CHAPTER 9

Estimating Treatment Effects: Matching Quantification to the Question

THOMAS A. LOUGHRAN AND EDWARD P. MULVEY

INTRODUCTION

In criminal justice as well as other areas, practitioners and/or policy makers often wish to know whether something "works" or is "effective." Does a certain form of family therapy reduce troubled adolescents' involvement in crime more than what would be seen if they were on probation? Does a jail diversion policy substantially increase indicators of community adjustment for mentally ill individuals who are arrested and processed under this policy? If so, by how much?

Trying to gauge the impact of programs or policies is eminently logical for several reasons. Obviously, this type of information is important from a traditional cost-benefit perspective. Knowing the overall impact of a program in terms of tangible and measurable benefits to some target group of interest is necessary to assess whether an investment in the program buys much. For instance, a drug rehabilitation program, which requires a large fixed cost of opening plus additional considerable operating expenses, should be able to show that this investment is worth it in terms of reduced drug use or criminal activity among its clients. Quantifiable estimates about the impact of policies or programs are also important in assessing the overall social benefit of particular approaches; it is often useful to know how much a recent change in policy has affected some subgroup in an unintended way. For instance, more stringent penalties for dealing crack, rather than powdered cocaine, appears to have provided only a marginal decrease in drug trafficking at the expense of considerable racial disparity in sentencing. Informed practice and policy rests on empirical quantifications of how much outcomes shift when certain approaches or policies are put into place.

These estimates of program or policy impact are rarely easy to obtain or to trust fully. It is often too easy to adopt a traditional empirical method of analysis (e.g., regression, odds ratios) to provide a metric of program effectiveness, without considering the limitations or restrictions of such approaches. Finding that two groups, on average, significantly differ statistically on some outcome after some statistical controls have been introduced, or that a certain group

A.R. Piquero and D. Weisburd (eds.), *Handbook of Quantitative Criminology*,
DOI 10.1007/978-0-387-77650-7_9, © Springer Science + Business Media, LLC 2010

affected by an intervention, is several times more likely to have a particular outcome than a non-affected comparison group, is far from definitive proof of the intervention's effectiveness. Relying on findings like these, analysts can often provide biased estimates of success or harm, estimates of outcome effects that are irrelevant for answering the appropriate policy question at hand, or, in some cases, both. Obtaining and interpreting quantitative results about outcomes in a manner which is both correct and germane, requires delicate consideration of both the question at hand and the methods used to generate the numbers considered.

There are a variety of ways to explicitly estimate treatment effects, including several that are formally outlined within this edition. This chapter does not focus on actual estimation methods per se, but rather addresses the broader question of interpreting these effects, once estimated. We attempt to provide the reader with some clarity regarding the use and interpretation of estimates of *treatment effects*, a term which we believe is often used in too generic a manner. Specifically, we focus on understanding which type of treatment effect estimate is of interest to particular types of policy problems.

This chapter is organized as follows: Section 2 presents some definitions and two important issues that have to be addressed in any effort to estimate the impact of a practice or policy. Section 3 provides some technical notation for defining several different quantities that can be calculated, each of which is often referred to as a treatment effect. Section 4 describes the issues connected with interpreting treatment effects when the researcher is capable of experimentally manipulating treatment assignment. Section 5 introduces additional considerations regarding the interpretation of treatment effects with observational data. Section 6 concludes and offers a general discussion of inferring causality for program and policy discussions.

COMMON TERMS AND CONSIDERATIONS

We should first be clear about several terms that are used throughout the ensuing discussion. A *treatment* refers to a policy or intervention, which may be administered to (or, alternatively, withheld from) an individual in a population of interest. For instance, an offender being sent to prison or entering drug rehabilitation would constitute examples of individuals "being treated," as compared to those individuals within the same population who are not sent to prison or do not enter drug rehabilitation. We denote those in the former conditions as members of a *treatment group*; those in the latter conditions are considered members of the *comparison group*. Researchers are typically interested in estimating a *treatment effect*, that is, the effect of a policy or intervention on some related outcome of interest. For instance, criminologists are typically interested in determining the effect of being sent to prison on an individual's subsequent recidivism, or the effect of a drug rehabilitation program on the likelihood that an individual will relapse into drug use.

In a broader scientific sense, researchers in these situations are interested in determining a *causal* effect of the treatment. The effect *caused* by a treatment is a comparison of an outcome a subject reveals after involvement with some treatment the subject *actually received* with the latent, and unobserved outcome the subject *would have* exhibited under the alternative treatment. This unobservable outcome for the individual under the alternative treatment is known as a *counterfactual* outcome. However, as considered in more depth below, it can often be very difficult, if not impossible, to make a strong causal inference regarding the effect of a treatment.

There are two main impediments to coming to a strong conclusion about the size of the treatment effect, namely *selection bias* and *response heterogeneity*. These two issues take a variety of forms in evaluation designs, and there are a number of different strategies to address them. Depending on how well and appropriately they are addressed, we gain more or less confidence that we are getting a valid estimate of the desired effect of a treatment.

Selection bias (or a *selection effect*) is the result of a situation in which a subset of the population is more likely to be assigned to or select into some treatment than the rest of the population. The problem is that the factors, which make this subset more likely to select into the treatment, may also be directly influencing some outcome of interest. In this case, a comparison of treatment and control groups may be inappropriate as it involves a comparison of groups which are dissimilar in important ways *prior* to entering the treatment. It is, thus, difficult to disentangle whether differences in outcomes after treatment between the treatment and control groups are caused explicitly by the treatment, or if they may be due to these preexisting differences, between groups. Such preexisting differences which mask the true causal nature of the relationship between treatment and outcome, are often referred to as *confounders*. For example, if we are trying to determine the true causal effect of being sent to prison on an individual's subsequent criminality, the problem is made difficult by the fact that those offenders who are likely to be sent to prison are inherently likely to be more criminally active after being released than those who are not sent to prison. Inherent differences between groups bias our estimate of a true causal effect to an unknown degree, and while this problem is typically well recognized, it is not always properly controlled for in many empirical designs.

