CHAPTER 26

Propensity Score Matching in Criminology and Criminal Justice

ROBERT J. APEL AND GARY SWEETEN

INTRODUCTION

Researchers in the discipline of criminology have long been interested in the "treatment effect" of discrete events on an individual's delinquent and criminal behavior. Is stable employment associated with less delinquency and crime? What is the effect of marriage to a noncriminal spouse on desistance? Is high school dropout related to a higher rate of offending? Notice the conspicuous absence of explicitly causal terminology in the way that these questions are posed. Criminological research is typically reliant on nonexperimental data, or data in which the treatment of interest (e.g., employment, marriage, dropout) is often not randomized for ethical or practical reasons. Yet this creates a situation known as the selection problem – a situation in which individuals are free to exercise some degree of discretion or choice with regard to the event(s) that they experience. The selection problem arises when there is something peculiar to the choice set, the choice maker, or some combination thereof, which partially determines the individual's choice as well as his or her behavioral response to that choice. Moreover, researchers are rarely privy to all of the relevant factors that go into the choice-making process.

In the absence of randomization, researchers interested in point identification of causal effects are forced to approximate the conditions of a controlled experiment. The evaluation literature is replete with examples of different quasi-experimental techniques, so called because they attempt to create a situation in which treatment is "as good as randomly assigned." In this chapter, we are interested in a procedure generally known as matching, which is a kind of *selection on observables* approach to causal effect estimation (see Heckman and Hotz 1989). Matching refers to a broad class of techniques that attempts to identify, for each individual in a treatment condition, at least one other individual in a comparison condition that "looks like" the treated individual on the basis of a vector of measured characteristics that may be relevant to the treatment and response in question. Matching is not new to criminology. For example, Glueck and Glueck (1950) matched each institutionalized youth in their classic study to a

A.R. Piquero and D. Weisburd (eds.), *Handbook of Quantitative Criminology*,
DOI 10.1007/978-0-387-77650-7_26,

school-going youth on the basis of age, ethnicity, intelligence, and neighborhood socioeconomic status. Widom (1989) matched each youth who was abused or neglected as a child to an officially nonabused youth by age, gender, race, and neighborhood socioeconomic status. These two prominent studies are examples of *covariate matching*, *categorical matching*, or *exact matching* (there are several naming conventions) on the basis of a handful of observed characteristics.

Exact matching, however, can become intractable as the list of measured characteristics grows, a problem known as the "curse of dimensionality." Fortunately, within the larger class of matching methods exists a technique known as *scalar matching* or *propensity score matching*. This is a data reduction technique that has the advantage of allowing researchers to match treated and comparison individuals on a very large number of measured characteristics, including pretreatment outcomes. In the remainder of this chapter, we will be specifically concerned with the details of propensity score matching. First, we will introduce the counterfactual framework on which the method is based. Second, we will turn to technical issues that arise in the use of propensity scores in applied research. Third, we will provide an empirical example of our own and describe other studies in criminology that employ the propensity score methodology. Finally, we will close the chapter with an outline to guide practice and some concluding comments.

PROPENSITY SCORE MATCHING IN THEORY

An understanding of propensity score matching is aided by familiarity with the language and logic of counterfactual estimation, known also as the analysis of *potential outcomes* (see Rubin 1974, 1977).[1] According to this framework, the causal effect of a binary treatment is the difference between an individual's value of the response variable when he or she is treated and that same individual's value of the response variable when he or she is not treated. Under this definition of causality, then, each individual experiences a response under *two simultaneous conditions*. As a straightforward example, consider the question of the effect of marriage on crime. For each individual, one must imagine outcomes under two alternate states: Crime patterns when an individual is married, and crime patterns when that individual is simultaneously not married.[2] The treatment effect of marriage is the difference between the crime outcomes for the same individual in these two simultaneous states. The "fundamental problem of causal inference" (Holland 1986) is that, for each individual, only one of the two potential outcomes is observed at any given time. The other is purely hypothetical, making direct estimation of the causal effect of treatment impossible.

To formalize the counterfactual approach to treatment effect estimation, suppose that all individuals in a target population have information on Y_i^1, their potential outcome under treatment; Y_i^0, their potential outcome under non-treatment; and T_i, their assigned treatment status that takes the values of 1 and 0, respectively, when treatment either is or is not received.

[1] Readers desiring a more thorough survey of the counterfactual framework generally and the propensity score method specifically are referred to Cameron and Trivedi (2005: chap. 25), Imbens (2004), Morgan and Harding (2006), and Wooldridge (2002: chap. 18).

