the relationship between benefit levels and the forcing variable changes. If we are willing to assume that in the absence of the kink in the benefit system, the derivative of the expected duration of unemployment would be smooth in lagged earnings, then the change in the derivative of the expected duration with respect to lagged earnings is informative about the relation between the expected duration and the benefit schedule.

Synthetic Control Methods and Difference-In-Differences

Difference-in-differences methods have been an important tool for empirical researchers since the early 1990s. These methods are typically used when some groups, like cities or states, experience a treatment, such as a policy change, while others do not. In this situation, the selection of which groups experience the treatment is not necessarily random, and outcomes are not necessarily the same across groups in the absence of the treatment. The groups are observed before and after the treatment. The challenge for causal inference is to come up with a credible estimate of what the outcomes would have been for the treatment group in the absence of the treatment. This requires estimating a (counterfactual) change over time for the treatment group if the treatment had not occurred. The assumption underlying difference-in-differences strategies is that the change in outcomes over time for the control group is informative about what the change would have been for the treatment group in the absence of the treatment. In general, this requires functional form assumptions. If researchers make a linearity assumption, they can estimate the average treatment effect as the difference between the change in average outcomes over time for the treatment group, minus the change in average outcomes over time for the control group.

Here we discuss two recent developments to the difference-in-differences approach: the synthetic control approach and the nonlinear changes-in-changes method. The synthetic control approach developed by Abadie, Diamond, and Hainmueller (2010, 2014) and Abadie and Gardeazabal (2003) is arguably the most important innovation in the policy evaluation literature in the last 15 years. This method builds on difference-in-differences estimation, but uses systematically more attractive comparisons. To gain some intuition about these methods, consider the classic difference-in-differences study by Card (1990; see also Peri and Yasenov 2015). Card is interested in the effect of the Mariel boatlift, which brought low-skilled Cuban workers to Miami. The question is how the boatlift affected the Miami labor market, and specifically the wages of low-skilled workers. He compares the change in the outcome of interest for the treatment city (Miami) to the corresponding change in a control city. He considers various possible control cities, including Houston, Petersburg, and Atlanta.

In contrast, the synthetic control approach moves away from using a single control unit or a simple average of control units, and instead uses a weighted average of the set of controls. In other words, instead of choosing between Houston, Petersburg, or Atlanta, or taking a simple average of outcomes in those cities, the synthetic control approach chooses weights for each of the three cities so that the weighted average is more similar to Miami than any single city would be. If pre-boatlift wages are higher in Houston than in Miami, but lower in Atlanta than Miami, it would make sense to compare Miami to the average of Houston and Atlanta rather than to either Houston or Atlanta. The simplicity of the idea, and the obvious improvement over the standard methods, have made this a widely used method in the short period of time since its inception.

The implementation of the synthetic control method requires a specific choice for the weights. The original paper, Abadie, Diamond, and Hainmueller (2010), uses a minimum distance approach, combined with the restriction that the resulting

weights are nonnegative and sum to one. This approach often leads to a unique set of weights. However, if a certain unit is on the extreme end of the distribution of units, then allowing for weights that sum up to a number different from one or allowing for negative weights may improve the fit. Doudchenko and Imbens (2016) explore alternative methods for calculating appropriate weights for a synthetic control approach, such as best subset regression or LASSO (the least absolute shrinkage and selection operator) and elastic nets methods, which perform better in settings with a large number of potential control units.

Functional form assumptions can play an important role in difference-in-differences methods. For example, in the extreme case with only two groups and two periods, it is not clear whether we should assume that the percentage change over time in average outcomes would have been the same in the treatment and control groups in the absence of the treatment, or whether we should assume that the level of the change over time would have been the same. In general, a treatment might affect both the mean and the variance of outcomes, and the impact of the treatment might vary across individuals.

For the case where the data includes repeated cross-sections of individuals (that is, the data include individual observations about many units within each group in two different time periods, but the individuals cannot be linked across time periods or may come from a distinct sample such as a survey), in Athey and Imbens (2006), we propose a nonlinear version of the difference-in-differences model. This approach, which we call changes-in-changes, does not rely on functional form assumptions, while still allowing the effects of time and treatment to vary systematically across individuals. For example, one can imagine a situation in which the returns to skill are increasing over time, or in which a new medical treatment holds greater benefit for sicker individuals. The distribution of outcomes that emerges from the nonlinear difference-in-differences model is of direct interest for policy implications, beyond the average effect of the treatment itself. Further, a number of authors have used this approach as a robustness check, or what we will call in the next main section a supplementary analysis, for the results from a linear model.

Estimating Average Treatment Effects in Settings with Multivalued Treatments

Much of the earlier econometric literature on treatment effects focused on the case with binary treatments, but a more recent literature discusses the issues posed by multivalued treatment, which is of great relevance as, in practice, many treatments have multiple versions. For example, a get-out-the-vote campaign (or any advertising campaign) might consider a variety of possible messages; or a firm might consider several different price levels. In the case of a binary treatment, there are a variety of methods for estimating treatment effects under the unconfoundedness assumption, which requires that the treatment assignment is as good as random conditional on covariates. One method that works well when the number of covariates is small is to model average outcomes as a function of observed covariates, and then use the model to adjust for the extent to which differences in the treatment and control group are accounted for by observables.

However, this type of modeling performs less well if there are many covariates, or if the differences between the treatment and control group in terms of covariates are large, because errors in estimating the impact of covariates lead to large biases. An alternative set of approaches relies on the concept of a *propensity score* (Rosenbaum and Rubin 1983a), which is the probability that an individual gets a treatment, conditional on the individual's observable characteristics. In environments where unconfoundedness holds, it is sufficient to control for the propensity score (a single-dimensional variable that summarizes how observables affect the treatment probability), and it is not necessary to model outcomes as a function of all observables. That is, a comparison of two people with the same propensity score, one of whom received the treatment and one who did not, should in principle adjust for confounding variables. In practice, some of the most effective causal estimation methods in nonexperimental studies using observable data appear to be those that combine some modeling of the conditional mean of outcomes (for example, using regression adjustments) with a covariate balancing method such as subclassification, matching, or weighting based on the propensity score (Imbens and Rubin 2015), making them doubly robust (Bang and Robins 2005).

Substantially less attention has been paid to extensions of these methods to the case where the treatment takes on multiple values (exceptions include Imbens 2000; Lechner 2001; Imai and Van Dyk 2004; Cattaneo 2010; Hirano and Imbens 2004; Yang et al. 2016). However, the recent literature shows that the dimension-reducing properties of a generalized version of the propensity score, and by extension the doubly robust properties, can be maintained in the multivalued treatment setting, but the role of the propensity score is subtly different, opening up the area for empirical research in this setting. Imbens (2000) introduced the concept of a generalized propensity score, which is based on an assumption of weak unconfoundedness, requiring only that the indicator for receiving a particular level of the treatment and the potential outcome for that treatment level are conditionally independent. Weak unconfoundedness implies similar dimension-reduction properties as are available in the binary treatment case. This approach can be used to develop matching or propensity score subclassification strategies (where groups of individuals whose propensity scores lie in an interval are compared as if treatment assignment was random within the band) (for example, Yang et al. 2016). The main insight is that it is not necessary to look for subsets of the covariate space where one can interpret the difference in average outcomes by all treatment levels as estimates of causal effects. Instead, subsets of the covariate space are constructed where one can estimate the marginal average outcome for a particular treatment level as the conditional average for units with that treatment level, one treatment level at a time.

Causal Effects in Networks and Social Interactions

Peer effects, and more generally causal effects of various treatments, in networks is an important area. For example, individuals in a social network may receive information, or may gain access to a product or service, and we wish to understand the impact of that treatment both on the treated individuals, but also their peers. This

area has seen much novel work in recent years, ranging from econometrics (Manski 1993) to economic theory (Jackson 2010). Here, we discuss some of the progress that has been made in econometrics. In general, this literature focuses on causal effects in settings where units, often individuals, interact in a way that violates the no-interference assumptions (more precisely, the SUTVA or Stable Unit Treatment Value Assumption as in Rosenbaum and Rubin 1983a; Imbens and Rubin 2015) that are routinely made in the treatment effects literature. In some cases, the way in which individuals interact is simply a nuisance, and the main interest continues to be on the direct causal effects of own treatments. In other cases, the magnitude of the interactions, or peer effects, is itself the subject of interest.