There is a second issue which is less mentioned, yet, is equally problematic for assessing treatment effects for their policy relevance. *Heterogeneity in response to treatment* among different subsets of the population presents issues when both estimating the causal impact of a treatment as well as when applying the observed effects to program or policy improvements. In some instances, even if one is able to get an unbiased estimate of the effect of a treatment on some subset of the population, it may not be generalizable to the population at large. In other words, the same treatment may have dramatically different effects on different segments of the population; that is, different subgroups may *respond* differently to exposure to the treatment. If this is the case, any attempt to generalize the effects learned from one particular segment to another may rely on extrapolation and lead to nonsensical, or even worse, harmful conclusions. For instance, suppose we find evidence that one-on-one sessions with a school psychologist greatly benefit children who follow a high, chronic trajectory of conduct problems by reducing this behavior. It is likely the case that providing the same sessions to those children who do not exhibit such intense conduct problems may not only fail to benefit these children, but may actually work to impede some of their basic school outcomes by removing them from the classroom environment.

These issues are certainly not new. Researchers address them regularly, using a variety of quantitative methods ranging from randomized treatment assignment, sophisticated statistical and econometric methods, to testing the representativeness of samples to ensure generalizability. We mention them here as a backdrop for our discussion of the more common ways that treatment effects are estimated in the literature. Carefully considering how selection bias and heterogeneity of response have been addressed in a particular approach is essential for applying findings about treatment effects to policy questions appropriately.

DEFINING TREATMENT EFFECTS

There are several specific quantities which fall under the broader rubric of "treatment effect." We use the *potential-outcomes* conceptual framework, first introduced by Neyman (1923) and later developed by Rubin (1974, 1977, and 1978), to define these different quantities.[1] Due to scope and spatial constraints, we restrict our discussion exclusively to *point-treatment* studies, where the impact of a single treatment on some later outcome in an individual is considered, and there are no multiple time-dependant states of treatment status. In other words, we only consider such cases where there is neither treatment exposure nor covariate confounding which is time-dependent.[2] This seems to us to be the most common situation encountered when practitioners and policy makers question whether there is sufficient evidence to believe that something "works."

Bodies of literature regarding specific treatments are often analyzed and summarized using *meta-analyses*, which treat separate studies of treatments as data points in an analysis of overall effectiveness of a particular approach. We do not address this technique here, as there is a separate chapter in this volume specifically devoted to this topic. The discussion here, however, raises questions that might be considered when conducting these sorts of summary analyses, especially with regard to the types of treatment effects considered to be equivalent in the consideration of a set of investigations.

Some Basic Notation

Let y_1 denote some outcome with treatment and y_0 denote the outcome without treatment. Notice that since an individual cannot be in both states of treatment simultaneously, we *cannot* observe both y_1 and y_0. Let $Z = 1$ for each subject if that subject has been treated, and $Z = 0$ if the subject is assigned to the control condition. To measure the *effect of the treatment*, we are interested in the difference in the outcomes in each treatment condition, $y_1 - y_0$.

Notice that this treatment effect cannot be calculated for individuals, since any single individual's counterfactual outcome is never actually observed. In other words, we are only able to observe an individual's outcome either under treatment or under control, but never both simultaneously. Therefore, when evaluating the effect of a treatment, we are limited to estimating the *average* effect across either some predefined population or subpopulation.

Population Average Treatment Effect

We define the *population average treatment effect (PATE)* as

$$E(y_1 - y_0)$$

The *PATE* (or, alternatively, the *average treatment effect*, or *ATE*) is the expected effect of treatment on a randomly drawn person from the population. This answers the question, if

[1] For a thorough overview of the framework of the Rubin Causal Model, see Holland (1986).

[2] For a discussion of these more complicated situations involving time-dependency in treatment and covariate confounding, see Robins et al. (2000) and Robins et al. (1999).

we could randomly choose someone from the entire population and treat them, what would we expect the effect to be? Conceptually, this quantity may appear to be highly attractive to researchers, since it provides an estimate of the effect of a policy across the entire population. This same reason, however, also undermines the utility of this estimate in many situations. The *PATE* is sometimes not practical because as it averages precisely across the entire population in making an estimate, it may include units never eligible for particular types of treatment in the first place.

Consider an example where we are interested in estimating the effectiveness of a technical job training program on increasing the future labor market success of those who participate as measured by future wages. There are some individuals in the population such as those with a 4-year college degree, for whom this particular type of training might have zero, or perhaps even a negative impact on their future wages. Many of these individuals would likely never enter such a training program in the first place. Therefore, the *PATE*, which takes the entire population into consideration, is often not particularly useful in evaluating the effectiveness of a specific treatment, since it gives an estimate of a treatment effect that will never happen.

The *PATE* is well suited to provide information about the possible impact of universal prevention efforts. Because the PATE estimates the effect of a treatment on a population, it provides an estimate of the overall societal cost or benefit connected with particular policies or practices. For example, exposure to lead has been shown to have an effect on the development of early antisocial behavior (Needleman et al. 1996; Nevin 2000), and subsequent estimates have been made regarding the amount of delinquency in a particular locale that can be reasonably attributed to this exposure (Wright et al. 2008). These estimates of the *PATE* indicate the expected impact across the entire population of children from lead exposure. Thus, they give policy makers an indication of the potential general payoff from a broadly implemented strategy and the concomitant benefit and cost deriving to each taxpayer.