[2] Notice that the counterfactual definition of causality requires that the individual occupy two states at the same time, not two different states at two different times. If the latter condition held, panel data with a time-varying treatment condition would suffice to estimate a causal effect of treatment. In the marriage example, the period(s) in which the individual is not married would be the counterfactual for the period(s) in which the same individual is married.

Treatment is defined generally as any form of intentional intervention. For each individual, the causal effect of treatment is computed as $Y_i^1 - Y_i^0$. Aggregating across the target population, the *average treatment effect* (ATE) is defined as the expected effect of treatment on a randomly selected person from the target population:

$$\begin{aligned} \text{ATE} &= \text{E}\left(Y_i^1 - Y_i^0\right) \\ &= \text{E}\left(Y_i^1\right) - \text{E}\left(Y_i^0\right) \end{aligned} \tag{26.1}$$

Note that ATE may also be written as a function of two other types of treatment effects. The *average treatment effect on the treated* (ATT) is defined as the expected effect of treatment for those individuals actually assigned to the treatment group, or the "gain" from treatment among those in the treated group:

$$\begin{aligned} \text{ATT} &= \text{E}\left(Y_i^1 - Y_i^0 \middle| T_i = 1\right) \\ &= \text{E}\left(Y_i^1 \middle| T_i = 1\right) - \text{E}\left(Y_i^0 \middle| T_i = 1\right) \end{aligned} \tag{26.2}$$

Although rarely of policy interest, it is also possible to determine the *average treatment effect on the untreated* (ATU), defined as the expected effect of treatment for those individuals assigned to the non-treatment (i.e., control or comparison) group:

$$\begin{aligned} \text{ATU} &= \text{E}\left(Y_i^1 - Y_i^0 \middle| T_i = 0\right) \\ &= \text{E}\left(Y_i^1 \middle| T_i = 0\right) - \text{E}\left(Y_i^0 \middle| T_i = 0\right) \end{aligned} \tag{26.3}$$

Collecting terms, ATE can be rewritten as a weighted average of ATT and ATU:

$$\text{E}\left(Y_i^1 - Y_i^0\right) = \Pr\left(T_i = 1\right)\text{E}\left(Y_i^1 - Y_i^0 \middle| T_i = 1\right) + \Pr\left(T_i = 0\right)\text{E}\left(Y_i^1 - Y_i^0 \middle| T_i = 0\right) \tag{26.4}$$

The first term on the right-hand side is ATT, weighted by the probability of treatment, and the second term is ATU, weighted by the probability of nontreatment.[3]

The obvious problem for causal identification under this framework is that only one of the two potential outcomes is observed for all individuals in the target population. The rule for determining which potential outcome is observed for any given individual is written as:

$$Y_i = \begin{cases} Y_i^1 & \text{if} \quad T_i = 1 \\ Y_i^0 & \text{if} \quad T_i = 0 \end{cases}$$

where Y_i represents the observed (as opposed to potential) outcome and T_i represents the observed treatment.

The potential outcomes framework reveals that treatment effect estimation represents, fundamentally, a missing data problem. The problem can be illustrated more clearly by decomposing ATE in terms of the known factuals and unknown counterfactuals. Inserting (26.2) and (26.3) into (26.4) yields:

$$\begin{aligned} \text{E}\left(Y_i^1 - Y_i^0\right) = {} & \Pr\left(T_i = 1\right)\left[\text{E}\left(Y_i^1 \middle| T_i = 1\right) - \text{E}\left(Y_i^0 \middle| T_i = 1\right)\right] \\ & + \Pr\left(T_i = 0\right)\left[\text{E}\left(Y_i^1 \middle| T_i = 0\right) - \text{E}\left(Y_i^0 \middle| T_i = 0\right)\right] \end{aligned}$$

[3] In this chapter, we will be mostly concerned with estimation of ATE rather that its constituents, ATT and ATU.

The unknown quantities, or counterfactuals, are $\mathrm{E}\left(Y_i^1 \middle| T_i = 0\right)$ and $\mathrm{E}\left(Y_i^0 \middle| T_i = 1\right)$. These are, in words, the potential outcome under treatment for individuals in the untreated sample, and the potential outcome under non-treatment for individuals in the treated sample, respectively. The goal of treatment effect estimation is the imputation of a hypothetical value or range of values for these missing counterfactuals.