Networks and peer effects can operate through many scenarios, which has led to the literature becoming somewhat fractured and unwieldy. For example, there is a distinction between, on the one hand, settings where the population can be partitioned into subpopulations with all units within a subpopulation connected, as, for example, in classrooms (for example, Manski 1993; Carrell, Sacerdote, and West 2013), workers in a labor market (Crépon et al. 2013), or roommates in college (Sacerdote 2001). One can also consider settings with general networks, in which friends of friends are not necessarily friends themselves (Christakis and Fowler 2007). Another important distinction is between settings with many disconnected networks, where asymptotic arguments for consistency rely on the number of networks getting large, and settings with a single connected network. It may be reasonable in some cases to think of the links as symmetric, and in others of links operating only in one direction. Links can be binary, with links either present or not, or a network may contain links of different strengths.

A seminal paper in the econometric literature in this area focuses on Manski's linear-in-means model (Manski 1993; Bramoullé, Djebbari, and Fortin 2009; Goldsmith-Pinkham and Imbens 2013). Manski's original paper focuses on the setting where the population is partioned into groups (like classrooms), and peer effects are constant within the groups. The basic model specification is

$$Y_i = \beta_0 + \beta_{\overline{Y}} \cdot \overline{Y}_i + \beta_X' X_i + \beta_{\overline{X}}' \overline{X}_i + \beta_Z' Z_i + \varepsilon_i,$$

where i indexes the individual. Here Y_i is the outcome for individual i, say educational achievement; $\overline{Y}_i$ is the average outcome for individuals in the peer group for individual i; X_i is a set of exogenous characteristics of individual i, like prior test scores in an educational setting; $\overline{X}_i$ is the average value of the characteristics in individual i's peer group; and Z_i is a vector of group characteristics that is constant for all individuals in the same peer group, like quality of teachers in a classroom setting. Manski considers three types of peer effects that lead to correlations in outcomes between individuals. Outcomes for individuals in the same group may be correlated because of a shared environment. These effects are called correlated peer effects, and captured by the coefficient on Z_i. Next are the exogenous peer effects, captured by the coefficient on the group average $\overline{X}_i$ of the exogenous variables. The third type is the endogenous peer effect, captured by the coefficient on the group average outcomes $\overline{Y}_i$.

Manski (1993) concludes that separate identification of these three effects, even in the linear model setting with constant coefficients, relies on very strong assumptions and is unrealistic in many settings. In subsequent empirical work, researchers have often put additional structure on the effects (for example, by ruling out some of the effects) or brought in additional information (for example, by using richer network structures) to obtain identification. Graham (2008) focuses on a setting very similar to that of Manski's linear-in-means model. He considers restrictions on the within-group covariance matrix of the ε_i assuming homoskedasticity at the individual level. In that case, a key insight is that variation in group size implies restrictions on the within and between group variances that can be used to identify peer effects. Bramoullé, Djebbari, and Fortin (2009) allow for a more general network configuration than Manski, one in which friends of friends are not necessarily connected, and demonstrate the benefits of such configurations for identification of peer effects. Hudgens and Halloran (2008) start closer to the Rubin Causal Model or potential outcome setup. They focus primarily on the case with a binary treatment, and consider how the vector of treatments for the peer group affects the individual. They suggest various structures on these treatment effects that can aid in identification. Aronow and Samii (2013) allow for general networks and peer effects, investigating the identifying power from randomization of the treatments at the individual level.

Two other branches of the literature on estimation of causal effects in a context of network and peer effects are worth mentioning. One part focuses on developing models for network formation. Such approximations require the researcher to specify in what way the expanding sample would be similar to or different from the current sample, which in turn is important for deriving asymptotic approximations based on large samples. Recent examples of such work in economics include Jackson and Wolinsky (1996), Jackson (2010), Goldsmith-Pinkham and Imbens (2013), Christakis, Fowler, Imbens, and Kalyanaraman (2010), and Mele (2013). Chandrasekhar and Jackson (2016) develop a model for network formation and a corresponding central limit theorem in the presence of correlation induced by network links. Chandrasekhar (2016) surveys the general econometrics literature on network formation.

The other branch worth a mention is the use of randomization inference in the context of causal regressions involving networks, as a way of generating exact *p*-values. As an example of randomization inference, consider the null hypothesis that a treatment has no effect. Because the null of no effects is sharp (that is, if the null hypothesis is true, we know exactly what the outcomes would be in alternative treatment regimes after observing the individual in one treatment regime), it allows for the calculation of exact *p*-values. The approach works by simulating alternative (counterfactual) treatment assignment vectors and then calculating what the test statistic (for example, difference in means between treated and control units) would have been if that assignment had been the real one. This approach relies heavily on the fact that the null hypothesis is sharp, but many interesting null hypotheses are not sharp. In Athey, Eckles, and Imbens (forthcoming), we discuss a large class of

alternative null hypotheses: for example, hypotheses restricting higher order peer effects (peer effects from friends-of-friends) while allowing for the presence of peer effects from friends; hypotheses about whether a dense network can be represented by a simplified or *sparsified* set of rules; and hypotheses about whether peers are exchangeable, or whether some peers have larger or different effects. To test such hypotheses, in Athey, Eckles, and Imbens (forthcoming), we introduce the notion of an artificial experiment, in which some units have their treatment assignments held fixed, and we randomize over the remaining units. The artificial experiment starts by designating an arbitrary set of units to be focal. The test statistics considered depend only on outcomes for these focal units. Given the focal units, one derives the set of assignments that does not change the outcomes for the focal units. The exact distribution of the test statistic can then be inferred despite the original null hypothesis not being sharp. This approach allows us to test hypotheses about, for example, the effect of friends-of-friends, without making additional assumptions about the network structure and without resorting to asymptotics in the size of the network.

External Validity

Even when a causal study is done carefully, both in analysis and design, there is often little assurance that the causal effects are valid for populations or settings other than those studied. This concern has been raised particularly forcefully in experimental studies (for examples, see the discussions in Deaton 2010; Imbens 2010; Manski 2013). Some have emphasized that without internal validity, little can be learned from a study (Shadish, Cook, and Cambell 2002; Imbens 2013). However, Deaton (2010), Manski (2013), and Banerjee, Chassang, and Snowberg (2016) have argued that external validity should receive more emphasis.

In some recent work, approaches have been proposed that allow researchers to directly assess the external validity of estimators for causal effects. A leading example concerns settings with instrumental variables (for example, Angrist 2004; Angrist and Fernandez-Val 2010; Dong and Lewbel 2015; Angrist and Rokkanen 2015; Bertanha and Imbens 2014; Kowalski 2016; Brinch, Mogstad, and Wiswall 2015). An instrumental variables estimator is often interpreted as an estimator of the local average treatment effect, that is, the average effect of the treatment for individuals whose treatment status is affected by the instrument. So under what conditions can these estimates be considered representative for the entire sample? In this context, one can partition the sample into several groups, depending on the effect of the instrumental variable on the receipt of the treatment. There are two groups that are unaffected by the instrumental variable: *always-taker*s, who always receive the treatment, and *never-takers,* who never receive the treatment, no matter the value of the instrumental variable. *Compliers* are those whose treatment status is affected by the instrumental variable. In that context, Angrist (2004) suggests testing whether the difference in average outcomes for always-takers and never-takers is equal to the average effect for compliers. Bertanha and Imbens (2014) suggest testing a combination of two equalities: whether the average outcome for untreated compliers is equal to the average outcome for never-takers; and whether the average outcome

for treated compliers is equal to the average outcome for always-takers. Angrist and Fernandez-Val (2010) seek to exploit the presence of other exogenous covariates using *conditional effect ignorability*, which is that, conditional on these additional covariates, the average effect for compliers is identical to the average effect for never-takers and always-takers.