Average Effect of Treatment on the Treated

A second quantity of interest is known as the *average effect of the treatment on the treated (ATT)*, which is defined as

$$E(y_1 - y_0 | Z = 1)$$

The *ATT* is the mean effect for those who *actually participated* in the treatment or program. It answers the question, for a random individual who *actually received the treatment*, what would we expect this effect to be? The *PATE* and *ATT* generally differ, but are equivalent in some special cases discussed below. Return to the job training example from above. A program evaluator charged with understanding the effectiveness of such a program would likely wish to know how the treatment affected those who actually underwent the training, with no concern for those who had no intention of doing so. Thus, the *ATT* is a much more appealing quantity in this case and in many others. Ridgeway (2006) provides an empirical example of *ATT* in his analysis of the role of racial biases in traffic stop outcomes, by using propensity score *weighting* (McCaffrey et al. 2004). For the purposes of estimation, Ridgeway notes that in general, propensity scores methods (chapter in edition; see also Rosenbaum and Rubin 1983) are closely linked to the potential-outcomes framework.

Within-Group Treatment Effect

It certain instances, we might be interested in how some treatment affects a certain subgroup of the population. Rather than just being concerned with how a program has an effect across the whole population or those who were enrolled in the program, we might want to know how the program affects a particular group of policy interest (e.g., females vs. males). Both the *PATE* and *ATT* can be redefined to generalize to a specific subset of the population. We may do this by simply expanding these definitions to condition on some covariate x or vector of covariates **x**.

The *PATE* conditional on **x** is simply

$$E(y_1 - y_0|\mathbf{x})$$

This quantity answers the question, if we randomly treat an individual from the entire population with characteristic **x**, what would we expect the effect to be? For example, consider the example of estimating the effects of lead exposure presented above. The *PATE* obtained above regarding the effects of lead exposure may be valuable for estimating the overall impact of a program limiting the use of certain materials in a community. By examining the *PATE* for those individuals living in housing projects versus the rest of the community, however, we would get a picture of the relative effect of focusing efforts at control in just those settings. The *PATE* estimates conditioned on certain individual level characteristics of a sample can thus provide guidance about ways to maximize the impact of an intervention or policy change.

Similarly, the *ATT* conditional on **x** is

$$E(y_1 - y_0|\mathbf{x}, \quad Z = 1)$$

This quantity answers the question, for a random individual with characteristic **x** who *actually received the treatment*, what would we expect this effect to be? By suitably choosing **x**, we may define the *PATE* and *ATT* for various subsets of the population. This may or may not be important depending on the policy question one is interested in addressing. For example, we might be interested in how incarceration specifically affects women as opposed to men, or how a drug rehabilitation program helps chronic users as opposed to less serious users. We explore differences in these quantities and their potential utility in more detail below.

SUMMARY

Clearly there are multiple ways to represent a causal effect of treatment. The different quantities used, however, are not all the same. Each answers specific, and sometimes very different questions and they may give very different estimates under different conditions. Given this, it becomes critical to recognize precisely which of these quantities is most appropriate to the policy question of interest in any inquiry. Confusion about which of these estimates is most appropriate, or worse, generic confusion in estimation, can lead to incorrect and potentially harmful policy conclusions.

These considerations drive the remainder of the discussion in this chapter. We continue by considering the two distinct situations under which researchers must estimate and interpret effects of treatment. In one, researchers are allowed to randomly decide who receives a treatment, and, in the other, individuals are allowed to self-select into treatment through their own means or via some other, nonrandom mechanism. We consider the latter case, which is much more prevalent in the social sciences, in deeper detail.

INTERPRETING TREATMENT EFFECTS UNDER RANDOMIZATION

Causal Effects in Randomized Experiments

The universally accepted best method for determining a treatment effect is to conduct a controlled application of a treatment to two groups, composed at random, who either receive or do not receive the treatment. This strategy of *randomized controlled trials*, or *RCTs*, has been widely accepted in medicine for a long time and is currently coming more into vogue as the sine qua non of scientific proof in the social sciences (Weisburd et al. 2001; see also evidence from the Campbell Collaboration, http://www.campbellcollaboration.org/index.asp). In this approach, the researcher is conducting an *experiment* and has the power to control the assignment of the treatments to subjects. Thus, the treatment can be *randomly* assigned. This is critically important, then, as the treatment assignment will not be correlated in any way with the outcome in question. This tends to produce relatively comparable, or *balanced* treatment groups in large experiments, meaning that the treatment and control groups are similar in terms of the distribution of both *observable* and *unobservable* individual characteristics, which are set prior to the treatment. Thus, randomization essentially rules out the presence of selection biases confounding the estimated effect. Although we draw caution to some important considerations below, we cannot stress strongly enough that randomization with a sufficiently large sample is *the absolute best* condition under which to determine causal effects of treatment.

In the case of treatment randomization, the PATE can be thought of in the potential-outcomes framework as a simple difference in means of treatment and comparisons:

$$\text{PATE} = \bar{y}_1 - \bar{y}_0$$

Under randomization, this simple difference in means yields an *unbiased* and *consistent* estimate of the causal treatment effect – a very powerful result. Also note that randomization implies that all individuals within the population are equally likely to be treated (say, with probability = .5), and thus, the *PATE* and the *ATT* will be equivalent.

Furthermore, randomization can be applied usefully in more basic regression contexts as well. Consider the following simple regression model

$$y_i = \alpha + \beta Z_i + u_i$$

If the treatment is randomly assigned, then it is independent of all other factors, or formally, $Cov(Z, u) = 0$, meaning that, an Ordinary Least Squares (OLS) estimate of β in the above equation will too yield an unbiased and consistent estimate of the *PATE*.