The underlying issue for estimation of the causal effect of treatment is *balance*, or ensuring that treated individuals are statistically equivalent to untreated individuals on all background factors that are relevant for estimating the causal effect of interest. To the extent that balance is achieved, treatment is said to be exogenous or *ignorable*, meaning that treatment assignment is independent of the potential outcomes. The selection problem, of course, arises when treatment is endogenous or non-ignorable. In the case of the marriage-crime relationship, for example, individuals who get married might be endowed with a variety of other characteristics that are correlated with low crime, such as higher education and income, better long-term employment prospects, and no arrest history. They might also differ in ways that are challenging to observe and measure, for example, their intelligence, orientation to family, desire for children, career ambition, and ability to delay gratification. The goal of treatment effect estimation is to confront this endogeneity in a tractable and compelling manner.

Randomization of treatment ensures that treatment assignment is independent of potential outcomes (formally, $Y_i^1, Y_i^0 \perp T_i$), rendering the treatment assignment process ignorable. By virtue of its design, randomization also achieves balance (in expectation) on all potential confounding variables, observed and unobserved.[4] Consequently, the control group can be used as a valid counterfactual source for the experimental group.[5] Because the experimenter exercises *physical control* over the treatment conditions that participants experience, adjustment for additional covariates is technically unnecessary because the selection problem is eliminated (although covariate adjustment is still good practice, even in an experiment). Returning to the marriage-crime example, to the extent that a researcher can assign individuals to marriage and non-marriage with the flip of a fair coin, for instance, the treatment effect of marriage on crime can be readily estimated.[6] The average treatment effect is simply the mean of the response variable for the treated (married) less the mean of the response variable for the untreated (unmarried).

---

[4] Because it renders treatment ignorable, randomization is sufficient to identify the average treatment effect in the following manner:

$$\begin{aligned}\mathrm{ATE} &= \mathrm{E}\left(Y_i^1 \middle| T_i = 1\right) - \mathrm{E}\left(Y_i^0 \middle| T_i = 0\right) \\ &= \mathrm{E}\left(Y_i \middle| T_i = 1\right) - \mathrm{E}\left(Y_i \middle| T_i = 0\right)\end{aligned}$$

Notice that this is simply the mean difference in the outcome for treated and untreated individuals in the target population, as the potential outcomes notation in the first equality can be removed. The second equality necessarily follows because treatment assignment independent of potential outcomes ensures that:

$$\mathrm{E}\left(Y_i^1 \middle| T_i = 1\right) = E\left(Y_i^1 \middle| T_i = 0\right) = E\left(Y_i \middle| T_i = 1\right)$$

and

$$\mathrm{E}\left(Y_i^0 \middle| T_i = 1\right) = E\left(Y_i^0 \middle| T_i = 0\right) = E\left(Y_i \middle| T_i = 0\right)$$

As an interesting aside, in the case of a randomized experiment, it is also the case that ATT and ATU are equivalent to ATE by virtue of these equalities.

[5] To be perfectly accurate, randomization may in fact produce imbalance, but the imbalance is attributable entirely to chance. However, asymptotically (i.e., as the sample size tends toward infinity) the expected imbalance approaches zero.

[6] Aside from ethical and practical concerns, this experiment would be unable to assess the effect of marriage as we know it, as marriages entered into on the basis of a coin flip would likely have very different qualities than those freely chosen.

In observational studies, on the other hand, the researcher exercises no physical control over the treatment status. In this setting, the researcher strives to achieve *statistical control* over treatment assignment in a way that approximates the conditions of randomization. In the case of marriage and crime, the challenge is to identify one or more unmarried individuals that may serve as a counterfactual source for each married individual. One method of doing so is to choose, as a counterfactual source, a group of untreated individuals who most closely resemble the individuals in the treatment group as measured by a large number of potential confounding variables. If this requirement is satisfied, treatment is then said to be independent of potential outcomes, conditional on the confounding variables, a situation known as the *conditional independence assumption*, a scenario that can be formalized as $Y_i^1, Y_i^0 \perp T_i \mid x_i$. In other words, balance is achieved (i.e., treatment assignment is ignorable) once the relevant covariates are properly controlled. Rosenbaum and Rubin (1983) additionally show that conditional independence, given a vector of covariates, implies conditional independence, given a particular function of the covariates known as a propensity score, to which we now turn.