In the context of regression discontinuity designs, concerns about external validity are especially salient. In that setting, the estimates are in principle valid only for individuals with values of the forcing variable near the threshold. There have been a number of approaches to assess the plausibility of generalizing those local estimates to other parts of the population. Some of them apply to both sharp and fuzzy regression discontinuity designs, and some apply only to fuzzy designs. Some require the presence of additional exogenous covariates, and others rely only on the presence of the forcing variable. For example, Dong and Lewbel (2015) observe that in general, in regression discontinuity designs with a continuous forcing variable, one can estimate the magnitude of the discontinuity as well as the magnitude of the change in the first derivative of the regression function, or even higher-order derivatives, which allows one to extrapolate away from values of the forcing variable close to the threshold. In another approach, Angrist and Rokkanen (2015) suggest testing whether conditional on additional covariates, the correlation between the forcing variable and the outcome vanishes. Such a finding would imply that the treatment assignment can be thought of as unconfounded conditional on the additional covariates, which again allows for extrapolation away from the threshold. Finally, Bertanha and Imbens (2014) propose an approach based on a fuzzy regression discontinuity design. They suggest testing for continuity of the conditional expectation of the outcome conditional on the treatment and the forcing variable at the threshold, adjusted for differences in the covariates.

Leveraging Experiments

In some cases, we wish to exploit the benefits of the experimental results, in particular the high degree of internal validity, in combination with the external validity and precision from large-scale representative observational studies. Here we discuss three settings in which experimental studies can be leveraged in combination with observational studies to provide richer answers than either design could provide on its own. In the first example, the surrogate variables case, the primary outcome was not observed in the experiment, but an intermediate outcome was observed. In a second case, both the intermediate outcome and the primary outcome were observed. In the third case, multiple experiments bear on a common outcome. These examples do not exhaust the settings in which researchers can leverage experimental data more effectively, and more research in this area is likely to be fruitful.

In the case of surrogate variables, studied in Athey, Chetty, Imbens, and Kang (2016), the researcher uses an intermediate variable as a surrogate for the treatment variable. For example, in medical trials there is a long history of attempts to use intermediate health measures as surrogates (Prentice 1989). The key condition for an intermediate variable to be a valid surrogate is that, in the experimental sample,

conditional on the surrogate and observed covariates, the (primary) outcomes and the treatment are independent (Prentice 1989; Begg and Leung 2000; Frangakis and Rubin 2002). In medical settings, where researchers often used single surrogates, this condition was often not satisfied in settings where it could be tested. But it may be more plausible in other settings. For example, suppose an internet company is considering a change to the user experience on the company's website. It is interested in the effect of that change on the user's purchases over a year-long period. The firm carries out a randomized experiment over a month, during which it measures details concerning the customer's engagement like the number of visits, webpages visited, and the length of time spent on the various webpages. The firm may also have historical records on user characteristics, including past engagement. The combination of the pretreatment variables and the surrogates may be sufficiently rich so that, conditional on the combination, the primary outcome is independent of the treatment.

In administrative and survey research databases used in economics, a large number of intermediate variables are often recorded that lie on or close to the causal path between the treatment and the primary outcome. In such cases, it may be plausible that the full set of surrogate variables satisfies at least approximately the independence condition. In this setting, in Athey, Chetty, Imbens, and Kang (2016), we develop multiple methods for estimating the average effect. One method corresponds to estimating the relation between the outcome and the surrogates in the observational data and using that to impute the missing outcomes in the experimental sample. Another corresponds to estimating the relation between the treatment and the surrogates in the experimental sample and using that to impute the treatment indicator in the observational sample. Yet another exploits both methods, using the efficient influence function. In the same paper, we also derive the biases from violations of the surrogacy assumption.

In the second setting for leveraging experiments, studied in Athey, Chetty, and Imbens (2016), the researcher has data from a randomized experiment, in this case containing information on the treatment and the intermediate variables, as well as pretreatment variables. In an observational study, the researcher observes the same variables plus the primary outcome. One can then compare the estimates of the average effect on the intermediate outcomes based on the observational sample, after adjusting for pretreatment variables, with those from the experimental sample. The latter are known to be consistent, and so if one finds substantial and statistically significant differences, then unconfoundedness need not hold. For that case, in Athey, Chetty, and Imbens (2016), we develop methods for adjusting for selection on unobservables, exploiting the observations on the intermediate variables.

The third setting, involving the use of multiple experiments, has not received as much attention, but provides fertile ground for future work. Consider a setting in which a number of experiments were conducted that vary in terms of the population from which the sample is drawn or in the exact nature of the treatments included. The researcher may be interested in combining these experiments to obtain more efficient estimates, perhaps for predicting the effect of a treatment in another population or estimating the effect of a treatment with different characteristics. These

issues are related to external validity concerns but include more general efforts to decompose the effects from experiments into components that can inform decisions on related treatments. In the treatment effects literature, aspects of these problems have been studied in Hotz, Imbens, and Mortimer (2005), Imbens (2010), and Allcott (2015). They have also received some attention in the literature on structural modeling, where experimental data are used to anchor aspects of the structural model (for example, Todd and Wolpin 2006).

Supplementary Analyses

Primary analyses focus on point estimates of the primary estimands along with standard errors. In contrast, supplementary analyses seek to shed light on the credibility of the primary analyses. These supplementary analyses do not seek a better estimate of the effect of primary interest, nor do they (necessarily) assist in selecting among competing statistical models. Instead, the analyses exploit the fact that the assumptions behind the identification strategy often have implications for the data beyond those exploited in the primary analyses. Supplementary analyses can take on a variety of forms, and we are not aware of a comprehensive survey to date. This literature is very active, both in theoretical and empirical studies and likely to be growing in importance in the future. Here, we discuss some examples from the empirical and theoretical literatures, which we hope provide some guidance for future work.

We will discuss four forms of supplementary analysis: 1) placebo analysis, where pseudo-causal effects are estimated that are known to be equal to zero based on a priori knowledge; 2) sensitivity and robustness analyses that assess how much estimates of the primary estimands can change if we weaken the critical assumptions underlying the primary analyses; 3) identification and sensitivity analyses that highlight what features of the data identify the parameters of interest; and 4) a supplementary analysis that is specific to regression discontinuity analyses, in which the focus is on whether the density of the forcing variable is discontinuous at the threshold, which would suggest that the forcing variable is being manipulated.

Placebo Analyses

In a placebo analysis, the most widely used of the supplementary analyses, the researcher replicates the primary analysis with the outcome replaced by a pseudo-outcome that is known not to be affected by the treatment. Thus, the true value of the estimand for this pseudo-outcome is zero, and the goal of the supplementary analysis is to assess whether the adjustment methods employed in the primary analysis, when applied to the pseudo-outcome, lead to estimates that are close to zero. These are not standard specification tests that suggest alternative specifications when the null hypothesis is rejected. The implication of rejection here is that it is possible the original analysis was not credible at all.

One type of placebo test relies on treating lagged outcomes as pseudo-outcomes. Consider, for example, the dataset assembled by Imbens, Rubin, and Sacerdote

(2001), which studies participants in the Massachusetts state lottery. The treatment of interest is an indicator for winning a big prize in the lottery (with these prizes paid out over a 20-year period), with the control group consisting of individuals who won one small, one-time prize. The estimates of the average treatment effect rely on an unconfoundedness assumption, namely that the lottery prize is as good as randomly assigned after taking out associations with some pre-lottery variables: for example, these variables include six years of lagged earnings, education measures, gender, and other individual characteristics. Unconfoundedness is certainly a plausible assumption here, given that the winning lottery ticket is randomly drawn. But there is no guarantee that unconfoundedness holds. The two primary reasons are: 1) there is only a 50 percent response rate for the survey; and 2) there may be differences in the rate at which individuals buy lottery tickets. To assess unconfoundedness, it is useful to estimate the average causal effect with pre-lottery earnings as the outcome. Using the actual outcome, we estimate that winning the lottery (with on average a $20,000 yearly prize), reduces average post-lottery earnings by $5,740, with a standard error of $1,400. Using the pseudo-outcome we obtain an estimate of minus $530, with a standard error of $780. This finding, along with additional analyses, strongly suggests that nonconfoundedness holds.