While randomization is unequivocally the de facto gold standard in evaluation research for the reason described above, we still must be careful about simply generalizing this effect to a wide population. It may be that not all members of the population were equally eligible to be included in the treatment allocation process. Thus, simply because we employ randomization of treatment, this does not mean we have a universally generalizable treatment effect. If the group eligible to be selected for treatment is not the same as the population in general, then any attempt to extend the results to the population at large is potentially problematic. This is sometimes referred to as a lack of *external validity* (Shadish et al. 2001), and it can present serious complications to the conclusions, even if there is pure randomization of treatment assignment.

This issue becomes relevant when implementing "evidence based practices" in different locales. Oftentimes, researchers conduct controlled studies of a treatment program in several locales, documenting impressive treatment effects for a particular intervention approach. This treatment effect may or may not be found, however, when the intervention is then applied to locales that differ substantially from the demonstration locales (e.g., in the demographics of the adolescents/families served, the history of the individuals referred to the intervention). The generation of a treatment effect is certainly a different process than demonstrating the applicability of that effect to a broadly defined group of individuals such as serious adolescent offenders in general. While concerns such as these are most common, there are other factors with randomization, which need be considered. Heckman and Smith (1995) offer a thoughtful and detailed discussion of some other important considerations and limitations of experiments and randomization in the social sciences.

Another important caveat of randomization, which poses a potential threat to the external validity of measured treatment effects, deals with *treatment compliance*, or more precisely, the lack thereof. It cannot simply be assumed that all individuals who are randomly assigned to receive some treatment or control actually do receive it, or, in other words, *comply* with their assigned treatment status.[3] Noncompliance can occur in two general forms. First, an individual who is randomized to the treatment group can end up not receiving the treatment. This is known as *treatment dilution*. Second, some subjects who are assigned to the control group could still potentially end up receiving treatment. In this case, we have *treatment migration*. Both of these occurrences are potentially problematic, as they could possibly reintroduce selection biases into the interpretation of a randomized experiment if not properly considered.

One initial strategy to deal with noncompliance (which at first seems rather intuitively appealing) is to simply ignore those who did not properly comply, and estimate treatment effects from only those who did properly comply. However, the exclusion of noncompliant individuals will likely not be random, if the reason for noncompliance is correlated with the outcome in question. The result is a nonrandom compliance group that no longer balances overall pretreatment characteristics, and thus, might not reveal the realistic effect of the treatment. For example, consider a hypothetical therapy intervention for terminally ill cancer patients aimed at prolonging their survival time. Some of the most seriouslyill individuals might die prior to receiving treatment. This is an extreme example of noncompliance; however, by excluding these individuals from the analysis, we are, in all likelihood, limiting

[3] The concept of noncompliance should be thought of in a purely statistical interpretation in this case, where it literally means not adhering to the randomly assigned treatment. Often, particularly in some clinical applications, the term noncompliant can have a negative connotation, as in lack of willingness to accept a helpful therapy. Noncompliance can occur for a variety of reasons, not simply lack of insight or stubbornness, and should therefore not be thought to indicate anything negative about an individual when used in this context. For instance, if a chronic headache sufferer is randomized into the treatment group testing the effectiveness of a new drug and chooses to not take the drug for the simple reason that there is no pain at the time of treatment, then this individual is a non-complier as defined here.

the most severelyill individuals from the treatment group but not the control group (in which case, we are still capable of observing their outcome, survival time). As such, we gain a biased estimate of the treatment effect.

An alternative is to conduct an analysis based on *Intention to Treat (ITT)*, which requires the inclusion of individuals as members of the treatment group to which they are assigned, regardless of compliance or noncompliance. In contrast to the above strategy of excluding noncompliers, this approach may, at first glance, seem counterintuitive; for example, including people who refuse to be treated in a treatment group does not seem very logical. However, the ITT framework is actually critical in that it preserves randomization, which is in direct contrast to the approach of excluding those who do not comply. Also, it can be interpreted as a much more practical assessment of external validity. Assessing the size of a treatment effect in an ITT framework provides a picture of what the impact of a treatment is likely to be when the approach is implemented in the real world. It builds in the attrition that a particular intervention might precipitate into the estimate, thus, in many ways allowing for both the potentially positive and negative aspects of an intervention to be considered. Since it retains randomization, it tests whether the overall effect of the intervention is likely to be positive when it is implemented with individuals like those who are enrolled in the study.

However, there are some limitations to an ITT analysis; most notably, it may reveal a more conservative, or muted estimated treatment effect because of dilution from noncompliance. This is problematic, particularly if one wishes to test for the inequality of different treatment effects, since it will be harder to reject a null hypothesis of no difference. Also, if there is an unusually high degree of noncompliance, then it can become complicated to interpret the estimated treatment effect.

Local Average Treatment Effect

Although randomization is a powerful method to deal with selection bias, opportunities for pure experimental randomization are rare in many of the social sciences. There are, however, situations where randomization of treatment may naturally occur through some other mechanism for some segment of the population, and thus, preserve many of the benefits of experiments. Such situations, known as *natural experiments*, may be exploited in order to circumvent some issues of noncompliance as well as provide useful estimates of treatment effects when explicit randomization of treatment assignment is generally unfeasible or impossible.

In such cases, instrumental variable (IV) methods may be employed. IV methodology relies on pockets of exogenous variation which affect treatment assignment in ways otherwise unrelated to the outcome of interest. Moreover, it yields another sound solution to the problem of noncompliance. For instance, Angrist (2006) shows how IV methodology can be applied to the Minneapolis domestic violence experiment (Sherman and Berk 1984; Berk and Sherman 1988) to counteract the problems of noncompliance in arrests for domestic disputes.

When using IV to estimate causal treatment effects, however, it is important to note that we identify yet another quantity. Recall that, with IV, the source of exogenous variation critical to identification of the treatment effect, centers on individuals being induced to receive treatment based on their having different values of the instrumental variable. As such, the treatment group can be broken into two distinct categories as defined by Angrist et al. (1996): *always-takers*, or those who select into the treatment regardless of their individual value of the instrument, and *compliers*, or those individuals who would not have selected into the treatment had the instrument not induced them to do so.[4]

With the assumption of monotincity, (that is, some binary instrumental variable, D, makes everyone either more or less likely to select into some binary treatment, Z, but not both), Imbens and Angrist (1994) define what they call the *local average treatment effect (LATE)*, which can be written as:

$$\frac{E(y|D=1) - E(y|D=0)}{P(Z=1|D=1) - P(Z=1|D=0)}$$

Notice that numerator in this expression, which is commonly known as the *Wald estimator*, is the difference in outcome between the two groups split by the binary instrument, D. However, since not everyone receiving a value of $D = 1$ will also select into treatment (i.e., have $Z = 1$), this difference in outcomes for the two groups must be adjusted by the probability that individuals in each group select into treatment. As such, having a value $D = 1$ must induce some additional subset of individuals to select into treatment, $Z = 1$, than would not select in if $D = 0$, or else there will be no identification, as the denominator of this quantity would be equal to zero.

However, this identification comes at a price. Since it is identified only off of the compliers, without the very strong assumption of treatment effect homogeneity, the *LATE* will not equal either the *PATE* or *ATT*. Instead, the *LATE* estimator answers the question, what is the expected effect for the compliers, or those who received treatment explicitly because the instrument induced them to do so but otherwise would not have selected into the treatment?

For example Angrist (1990) employs IV methodology in an attempt to determine the causal effects of military service on future civilian labor market outcomes by using the Vietnam draft lottery number IV. The treatment group, that is, those who joined the military, included two distinct groups of individuals: those who would have joined the military regardless (i.e., the always-takers), and those who only joined because their low draft number induced them to do so (i.e., the compliers). Allowing that the draft number met the assumptions to be used as an instrument (i.e., it predicts military service but is otherwise random), Angrist is able to estimate the causal effect of military service on future wages in the civilian labor market, but is only able to generalize with regard to those who only joined the military *because they had a low draft number*. There is no way to estimate the effect for those who joined the military regardless of draft number (which likely includes those with the worst potential civilian labor market outcomes, an interesting subgroup in this context) This inherent inability of the *LATE* estimator to generalize to a broader set of the population is one of its main criticisms, and it has been argued that natural experiments and IV methods are best suited when the response to treatment is *homogeneous* (Heckman 1997; Angrist 2004).

There are, however, some instances where the exogenous variation exploited by an IV answers *precisely* the policy question of interest, and hence, the *LATE* estimator may actually be the most preferred method. Consider an example of two adjacent areas, say neighboring

[4] Angrist, Imbens and Rubin also defines a group known as *never-takers*, or those who, regardless of the instrument, never select into treatment, and therefore are not included as part of the treatment group. Furthermore, the assumption of monotonicity effectively rules out the existence of *defiers*, or those who would have selected into treatment had the instrument made them less likely to do so, but not selected into treatment had their value of the instrument made them more likely to do so.

counties, which are relatively homogenous, except that one adopts a policy lowering the legal limit of blood alcohol content for driving in hopes of deterring Driving Under the Influence (DUI)'s. Clearly, those who are the worst offenders of driving under the influence of alcohol (i.e., those who choose to drive no matter how much they have to drink) will not be deterred by such a change. Yet, such a policy is not aimed at curbing these individuals anyway, but rather, those at the margin, who may be deterred from driving given the lowered legal limit. It is the impact on these individuals that the *LATE* estimator provides, and it can be argued that in this context, the *LATE* estimate is most relevant for evaluating the effectiveness of the policy. The *ATT*, often attractive in other instances, would not be suitable here, since it would consider all drunk drivers, including those who would never respond to such a policy shift.

Summary

True randomization of treatment assignment is undoubtedly the best way to evaluate causal effects of treatments. It should be noted that no statistical or econometric methodology, no matter how sophisticated, can do better in terms of estimating treatment effects, and oftentimes yield substantial bias in the presence of specification errors (see LaLonde 1986, for an assessment of several nonexperimental estimators). Furthermore, even in situations where pure randomization is infeasible or impossible, pockets of exogenous variation, such as abrupt law changes or variation in policy in otherwise similar areas may present the possibility of a convincing natural experiment. This high regard for randomization or the estimation of exogenous effects should be tempered, however, by skepticism about whether the requirements of randomization are really met in any investigation. Randomization is often difficult to generate convincingly, and the methods used in a study to achieve randomization matter in terms of the estimate of the treatment effects. For example, Weisburd et al. (2001) compared studies involving randomized and nonrandomized interventions in the National Institute of Justice Studies, and found study design to influence its conclusions substantially. Without a demonstration of effective randomization, the treatment effects generated in such studies should be examined closely for the possible impact of factors such as compliance on the estimates obtained.

In many situations, randomization is simply impossible. Interventions are rarely withheld at random, experiences do not occur at random to individuals, and policies are not applied to only a randomly selected subset of population. We now consider such situations, which are of much more importance to criminologists.

TREATMENT EFFECTS IN OBSERVATIONAL DATA

In most evaluation research in social science, in general, and in criminology in particular, randomization of treatment assignment is neither possible (e.g., incarceration) nor ethical (e.g., drug use), meaning that people may *self-select* into the treatment they wish to receive. Consequently, one must rely on data from an *observational study* in order to study these treatments and their effects. An observational study is an empirical analysis of treatments or policies and the effects that they cause, which differs from an experiment in that the investigator has no control over the treatment assignments (Rosenbaum 2002). As mentioned earlier, these situations involve consideration of selection effects, in which differences in outcomes between treatment and control groups may be due to preexisting differences of those who are and are not selected for treatment as opposed to an actual causal effect of the treatment. There are multiple methods that can be employed to correct for selection bias and these methods work with varying degrees of success to eliminate these effects.