However, using the same placebo analysis approach with the LaLonde (1986) data on job market training that are widely used in the econometric evaluation literature (for example, Heckman and Hotz 1989; Dehejia and Wahba 1999; Imbens 2015), the results are quite different. Imbens (2015) uses 1975 (pretreatment) earnings as the pseudo-outcome, leaving only a single pretreatment year of earnings to adjust for the substantial difference between the trainees and comparison group from the Current Population Survey. Imbens first tests whether the simple average difference in adjusted 1975 earnings is zero. Then he tests whether both the level of 1975 earnings and the indicator for positive 1975 earnings are different in the trainees and the control groups, using separate tests for individuals with zero and positive 1974 earnings. The null is clearly rejected, casting doubt on the unconfoundedness assumption.

Placebo approaches can also be used in other contexts, like regression discontinuity design. Covariates typically play only a minor role in the primary analyses there, although they can improve precision (Imbens and Lemieux 2008; Calonico, Cattaneo, and Titiunik 2014a, b). However, these exogenous covariates can play an important role in assessing the plausibility of the regression discontinuity design. According to the identification strategy, they should be uncorrelated with the treatment when the forcing variable is close to the threshold. We can test this assumption, for example by using a covariate as the pseudo-outcome in a regression discontinuity analysis. If we were to find that the conditional expectation of one of the covariates is discontinuous at the threshold, such a discontinuity might be interpreted as evidence for an unobserved confounder whose distribution changes at the boundary, one which might also be correlated with the outcome of interest. We can illustrate this application with the election data from Lee (2008), who is interested in estimating the effect of incumbency on electoral outcomes. The treatment is a Democrat winning a congressional election, and the forcing variable is the

Democratic vote share minus the Republication vote share in the current election, and so the threshold is zero. We look at an indicator for winning the next election as the outcome. As a pretreatment variable, we consider an indicator for winning the previous election to the one that defines the forcing variable. Our estimates for the actual outcome (winning the next election) are substantially larger than those for the pseudo-outcome (winning the previous election), where we cannot reject the null hypothesis that the effect on the pseudo-outcome is zero.

One final example of the use of placebo regressions is Rosenbaum (1987), who is interested in the causal effect of a binary treatment and focuses on a setting with multiple comparison groups (see also Heckman and Hotz 1989; Imbens and Rubin 2015). In Rosenbaum's case, there is no strong reason to believe that one of the comparison groups is superior to another. Rosenbaum proposes testing equality of the average outcomes in the two comparison groups after adjusting for pretreatment variables. If one finds that there are substantial differences left after such adjustments, it shows that at least one of the comparison groups is not valid, which makes the use of either of them less credible. In applications to evaluations of labor market programs, one might implement such methods by comparing a control group of individuals who are eligible but choose not to participate with another control group of individuals who are not eligible, as in Heckman and Hotz (1989). The biases from evaluations based on the first control group might correspond to differences in motivation, whereas evaluations based on the second control group could be biased because of direct associations between eligibility criteria and outcomes.

Robustness and Sensitivity

The classical frequentist statistical paradigm suggests that a researcher specifies a single statistical model, estimates this model on the data, and reports estimates and standard errors. This is of course far from common practice, as pointed out, for example, in Leamer (1978, 1983). In practice, researchers consider many specifications and perform various specification tests before settling on a preferred model. Standard practice in modern empirical work is to present in the final paper estimates of the preferred specification of the model in combination with assessments of the robustness of the findings from this preferred specification. These alternative specifications are intended to convey that the substantive results of the preferred specification are not sensitive to some of the choices in that specification, like using different functional forms of the regression function or alternative ways of controlling for differences in subpopulations.

Some recent work has sought to make these efforts at assessing robustness more systematic. In Athey and Imbens (2015), we propose one approach to this problem, which we illustrate here in the context of regression analyses, although it can also be applied to more complex nonlinear or structural models. In the regression context, suppose that the object of interest is a particular regression coefficient that has an interpretation as a causal effect. We suggest considering a set of different specifications based on splitting the sample into two subsamples, and estimating them separately. (Specifically, we suggest splitting the original sample once for each

of the elements of the original covariate vector Z_i, and splitting at a threshold that optimizes fit by minimizing the sum of squared residuals.) The original causal effect is then estimated as a weighted average of the estimates from the two split specifications. If the original model is correct, the augmented model still leads to a consistent estimator for the estimand. Notice that the focus is *not* on finding an alternative specification that may provide a better fit; rather, it is on assessing whether the estimate in the original specification is robust to a range of alternative specifications. This approach has some weaknesses. For example, adding irrelevant covariates to the procedure might decrease the standard deviation of estimates. If there are many covariates, some form of dimensionality reduction may be appropriate prior to estimating the robustness measure. Refining and improving this approach is an interesting direction for future work. For example, the theoretical literature has developed many estimators in the setting with unconfoundedness. Some rely on estimating the conditional mean, others rely on estimating the propensity score, and still others rely on matching on the covariates or the propensity score (for a review of this literature, see Imbens and Wooldridge 2009). We recommend that researchers should report estimates based on a variety of methods to assess robustness, rather than estimates based on a single preferred method.

In combination with reporting estimates based on the preferred specification, it may be useful to report ranges of estimates based on substantially weaker assumptions. For example, Rosenbaum and Rubin (1983b, see also Rosenbam 2002) suggest starting with a restrictive specification, and then assessing the changes in the estimates that result from small to modest relaxations of the key identifying assumptions such as unconfoundedness. In the context Rosenbaum and Rubin consider, that of estimating average treatment effects under selection on observables, they allow for the presence of an unobserved covariate that should have been adjusted for in order to estimate the average effect of interest. They explore how strong the correlation between this unobserved covariate and the treatment, and the correlation between the unobserved covariate and the potential outcomes, would have to be in order to substantially change the estimate for the average effect of interest. Imbens (2003) builds on the Rosenbaum and Rubin approach by developing a data-driven way to obtain a set of correlations between the unobserved covariates and treatment and outcome.

In other work along these lines, Arkhangelskiy and Drynkin (2016) study sensitivity of the estimates of the parameters of interest to misspecification of the model governing the nuisance parameters. Tamer (2010) reviews how to assess robustness based on the partial indentification or bounds literature originating with Manski (1990).

Altonji, Elder, and Taber (2008) and Oster (2015) focus on the correlation between the unobserved component in the relation between the outcome and the treatment and observed covariates, and the unobserved component in the relation between the treatment and the observed covariates. In the absence of functional form assumptions, this correlation is not identified. These papers therefore explore the sensitivity to fixed values for this correlation, ranging from the case where the correlation is zero (and the treatment is exogenous), to an upper limit chosen to match

the correlation found between the observed covariates in the two regression functions. Oster takes this further by developing estimators based on this equality. This useful approach provides the researcher with a systematic way of doing the sensitivity analyses that are routinely done in empirical work, but often in an unsystematic way.

Identification and Sensitivity

Gentzkow and Shapiro (2015) take a different approach to sensitivity. They propose a method for highlighting what statistical relationships in a dataset are most closely related to parameters of interest. Intuitively, the idea is that simple correlations between particular combinations of variables identify particular parameters. To operationalize this, they investigate, in the context of a given model, how the key parameters of interest relate to a set of summary statistics. These summary statistics would typically include easily interpretable functions of the data such as correlations between subsets of variables. Under mild conditions, the joint distribution of the model parameters and the summary statistics should be jointly normal in large samples. If the summary statistics are in fact asymptotically sufficient for the model parameters, the joint distribution of the parameter estimates and the summary statistics will be degenerate. More typically, the joint normal distribution will have a covariance matrix with full rank. For example, when estimating the average causal effect of a binary treatment under unconfoundedness, one would expect the parameter of interest to be closely related to the correlation between the outcome and the treatment, and, in addition, to the correlations between some of the additional covariates and the outcome, or to the correlations between some of those covariates and the treatment. Gentzkow and Shapiro discuss how to interpret the covariance matrix in terms of sensitivity of model parameters to model specification. More broadly, their approach is related to proposals in different settings by Conley, Hansen, and Rossi (2012) and Chetty (2009).