As mentioned above, though, this is not the end of the story. Even in instances where we may reasonably believe we have eliminated all or most of the bias due to selection for some subgroup of the population, it still may be the case that the treatment effect we estimate is not necessarily generalizable to the population at large due to population heterogeneity in the response to treatment. While this residual bias or apparent lack of external validity may appear to be problematic on the surface, we develop an argument below as to why this is not always the case. Instead, we see this variability as an analytic opportunity. In particular, we posit several situations in which the *global treatment effect*, that is, an effect describing the treatment effect for the entire population (or treatment population) in question, is actually *less* interesting and relevant for substantive policy applications than the variable effects that might exist within subpopulations of the group examined. Before we get to an illustration of this latter point, however, we will examine the issues related to constructing valid estimates of treatment effects when confronted with observational data.

Common Support

When attempting to estimate treatment effects with observational data, it is critical to determine whether proper counterfactual outcomes can be found within the control group data. In an ideal situation, a counterfactual outcome can be generated in the control group data to directly assess the impact of the treatment in question. In many situations, though, this is impossible. Consider the example of testing the discharge practices used in a forensic psychiatric hospital. There are simply some individuals who will never be released because of the bizarre and violent nature of their crimes and their lack of responsiveness to medications. Finding the counterfactual of what would happen if a person with such a severe criminal and mental health profile were released is simply not possible.

In the absence of such counterfactual outcomes, depending on the method of estimation, we may be either estimating a biased effect, a fractional effect, which is not necessarily generalizable to the population of interest, or an extrapolated effect, which has no meaning at all. It is, therefore, necessary to assess the overall impact of these situations that cannot be represented in any control condition to determine how applicable any observed treatment effect might be to the policy in question. To do this, we may examine the conditional treatment and control group distributions over some covariate x or covariates $\mathbf{x}$ to make sure there is sufficient overlap, or *common support*. If there is a portion of treated individuals whose values of x (or vector $\mathbf{x}$) are so different that they are unlike any control individuals, then we must be cautious in how we interpret our estimated effect.

An illustration of how some situations indicating different levels of common support helps us to see the importance of this issue for later interpretation of any treatment effect generated on nonrandomized treatment and control groups. Suppose we examine frequency histograms of treatment and control group membership, based on some covariate x, which is important in treatment selection, as is done in Fig. 9.1. Note that the dimension of x may

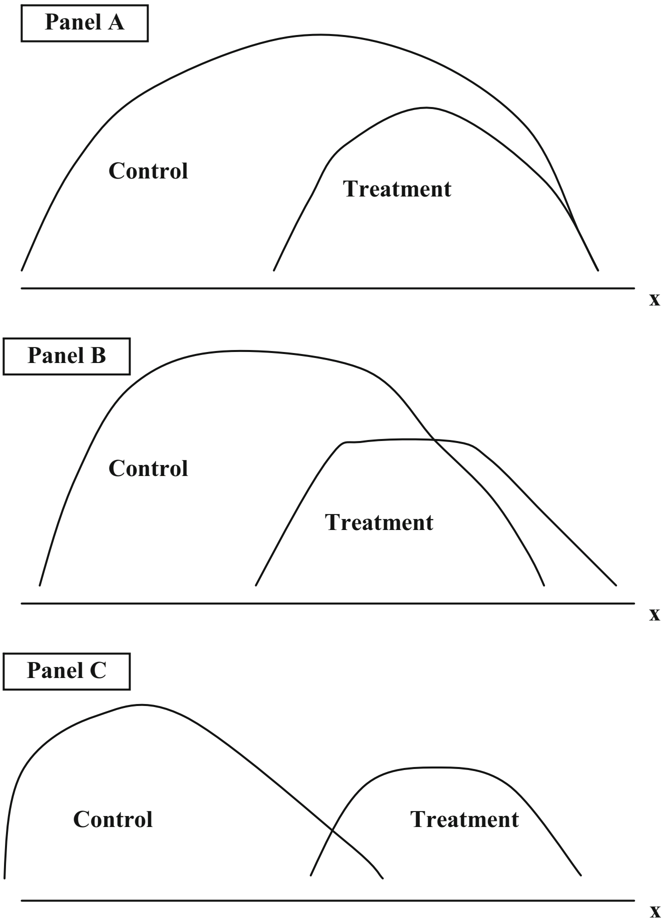

FIGURE 9.1. Some examples of common support.

be easily increased to multiple covariates important to treatment selection, in which case, the scale can then be thought of as some scalar combination of these values such as a propensity score (Rosenbaum and Rubin 1983).

First consider Fig. 9.1, Panel A. In this case, the entire distribution of the treatment group is overlapped by the distribution of controls on x. Thus, we have common support over this entire distribution of treated individuals, and hence, a proper counterfactual for inference can be generated in each case. Suppose we estimate the *ATT* in this case. We then might think of the *ATT* as a *global effect of treatment*, that is, a treatment effect which is an average of, and thus, generalizable to the entire treatment group.

Now consider Fig. 9.1, Panel B. Notice that much of the distribution of the treatment group overlaps with that of the control group on the covariate of interest, although the upper tail of the treatment distribution does not overlap. In this case, while most treated individuals do have a proper counterfactual with which to compare, some in the extreme tail do not. Screening these individuals out could potentially lead to the introduction of more bias in estimating a global *ATT*, as there are treated individuals we are not considering (Rosenbaum and Rubin 1985). However, as developed below, while it may be the case that we have a biased estimate of the global effect of treatment, in some instances, we may have a quantity which is still relevant for some specific policy of interest.

Finally, consider Fig. 9.1, Panel C, which appears to have very little overlap between the distributions of treatment and control groups. In this case, there may be very little we can infer regarding the effect of this treatment as there is simply no proper comparison in the data. In an instance such as this, we are powerless to do much. It is worth noting that this problem *may not* disappear with a larger sample size. In econometric parlance, this is what is known as a fundamental identification problem, that is to say, given the parameters of the problem, we are unable to implicitly identify the treatment effect of interest (see Manski 1995, for a general discussion of identification). Any attempt to quantify a treatment effect in such a case would need to rely on extrapolation, and by extension, be more conjecture than conclusion.

Global vs. Limited Effects

In an instances where we are not fully capable of estimating the global effect of a treatment (as in Fig. 9.1, Panel B above), it is often accepted that what we are able to estimate is merely a "biased" treatment effect and generally not applicable in an evaluation sense (Rosenbaum and Rubin 1985). However, it may be the case that the effect which we have estimated is actually interpretable and valuable, from a policy perspective, despite not being a valid estimate of the global treatment effect. For instance, if there are some individuals within the population who are likely to be treated with high probability, or probability equal to one, then the effect of the treatment on these individuals, at least in a policy evaluation sense, is generally likely to be irrelevant. The policy question is not whether it is worth treating these individuals; they are going to be treated anyway. Thus, the fact that no counterfactual exists in the data for these individuals is not necessarily problematic. We would likely be more interested in answering a policy question relating to those individuals who have some chance or none of being treated. In many instances, these are the cases falling in the area of common support, as they are the individuals "in play" in the policy sense.

Consider our earlier example of releasing individuals from a forensic psychiatric hospital. Even with variability in governing state laws and clinical practice, under which offenders get released, an individual with multiple arrests who has committed murder and does not respond to medication will not be released; that is, his probability of release is equal to 0, or conversely, the probability of being retained equal to 1. Conversely, a young individual with a strong response to treatment, a history of public nuisance crimes, and no prior arrests will be released with probability equal to 1 (again, that is, he or she will likely have no chance of being retained indefinitely in the hospital). Therefore, attempting to compute a global *PATE* or *ATT*, it can be argued, is conceptually irrelevant in a practical sense, as there will never be a situation where the laws and practice will be changed so as to release the murderer mentioned above. Again, conversely, it is also difficult to imagine a world where the laws are expanded to the point where the young, treatment-responsive misdemeanant will be retained for a long period.

In cases like this, a more relevant policy question might instead be to ask, if we redraw the line of demarcation for release from these facilities, where is the logical place to put it? Few would argue that all of the murderers portrayed above should be released. Instead if we look at the effects of release (say, on subsequent recidivism as a measure of future criminality), *only on those where there is discretion involved*, we may evaluate release policy in a more pragmatic sense. Thus, the absence of a global treatment effect in this case is not necessarily handicapping to the analysis, but instead, potentially conceptually preferable.

Global vs. Within-Group Effects

Another important instance of the potential of delimited effects occurs when we are capable of estimating a global effect, yet the particular policy problem of interest is better addressed by examining more localized, within-stratum effects. This situation occurs when one appreciates the possibilities created by the issue of heterogeneity in a global treatment effect, introduced earlier. In some instances, there may be a global null effect of a treatment, and one may be tempted to say that the treatment in question is useless. Further consideration of possible heterogeneous treatment effects, though, can lead to more focused analyses beyond the simple assessment of the overall treatment effect. It may be that there is considerable heterogeneity within various strata of the treatment group and the treatment is causing different groups to respond significantly, but in different ways.

The effects of institutional placement in juvenile justice provide relevant examples. In this area, there is often a global null effect of institutional care on the rate of rearrest in adolescent offenders. As a result, it may be tempting to dispel any consequences, positive or negative, of the use of these types of settings. It may be the case, however, that for certain subgroups within the larger population, say older, more-seasoned offenders, being placed in an institution serves no useful deterrent to future criminality. Conversely, for another subgroup, say younger, more impressionable offenders, such a placement actually exposes them to new people and situations, and thus, for such individuals, is actually criminogenic. Thus, depending on the "case mix" of the treatment group, we might see an overall null effect, even though there were actually strong opposite effects according to age group. If we feel that there is theoretical or practical heterogeneity in the treatment group, then it is possible, if not likely, that there is also heterogeneity in response to treatment. Depending on the policy question of interest, it is likely worth examining subgroup effects to focus future policy alternatives more specifically.

Summary

In the presence of observational data, it is typical of researchers to attempt to address selection, yet ignore heterogeneity, when quantifying treatment effects. We argue that both such problems are equally dangerous, and when considering them, a focus on the relevant policy problem or question should be employed to drive the analysis and results. Furthermore, often despite the efforts to control for a selection effect, in observational data, there may always be some unmeasured confounders, which we cannot account for because we have no observations of their values. If such unobservable factors exist, they may be responsible for bias but we have no knowledge of the direction or magnitude of that bias. That is why randomization is the gold standard, as it creates reasonable balance over all covariates, observable or not.

In the case of observational data, where we suspect such hidden biases may exist, it is useful to conduct a *sensitivity analysis* for hidden biases in order to address this possibility. A sensitivity analysis asks what some unmeasured covariate *would have to be like* in order to materially alter the conclusions of the study. Notice we are unable to actually prove (or disprove) the existence of such a confounder, but if our results are highly sensitive to potential hidden biases, then we may wish to reevaluate our general policy conclusions (for more information of measuring sensitivity to hidden biases, see Rosenbaum 2002). Such additional information about the robustness of treatment effects can be very informative, but is rarely provided.