Supplementary Analyses in Regression Discontinuity Designs

One of the most interesting supplementary analyses is the McCrary (2008) test in regression discontinuity designs (see also Otsu, Xu, and Matsushita 2013). What makes this analysis particularly interesting is the conceptual distance between the primary analysis and the supplementary analysis. The McCrary test assesses whether there is a discontinuity in the density of the forcing variable at the threshold. In a conventional analysis, it is unusual that the marginal distribution of a variable that is assumed to be exogenous is of any interest to the researcher: often, the entire analysis is conducted conditional on such regressors. However, the identification strategy underlying regression discontinuity designs relies on the assumption that units just to the left and just to the right of the threshold are comparable. That argument is difficult to reconcile if, say, there are substantially more units just to the left than just to the right of the threshold. Again, even though such an imbalance could easily be taken into account in the estimation, in many cases where one would find such an imbalance, it would suggest that the forcing variable is not a characteristic exogenously assigned to individuals, but rather that it is being manipulated in some way.

The classic example is that of an educational regression discontinuity design where the forcing variable is a test score. If the individual grading the test is aware of the importance of exceeding the threshold, and in particular if graders know the student personally, they may assign scores differently than if they were not aware of this. If there was such manipulation of the score, there would likely be a discontinuity in the density of the forcing variable at the threshold; there would be no reason to change the grade for an individual scoring just above the threshold.

Machine Learning and Econometrics

Supervised machine learning focuses primarily on prediction problems: given a dataset with data on an outcome Y_i, which can be discrete or continuous, and some predictors X_i, the goal is to estimate a model on a subset of the data, given the values of the predictors X_i. This subset is called the *training sample,* and it is used for predicting outcomes in the remaining data, which is called the *test sample*. Note that this approach is fundamentally different from the goal of causal inference in observational studies, where we observe data on outcomes and a treatment variable, and we wish to draw inferences about potential outcomes. Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015) argue that many important policy problems are fundamentally prediction problems; see also the article by Mullainathan and Spiess in this issue. A second class of problems, *unsupervised machine learning*, focuses on methods for finding patterns in data, such as groups of similar items, like clustering images into groups, or putting text documents into groups of similar documents. The method can potentially be quite useful in applications involving text, images, or other very high-dimensional data, even though these approaches have not had too much use in the economics literature so far. For an exception, see Athey, Mobius, and Pal (2016) for an example in which unsupervised learning is used to categorize newspaper articles into topics.

An important difference between many (but not all) econometric approaches and supervised machine learning is that supervised machine learning methods typically rely on data-driven model selection, most commonly through cross-validation, and often the main focus is on prediction performance without regard to the implications for inference. For supervised learning methods, the sample is split into a training sample and a test sample, where, for example, the test sample might have 10 percent of observations.

The training sample is itself partitioned into a number of subsamples, or cross-validation samples, often 10 of them. For each subsample, the cross-validation sample m is set aside. The remainder of the training sample is used for estimation. The estimation results are then used to predict outcomes for the left-out subsample m. The final choice of the tuning parameter is the one that minimizes the sum of the squared residuals in the cross-validation samples. Ultimate model performance is assessed by calculating the mean-squared error of model predictions (that is, the sum of squared residuals) on the held-out test sample, which was not used at all

for model estimation or tuning. Predictions from these machine learning methods are not typically unbiased, and estimators may not be asymptotically normal and centered around the estimand. Indeed, the machine learning literature places little emphasis on asymptotic normality, and when theoretical properties are analyzed, they often take the forms of worst-case bounds on risk criteria. However, the fact that model performance (in the sense of predictive accuracy on a test set) can be directly measured makes it possible to compare predictive models even when their asymptotic properties are not understood. Enormous progress has been made in the machine learning literature in terms of developing models that do well (according to the stated criteria) in real-world datasets. Here, we focus primarily on problems of causal inference, showing how supervised machine learning methods improve the performance of causal analysis, particularly in cases with many covariates.

Machine Learning Methods for Average Causal Effects

In recent years, researchers have used machine learning methods to help them control in a flexible manner for a large number of covariates. Some of these methods involved adaptions of methods used for the few-covariate case: for example, use of the weighting approach in Hirano, Imbens, Ridder, and Rubin (2001) in combination with machine learning methods such as LASSO and random forests for estimating the propensity score as in McCaffrey, Ridgeway, and Morral (2004) and Wyss et al. (2014). Such methods have relatively poor properties in many cases because they do not necessarily emphasize the covariates that are important for the bias, that is, those that are correlated both with the outcomes and the treatment indicator. More promising methods would combine estimation of the association between the potential outcomes and the covariates, and of the association between the treatment indicator and the covariates. Here we discuss three approaches along these lines (see also Athey, Imbens, Pham, and Wager 2017).

First, Belloni, Chernozhukov, Fernández, and Hansen (2013) propose a double selection procedure, where they first use a LASSO regression to select covariates that are correlated with the outcome, and then again to select covariates that are correlated with the treatment. In a final ordinary least squares regression, they include the union of the two sets of covariates, improving the properties of the estimators for the average treatment effect compared to simple regularized regression of the outcome on the covariates and the treatment.

A second line of research has focused on finding weights that directly balance covariates or functions of the covariates between treatment and control groups, so that once the data has been reweighted, it mimics a randomized experiment more closely. In the literature with few covariates, this approach has been developed in Hainmueller (2012) and Graham, Pinto, and Egel (2012, 2016); for discussion of the case with many covariates, some examples include Zubizarreta (2015) and Imai and Ratkovic (2014). In Athey, Imbens, and Wager (2016), we develop an estimator that combines the balancing with regression adjustment. The idea is that, in order to predict the counterfactual outcomes that the treatment group would have had in the absence of the treatment, it is necessary to extrapolate

from control observations. By rebalancing the data, the amount of extrapolation required to account for differences between the two groups is reduced. To capture remaining differences, the regularized regression just mentioned can be used to model outcomes in the absence of the treatment. In effect, the Athey et al. estimator balances the bias coming from imbalance between the covariates in the treated subsample and the weighted control subsample, with the variance from having excessively variable weights.

A third approach builds on the semiparametric literature on influence functions. In general, van der Vaart (2000) suggests estimating the finite dimensional component as the average of the influence function, with the infinite dimensional components estimated nonparametrically. In the context of estimation of average treatment effects this leads to "doubly robust estimators" in the spirit of Robins and Rotnitzky (1995), Robins, Rotnitzky, and Zhao (1995), and van der Laan and Rubin (2006). Chernozhukov et al. (2016) propose using machine learning methods for the infinite dimensional components and incorporate sample-splitting to further improve the properties.

In all three cases, procedures for trimming the data to eliminate extreme values of the estimated propensity score (as in Crump, Hotz, Imbens, and Mitnik 2009) remain important in practice.

Machine Learning for Heterogenous Causal Effects

In many cases, a policy or treatment might have different costs and benefits if applied in different settings. Gaining insight into the nature of such heterogenous treatment effects can be useful. Moreover, in evaluating a policy or treatment, it is useful to know the applications where the benefit/cost ratios are most favorable. However, when machine learning methods are applied to estimating heterogenous treatment effects, they in effect search over many covariates and subsets of the covariate space for the best fit. As a result, such methods may lead to spurious findings of treatment effect differences. Indeed, in clinical medical trials, pre-analysis plans must be registered in advance to avoid the problem that researchers will be tempted to search among groups of the studied population to find one that seems to be affected by the treatment, and may instead end up with spurious findings. In the social sciences, the problem of searching across groups becomes more severe when there are many covariates.

One approach to this problem is to search exhaustively for treatment effect heterogeneity and then correct for issues of multiple hypothesis testing, by which we mean the problems that arise when a researcher considers a large number of statistical hypotheses, but analyzes them as if only one had been considered. This can lead to *false discovery*, because across many hypothesis tests, we expect some to be rejected even if the null hypothesis is true. To address this problem, List, Shaikh, and Xu (2016) propose to give each covariate a "low" or "high" discrete value, and then loop through the covariates, testing whether the treatment effect is different when the covariate is low versus high. Because the number of covariates may be large, standard approaches to correcting for multiple testing may severely

limit the power of a (corrected) test to find heterogeneity. List et al. propose an approach based on bootstrapping that accounts for correlation among test statistics; this approach can provide substantial improvements over standard multiple testing approaches when the covariates are highly correlated, because dividing the sample according to each of two highly correlated covariates results in substantially the same division of the data. However, this approach has the drawback that the researcher must specify in advance all of the hypotheses to be tested, along with alternative ways to discretize covariates and flexible interactions among covariates. It may not be possible to explore these combinations fully.

A different approach is to adapt machine learning methods to discover particular forms of heterogeneity by seeking to identify subgroups that have different treatment effects. One example is to examine within subgroups in cases where eligibility for a government program is determined according to criteria that can be represented in a decision tree, similar to the situation when a doctor uses a decision tree to determine whether to prescribe a drug to a patient. Another example is to examine within subgroups in cases where an algorithm uses a table to determine which type of user interface, offer, email solicitation, or ranking of search results to provide to a user. Subgroup analysis has long been used in medical studies (Foster, Taylor, and Ruberg 2011), but it is often subject to criticism due to concerns of multiple hypothesis testing (Assmann, Pocock, Enos, and Kasten 2000).

Among the more common machine learning methods, regression trees are a natural choice for partitioning into subgroups (the classic reference is Breiman, Friedman, Stone, and Olshen 1984). Consider a regression with two covariates. The value of each covariate can be split so that it is above or below a certain level. The regression tree approach would consider which covariate should be split, and at which level, so that the sum of squared residuals is minimized. With many covariates, these steps of choosing which covariate to split, and where to split it, are carried out sequentially, thus resulting in a tree format. The tree eventually results in a partition of the data into groups, defined according to values of the covariates, where each group is referred to as a leaf. In the simplest version of a regression tree, we would stop this splitting process once the reduction in the sum of squared residuals is below a certain level.

In Athey and Imbens (2016), we develop a method that we call *causal trees*, which builds on earlier work by Su et al. (2009) and Zeileis, Hothorn, and Hornik (2008). The method is based on the machine learning method of regression trees, but it uses a different criterion for building the tree: rather than focusing on improvements in mean-squared error of the prediction of outcomes, it focuses on mean-squared error of treatment effects. The method relies on sample splitting, in which half the sample is used to determine the optimal partition of the covariates space (the tree structure), while the other half is used to estimate treatment effects within the leaves. The output of the method is a treatment effect and a confidence interval for each subgroup. In Athey and Imbens (2016), we highlight the fact that the criteria used for tree construction should differ when the goal is to estimate treatment effect heterogeneity rather than heterogeneity in outcomes. After

all, the factors that affect the level of outcomes might be quite different from those that affect treatment effects. Although the sample-splitting approach may seem extreme—ultimately only half the data is used for estimating treatment effects—it has several advantages. The confidence intervals are valid no matter how many covariates are used in estimation. In addition, the researcher is free to estimate a more complex model in the second part of the data, for example, if the researcher wishes to include fixed effects in the model, or model different types of correlation in the error structure.

A disadvantage of the causal tree approach is that the estimates are not personalized for each individual; instead, all individuals assigned to a given group have the same estimate. For example, a leaf might contain all male individuals aged 60 to 70, with income above $50,000. An individual whose covariates are near the boundary, for example a 70 year-old man with income of $51,000, might have a treatment effect that is different than the average for the whole group. For the problem of more personalized prediction, Wager and Athey (2015) propose a method for estimating heterogeneous treatment effects based on random forest analysis, where the method generates many different trees and averages the result, except that the component trees are now causal trees (and in particular, each individual tree is estimated using sample splitting, where one randomly selected subsample is used to build the tree while a distinct subsample is used to estimate treatment effects in each leaf). Relative to a causal tree, which identifies a partition and estimates treatment effects within each element of the partition, the causal forest leads to estimates of causal effects that change more smoothly with covariates, and in principle every individual has a distinct estimate. Random forests are known to perform very well in practice for prediction problems, but their statistical properties were less well understood until recently. Wager and Athey show that the predictions from causal forests are asymptotically normal and centered on the true conditional average treatment effect for each individual. They also propose an estimator for the variance, so that confidence intervals can be obtained. Athey, Tibshirani, and Wager (2016) extend the approach to other models for causal effects, such as instrumental variables, or other models that can be estimated using the generalized method of moments (GMM). In each case, the goal is to estimate how a causal parameter of interest varies with covariates.

An alternative approach, closely related, is based on Bayesian Additive Regression Trees (BART) (Chipman, George, and McCulloch 2010), which is essentially a Bayesian version of random forests. Hill (2011) and Green and Kern (2012) apply these methods to estimate heterogeneous treatment effects. Large-sample properties of this method are unknown, but it appears to have good empirical performance in applications.

Other machine-based approaches, like the LASSO regression approach, have also been used in estimating heterogenous treatment effects. Imai and Ratkovic (2013) estimate a LASSO regression model with the treatment indicator interacted with covariates, and uses LASSO as a variable selection algorithm for determining which covariates are most important. In using this approach, it may be prudent

to perform some supplementary analysis to verify that the method is not overfitting; for example, one could use a sample-splitting approach, using half of the data to estimate the LASSO regression and then comparing the results to an ordinary least squares regression with the variables selected by LASSO in the other half of the data. If the results are inconsistent, it could indicate that using half the data is not good enough, or it might indicate that sample splitting is warranted to protect against overfitting or other sources of bias that arise when data-driven model selection is used.

A natural application of personalized treatment effect estimation is to estimate optimal policy functions in observational data. A literature in machine learning considers this problem (Beygelzimer and Langford 2009; Beygelzimer et al. 2011); some open questions include the statistical properties of the estimators, and the ability to obtain confidence intervals on differences between policies obtained from these methods. Recently, Athey and Wager (2017) bring in insights from semiparametric efficiency theory in econometrics to propose a new estimator for optimal policies and to analyze the properties of this estimator. Policies can be compared in terms of their "risk," which is defined as the gap between the expected outcomes using the (unknown) optimal policy and the estimated policy. Athey and Wager derive an upper bound for the risk of the policy estimated using their method and show that it is necessary to use a method that is efficient (in the econometric sense) to achieve that bound.

Conclusion

In the last few decades, economists have learned to take very seriously the old admonition from undergraduate econometrics that "correlation is not causality." We have surveyed a number of recent developments in the econometrics toolkit for addressing causality issues in the context of estimating the impact of policies. Some of these developments involve a greater sophistication in the use of methods like regression discontinuity and differences-in-differences estimation. But we have also tried to emphasize that the project of taking causality seriously often benefits from combining these tools with other approaches. Supplementary analyses can help the analyst assess the credibility of estimation and identification strategies. Machine learning methods provide important new tools to improve estimation of causal effects in high-dimensional settings, because in many cases it is important to flexibly control for a large number of covariates as part of an estimation strategy for drawing causal inferences from observational data. When causal interpretations of estimates are more plausible, and inference about causality can reduce the reliance of these estimates on modeling assumptions (like those about functional form), the credibility of policy analysis is enhanced.

■ *We are grateful for comments by the editor and coeditors.*

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 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, Alberto, Alexis Diamond, and Jens Hainmueller. 2014. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59(2): 495–510.

Abadie, Alberto, and Javier Gardeazabal. 2003. "The Economic Costs of Conflict: A Case Study of the Basque Country." *American Economic Review* 93(1): 113–32.

Abadie, Alberto, and Guido W. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects." *Econometrica* 74(1): 235–67.

Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130(3): 1117–65.

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2008. "Using Selection on Observed Variables to Assess Bias from Unobservables When Evaluating Swan–Ganz Catheterization." *American Economic Review* 98(2): 345–50.

Andrews, Donald, and James H. Stock. 2006. "Inference with Weak Instruments." Unpublished paper.

Angrist, Joshua D. 2004. "Treatment Effect Heterogeneity in Theory and Practice." *Economic Journal* 114(494): C52–83.

Angrist, Joshua, and Ivan Fernandez-Val. 2010. "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." NBER Working Paper 16566.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91(434): 444–55.

Angrist, Joshua D., and Alan B. Krueger. 1999. "Empirical Strategies in Labor Economics." In *Handbook of Labor Economics*, edited by Orley C. Ashenfelter and David Card, 1277–1366. North Holland.