CONCLUSIONS

Empirical validation of the effects of practices and policies is central to the improvement of interventions in criminal justice. The magnitude and specific effect of an intervention can tell us whether that intervention is worth continuing from a cost-benefit perspective, or whether the intervention should even be implemented in the first place. Too often, however, this question is framed in a rather general manner, with little critical examination of whether the treatment effect calculated matches the question asked about the implementation of the policy or practice. The points raised in this chapter address the nuances of the different methods for calculating treatment effects and emphasize the fact that not all treatment effects are created equal. Different policy problems and research questions require different approaches to the quantification of treatment effects.

We have discussed several quantities, all of which fall under the rubric of "treatment effect," i.e., the *PATE*, *ATT*, *LATE*, and various extensions of these quantities, which condition on a specific subset of the population. None of these quantities is necessarily "right" or "better." The main idea we stress is that the burden falls on the researcher to determine which, if any, of these quantities is most relevant for the purpose of a particular policy application, Naive use of any these quantities when another is better suited holds the potential for inappropriate conclusions.

We urge the reader to focus on two central issues, selection bias and response heterogeneity when assessing the utility of any particular treatment effect estimate.

As we note, selection bias is a common, well-recognized nemesis of the social science researcher, and we have oftentimes become somewhat complacent about documenting its effects. In the absence of a selection bias (or, alternatively, if one feels that it has been completely and properly accounted for), one is often tempted to invoke an argument of causality. Although tempting, we urge the reader to treat usage of the term "causal effect" with the same level of concern that one might adopt in handling a container of highly explosive material. Indeed, we may even have been too casual in how we discussed causality above, in the context of our examples, providing insufficient attention to the key assumption of *ignorability of treatment* in most treatment effect literature. That is, when attempting to make causal inference, we must be sure there is nothing unobservable, which is potentially biasing our estimates despite our best efforts to control for observables. This is why randomized trials, if done correctly, remain the undisputed champion for inferring causality, since they are able to rule out both observable and unobservable confounders. In the absence of randomization, our ability to completely rule out the latter is oftentimes tenuous.

Finally, we urge the reader to consider that, even in the situation where one feels most or all of the bias due to selection has been eliminated, the quantity still may not have a universal interpretation. The idea of heterogeneity of the effect of treatment is one that is given considerably less attention than the related problem of selection. It is, however, equally, if not in some cases, more important. Our position is that failure to consider the effects of heterogeneity of treatment may rob us of many opportunities to be more useful than we currently are to policy makers. The question may be posed as to whether something "works" or not, but there are multiple ways in which we can provide an answer that illuminates under what conditions and how it works best. Choosing the precise quantity for characterizing a treatment effect that is applicable to the policy question at hand is one key to making evaluation more informative in criminal justice.

REFERENCES

Angrist JD (1990) Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. Am Econ Rev 80:313–335

Angrist JD (2004) Treatment effect heterogeneity in theory and practice, The Royal Economic Society Sargan Lecture. Econ J 114:C52–C83

Angrist JD (2006) Instrumental variables methods in experimental criminological research: what, why, and how. J Exp Criminol 2:23–44

Angrist J, Imbens G, Rubin DB (1996) Identification of causal effects using instrumental variables. J Am Stat Assoc 91:444–455

Heckman JJ (1997) Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. J Hum Resour 32(2):441–462

Heckman JJ, Smith JA (1995) Assessing the case for social experiments. J Econ Perspect 9(2):85–110

Holland PW (1986) Statistics and causal inference. J Am Stat Assoc 81:945–960

Berk RA, Sherman LW (1988) Police response to family violence incidents: an analysis of an experimental design with incomplete randomization. J Am Stat Assoc 83(401):70–76

Imbens GW, Angrist JD (1994) Identification and estimation of local average treatment effects. Econometrica 62: 467–475

LaLonde RJ (1986) Evaluating the econometric evaluations of training programs with experimental data. Am Econ Rev 76:604–620

Manski CF (1995) Identification problems in the social sciences. Harvard University Press, Cambridge

McCaffrey DF, Ridgeway G, Morral AR (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychol Methods 9(4):403–425

Needleman HL, Riess JA, Tobin MJ, Biesecker GE, Greenhouse JB (1996) Bone lead levels and delinquent behavior. J Am Med Assoc 275(5):363–369

Nevin R (2000) How lead exposure relates to temporal changes in IQ, violent crime, and unwed pregnancy. Environ Res 83(1):1–22

Neyman JS (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci 4:465–480

Ridgeway G (2006) Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. J Quant Criminol 22(1):1–29

Robins JM, Greenland S, Hu F-C (1999) Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. J Am Stat Assoc 94:687–700

Robins JM, Hernan MA, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11(5):550–560

Rosenbaum PR (2002) Observational studies, 2nd edn. Springer-Verlag, New York

Rosenbaum P, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika 70:41–55

Rosenbaum PR, Rubin DB (1985) The bias due to incomplete matching. Biometrics 41:103–116

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 66:688–701

Rubin DB (1977) Assignment to treatment groups on the basis of a covariate. J Educ Stat 2:1–26

Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6:34–58

Shadish WR, Cook TD, Campbell DT (2001) Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin, Boston

Sherman LW, Berk RA (1984) The specific deterrent effects of arrest for domestic assault. Am Sociol Rev 49(2):261–272

Weisburd D, Lum C, Petronsino A (2001) Does research design affect study outcomes in criminal justice? Ann Am Acad Pol Soc Sci 578:50–70

Wright JP, Dietrich KN, Ris MD, Hornung RW, Wessel SD, Lanphear BP, Ho M, Rae MN (2008) Association of prenatal and childhood blood lead concentrations with criminal arrests in early adulthood. PLoS Med 5:e101