Angrist, Joshua D., and Miikka Rokkanen. 2015. "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff." *Journal of the American Statistical Association* 110(512): 1331–44.

Arkhangelskiy, Dmitry, and Evgeni Drynkin. 2016. "Sensitivity to Model Specification." Unpublished paper.

Aronow, Peter M., and Cyrus Samii. 2013. "Estimating Average Causal Effects under Interference between Units." arXiv: 1305.6156v1.

Assmann, Susan F., Stuart J. Pocock, Laura E. Enos, and Linda E. Kasten. 2000. "Subgroup Analysis and Other (Mis)uses of Baseline Data in Clinical Trials." *Lancet* 355 (9209): 1064–69.

Athey, Susan, Raj Chetty, and Guido Imbens. 2016. "Combining Experimental and Observational Data: Internal and External Validity." Unpublished paper.

Athey, Susan, Raj Chetty, Guido Imbens, and Hyunseung Kang. 2016. "Estimating Treatment Effects Using Multiple Surrogates: The Role of the Surrogate Score and the Surrogate Index." arXiv: 1603.09326.

Athey, Susan, Dean Eckles, and Guido Imbens. Forthcoming. "Exact *p*-Values for Network Interference." *Journal of the American Statistical Association*.

Athey, Susan, and Guido W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74(2): 431–97.

Athey, Susan, and Guido Imbens. 2015. "A Measure of Robustness to Misspecification." *American Economic Review* 105(5): 476–80.

Athey, Susan, and Guido Imbens. 2016. "Recursive Partitioning for Estimating Heterogeneous Causal Effects." *PNAS* 113(27): 7353–60.

Athey, Susan, Guido Imbens, Thai Pham, and Stefan Wager. 2017. "Estimating Average Treatment Effects: Supplementary Analyses and Remaining Challenges." arXiv: 1702.01250.

Athey, Susan, Guido Imbens, and Stefan Wager. 2016. "Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing." arXiv: 1604.07125.

Athey, Susan, Markus Mobius, and Jeno Pal. 2016. "The Impact of Aggregators on News Consumption." Unpublished paper.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2016. "Solving Heterogeneous Estimating Equations with Gradient Forests." arXiv: 1610.01271.

Athey, Susan, and Stefan Wager. 2017. "Efficient Policy Learning." arXiv: 1702.02896.

Banerjee, Abhijit, Sylvain Chassang, and Erik Snowberg. 2016. "Decision Theoretic Approaches to Experiment Design and External Validity." NBER Working Paper 22167.

Bang, Heejung, and James M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61(4): 962–73.

Begg, Colin B., and Denis H. Y. Leung. 2000. "On the Use of Surrogate End Points in

Randomized Trials." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 163(1): 15–28.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62(3): 657–81.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernández-Val, and Chris Hansen. 2013. "Program Evaluation and Causal Inference with High-Dimensional Data." arXiv: 1311.2645.

Bertanha, Marinho, and Guido Imbens. 2014. "External Validity in Fuzzy Regression Discontinuity Designs." NBER Working Paper 20773.

Beygelzimer, Alina, and John Langford. 2009. "The Offset Tree for Learning with Partial Labels." *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 129–38.

Beygelzimer, Alina, John Langford, Lihong Li, Lev Reyzin, and Robert E. Schapire. 2011. "Contextual Bandit Algorithms with Supervised Learning Guarantees." *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 19–26.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin. 2009. "Identification of Peer Effects through Social Networks." *Journal of Econometrics* 150(1): 41–55.

Breiman, Leo, Jerome Friedman, Charles J. Stone, and Richard A. Olshen. 1984. *Classification and Regression Trees.* CRC Press.

Brinch, Christian, Magne Mogstad, and Matthew Wiswall. 2015. "Beyond LATE with a Discrete Instrument: Heterogeneity in the Quantity-Quality Interaction in Children." Unpublished paper.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014a. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82(6): 2295–2326.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014b. "Robust Data-Driven Inference in the Regression-Discontinuity Design." *Stata Journal* 14(4): 909–46.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Industrial and Labor Relations Review* 43 (2): 245–57.

Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2015. "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica* 83 (6): 2453–83.

Carrell, Scott E., Bruce I. Sacerdote, and James E. West. 2013. "From Natural Variation to Optimal Policy? The Importance of Endogenous Peer Group Formation." *Econometrica* 81(3): 855–82.

Cattaneo, Matias D. 2010. "Efficient Semiparametric Estimation of Multi-valued Treatment Effects under Ignorability." *Journal of Econometrics* 155(2): 138–54.

Chamberlain, Gary, and Guido Imbens. 2004. "Random Effects Estimators with Many Instrumental Variables." *Econometrica* 72(1): 295–306.

Chandrasekhar, Arun G. 2016. "The Econometrics of Network Formation." Chap. 13 in *The Oxford Handbook on the Economics of Networks*, edited by Yann Bramoullé, Andrea Galeotti, Brian W. Rogers. Oxford University Press.

Chandrasekhar, Arun, and Matthew Jackson. 2016. "A Network Formation Model Based on Subgraphs." arXiv: 1611.07658.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2016. "Double Machine Learning for Treatment and Causal Parameters." arXiv: 1608.00060.

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge between Structural and Reduced-Form Methods." *Annual Review of Economics* 1(1): 451–87.

Chipman, Hugh A., Edward I. George, and Robert E. McCulloch. 2010. "BART: Bayesian Additive Regression Trees." *Annals of Applied Statistics* 4(1): 266–98.

Christakis, Nicholas A., and James H. Fowler. 2007. "The Spread of Obesity in a Large Social Network over 32 Years." *New England Journal of Medicine* (357): 370–79.

Christakis, Nicholas A., James H. Fowler, Guido W. Imbens, and Karthik Kalyanaraman. 2010. "An Empirical Model for Strategic Network Formation." NBER Working Paper 16039.

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi. 2012. "Plausibly Exogenous." *Review of Economics and Statistics* 94(1): 260–72.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. 2013. "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." *Quarterly Journal of Economics* 128(2): 531–80.

Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika* 96(1): 187–99.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2): 424–55.

Dehejia, Rajeev H., and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94(448): 1053–62.

Dong, Yingying. 2014. "Jump or Kink? Identification of Binary Treatment Regression Discontinuity Design without the Discontinuity." Unpublished paper.

Dong, Yingying, and Arthur Lewbel. 2015. "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models." *Review of Economics and Statistics* 97(5): 1081–92.

Doudchenko, Nikolay, and Guido W. Imbens. 2016. "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis." arXiv: 1610.07748.

Foster, Jared C., Jeremy M. G. Taylor, and Stephen J. Ruberg. 2011. "Subgroup Identification from Randomized Clinical Data." *Statistics in Medicine* 30(24): 2867–80.

Frangakis, Constantine E., and Donald B. Rubin. 2002. "Principal Stratification in Causal Inference." *Biometrics* 58(1): 21–29.

Gelman, Andrew, and Guido Imbens. 2014. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." NBER Working Paper 20405.

Gentzkow, Matthew, and Jesse Shapiro. 2015. "Measuring the Sensitivity of Parameter Estimates to Sample Statistics." Unpublished paper.

Goldberger, Arthur S. 1972. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Institute for Research on Poverty Discussion Paper 129-72.

Goldberger, Arthur S. 2008. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." In *Advances in Econometrics, Volume 21*, edited by Tom Fomby, R. Carter Hill, Daniel L. Millimet, Jeffrey A. Smith, and Edward J. Vytlacil, 1–31. Emerald Group Publishing Limited.

Goldsmith-Pinkham, Paul, and Guido W. Imbens. 2013. "Social Networks and the Identification of Peer Effects." *Journal of Business and Economic Statistics* 31(3): 253–64.

Graham, Bryan S. 2008. "Identifying Social Interactions through Conditional Variance Restrictions." *Econometrica* 76(3): 643–60.

Graham, Bryan S., Cristine Campos de Xavier Pinto, and Daniel Egel. 2012. "Inverse Probability Tilting for Moment Condition Models with Missing Data." *Review of Economic Studies* 79(3): 1053–79.

Graham, Bryan, Christine Campos de Xavier Pinto, and Daniel Egel. 2016. "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)." *Journal of Business and Economic Statistics* 34(2): 288–301.

Green, Donald P., and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491–511.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1): 201–09.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1): 25–46.

Heckman, James J., and V. Joseph Hotz. 1989. "Choosing among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training." *Journal of the American Statistical Association* 84(408): 862–74.

Heckman, James J., and Edward Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In *Handbook of Econometrics 6B*, edited by James Heckman and Edward Leamer, 4779–4874. Elsevier.

Hill, Jennifer L. 2011. "Bayesian Nonparametric Modeling for Causal Inference." *Journal of Computational and Graphical Statistics* 20(1): 217–40.

Hirano, Keisuke. 2001. "Combining Panel Data Sets with Attrition and Refreshment Samples." *Econometrica* 69(6): 1645–59.

Hirano, Keisuke, and Guido Imbens. 2004. "The Propensity Score with Continuous Treatments." In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, edited by Andrew Gelman and Xiao-Li Meng, 73–84. Wiley.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396): 945–60.

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." *Journal of Econometrics* 125(1–2): 241–70.

Hudgens, Michael G., and M. Elizabeth Halloran. 2008. "Toward Causal Inference with Interference." *Journal of the American Statistical Association* 103(482): 832–42.

Imai, Kosuke, and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *Annals of Applied Statistics* 7(1): 443–70.

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1): 243–63.

Imai, Kosuke, and David A. van Dyk. 2004. "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99(467): 854–66.

Imbens, Guido W. 2000. "The Role of the Propensity Score in Estimating Dose–Response Functions." *Biometrika* 87(3): 706–10.

Imbens, Guido W. 2003. "Sensitivity to

Exogeneity Assumptions in Program Evaluation." *American Economic Review* 93(2): 126–32.

Imbens, Guido W. 2004. "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review." *Review of Economics and Statistics* 86(1): 4–29.

Imbens, Guido W. 2010. "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature* 48(2): 399–423.

Imbens, Guido. 2013. "Book Review Feature: Public Policy in an Uncertain World." *Economic Journal* 123(570): F401–411.

Imbens, Guido W. 2014. "Instrumental Variables: An Econometrician's Perspective." *Statistical Science* 29(3): 323–58.

Imbens, Guido W. 2015. "Matching Methods in Practice: Three Examples." *Journal of Human Resources* 50(2): 373–419.

Imbens, Guido W., and Joshua D. Angrist. 1994. "Identification and Estimation of Local Average Treatment Effects." *Econometrica* 62(2): 467–75.

Imbens, Guido W., and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79(3): 933–59.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142(2): 615–35.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, Guido W., Donald B. Rubin, and Bruce I. Sacerdote. 2001. "Estimating the Effect of Unearned Income on Labor Earnings, Savings, and Consumption: Evidence from a Survey of Lottery Players." *American Economic Review* 91(4): 778–94.

Imbens, Guido W., and Jeffrey M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47(1): 5–86.

Jackson, Matthew O. 2010. *Social and Economic Networks*. Princeton University Press.

Jackson, Matthew, and Asher Wolinsky. 1996. "A Strategic Model of Social and Economic Networks." *Journal of Economic Theory* 71(1): 44–74.

Jacob, Brian A., and Lars Lefgren. 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* 86(1): 226–44.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105(5): 491–95.

Kowalski, Amanda. 2016. "Doing More When You're Running LATE: Applying Marginal Treatment Effect Methods to Examine Treatment Effect Heterogeneity in Experiments." NBER Paper 22363.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4): 604–20.

Leamer, Edward. 1978. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley.

Leamer, Edward E. 1983. "Let's Take the Con Out of Econometrics." *American Economic Review* 73(1): 31–43.

Lechner, Michael. 2001. "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption." In *Econometric Evaluation of Labour Market Policies,* vol. 13, edited by Michael Lechner and Friedhelm Pfeiffer, 43–58. Physica-Verlag Heidelberg.

Lee, David S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142(2): 675–97.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48(2): 281–355.

List, John A., Azeem M. Shaikh, and Yang Xu. 2016. "Multiple Hypothesis Testing in Experimental Economics." NBER Paper 21875.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *American Economic Review* 80(2): 319–23.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *Review of Economic Studies* 60(3): 531–42.

Manski, Charles F. 2013. *Public Policy in an Uncertain World: Analysis and Decisions*. Harvard University Press.

McCaffrey, Daniel F., Greg Ridgeway, and Andrew R. Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *Psychological Methods* 9(4): 403–25.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2): 698–714.

Mele, Angelo. 2013. "A Structural Model of Segregation in Social Networks." Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2294957.

Nielsen, Helena Skyt, Torben Sorensen, and Christopher Taber. 2010. "Estimating the Effect of Student Aid on College Enrollment: Evidence from a Government Grant Policy Reform." *American Economic Journal: Economic Policy* 2(2): 185–215.

Oster, Emily. 2015. "Diabetes and Diet: Behavioral Response and the Value of Health." NBER Working Paper 21600.

Otsu, Taisuke, Ke-Li Xu, and Yukitoshi Matsushita. 2013. "Estimation and Inference of Discontinuity in Density." *Journal of Business and Economic Statistics* 31(4): 507–24.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Peri, Giovanni, and Vasil Yasenov. 2015. "The Labor Market Effects of a Refugee Wave: Applying the Synthetic Control Method to the Mariel Boatlift." NBER Working Paper 21801.

Porter, Jack. 2003. "Estimation in the Regression Discontinuity Model." Unpublished paper, University of Wisconsin at Madison.

Prentice, Ross L. 1989. "Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria." *Statistics in Medicine* 8 (4): 431–40.

Robins, James M., and Andrea Rotnitzky. 1995. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data." *Journal of the American Statistical Association* 90(429): 122–29.

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1995. "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data." *Journal of the American Statistical Association* 90(429): 106–21.

Rosenbaum, Paul. 1987. "The Role of a Second Control Group in an Observational Study." *Statistical Science* 2(3): 292–306.

Rosenbaum, Paul R. 2002. "Observational Studies." In *Observational Studies*, 1–17. Springer.

Rosenbaum, Paul R., and Donald B. Rubin. 1983a. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41–55.

Rosenbaum, Paul R., and Donald B. Rubin. 1983b. "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome." *Journal of the Royal Statistical Society. Series B (Methodological)* 45(2): 212–18.

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics* 116(2): 681–704.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

Skovron, Chistopher, and Rocío Titiunik. 2015. "A Practical Guide to Regression Discontinuity Designs in Political Science." Unpublished paper.

Staiger, Douglas, and James H. Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3): 557–86.

Su, Xiaogang, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li. 2009. "Subgroup Analysis via Recursive Partitioning." *Journal of Machine Learning Research* 10: 141–58.

Tamer, Elie. 2010. "Partial Identification in Econometrics." *Annual Review of Economics* 2(1): 167–95.

Thistlewaite, D., and Donald Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex-post Facto Experiment." *Journal of Educational Psychology* 51(6): 309–17.

Todd, Petra, and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review* 96 (5): 1384–1417.

van der Klaauw, Wilbert. 2008. "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics." *Labour* 22(2): 219–45.

van der Laan, Mark J., and Daniel Rubin. 2006. "Targeted Maximum Likelihood Learning." *International Journal of Biostatistics* 2(1).

van der Vaart, Aad W. 2000. *Asymptotic Statistics*. Cambridge University Press.

Wager, Stefan, and Susan Athey. 2015. "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests." arXiv:1510.04342

Wyss, Richard, Allan Ellis, Alan Brookhart, Cynthia Girman, Michele Jonsson Funk, Robert LoCasale, and Til Strümer. 2014 "The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score." *American Journal of Epidemiology* 180(6): 645–55.

Yang, Shu, Guido W. Imbens, Zhanglin Cui, Douglas E. Faries, and Zbigniew Kadziola. 2016. "Propensity Score Matching and Subclassification in Observational Studies with Multi-level Treatments." *Biometrics* 72(4): 1055–65.

Zeileis, Achim, Torsten Hothorn, and Kurt Hornik. 2008. "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics* 17(2): 492–514.

Zubizarreta, Jose R. 2015. "Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data." *Journal of the American Statistical Association* 110(511): 910–22.