PROPENSITY SCORE MATCHING IN PRACTICE

In evaluation practice, it is often the case that a comparatively small proportion of the target population is treated relative to the proportion of individuals who are not treated. In addition, not all untreated individuals are equally desirable comparisons for treated individuals, as they may be quite different with respect to background characteristics, and thus, altogether inappropriate as a counterfactual source. Propensity score matching offers a way to select a subsample of treated and untreated cases that are observationally equivalent (or at least observationally similar) so that valid treatment effect estimates may be obtained. Rooted in the work of Rosenbaum and Rubin (1983, 1984, 1985), a propensity score is defined as "the conditional probability of assignment to a particular treatment, given a vector of observed covariates" (Rosenbaum and Rubin 1983: 41). It is equivalent to a *balancing score*, whereby, the conditional distribution of the covariates, given the propensity score, is the same for treated and untreated respondents. With the demonstration of balance, a researcher then has a stronger case for assuming that treatment is "as good as randomly assigned," at least with respect to the variables included in the estimation of the propensity score.

Propensity score matching is similar to standard regression in that it is assumed in both cases that selection into treatment is random conditional on observed characteristics ("selection on observables"). However, propensity score matching differs from regression techniques in two key respects. First, it does not rely on a linear functional form to estimate treatment effects. Although propensity scores are typically estimated using a parametric model, once these are obtained, individuals are usually matched nonparametrically. Second, propensity score matching highlights the issue of common support. It reveals the degree to which untreated cases actually resemble the treated cases on observed characteristics. Standard regression (known as covariate adjustment), on the other hand, obscures this issue, and can, in some situations, extrapolate treatment effect estimates, based solely on functional form when treated and untreated groups are actually not comparable at all. In many applications, only a subset of the treated and untreated populations will be useful (read, valid) for estimating treatment effects.

In practice, propensity score matching involves at least three steps. First, the propensity score must be estimated. Second, the conditional independence assumption must be evaluated

by demonstrating balance on potential confounders. Third, the treatment effect of interest (ATE, ATT, or ATU) must be estimated. Each step is considered in more detail below.

Estimation of the Propensity Score

Propensity score matching begins with the formal specification and estimation of a treatment status model. At this stage, great care must be taken to choose a specification that fully characterizes the treatment assignment mechanism with respect to variables that potentially confound the relationship between treatment status and the outcome variable. The goal is to model the non-random elements of the selection process using observed information.[7] A propensity score, $P(x_i)$, is defined as the probability of treatment, conditional on a vector of k observed confounders. [8] In most applications, it is conveniently estimated by way of the logistic distribution function:

$$\begin{aligned} P(x_i) &= \Pr(T_i = 1 | x_i) \\ &= \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})} \end{aligned}$$

The key output from the treatment status model is the predicted probability of receiving treatment, which is the propensity score. With the propensity score in hand, treatment status is presumed to be independent of potential outcomes conditional on the propensity score – this is the conditional independence assumption.

Once the propensity score is obtained, an important consideration is *common support*. This concerns the distribution of propensity scores for treated individuals versus the distribution for untreated individuals. Common support exists where the propensity score distributions for the two groups overlap. "Off-support" cases will typically be found in the tails of the propensity score distribution, since untreated individuals will tend to have a low probability of treatment, whereas treated individuals will have a comparatively high probability of treatment. Another support problem concerns those cases within the range of common support for which no comparable matches can be found due to a sparse propensity score distribution.

Absence of common support is valuable information, as it reveals conditions under which individuals *always* or *never* receive the treatment of interest within the sample (not necessarily within the population). There is some discretion in how many of the off-support cases to include in treatment effect estimates through the use of a bandwidth (or caliper), which is employed in several different kinds of matching protocols. The bandwidth specifies how far away a potential match can be on the propensity score metric and still be used as a counterfactual.[9]

---

[7] Researchers differ in their preferences for how exhaustive the treatment status model should be. In a *theoretically informed model*, the researcher includes only a vector of variables that are specified a priori in the theory or theories of choice. In a *kitchen sink model*, the researcher includes as many variables as are available in the dataset. In our view, a theoretically informed model is appealing only to the extent that it achieves balance on confounders that are excluded from the treatment status model but would have been included in a kitchen sink model.

[8] Some researchers also include functions of the confounders in the treatment status model, for example, quadratic and interaction terms.

[9] A useful sensitivity exercise is to estimate treatment effects using a number of different bandwidths to determine stability of the estimates. With smaller bandwidths, common support shrinks and fewer cases are retained. This

Demonstration of Covariate Balance

Once propensity scores have been estimated and support issues resolved, there are a number of ways in which one can test whether or not balance has been achieved. That is, whether conditioning on the propensity score makes treatment appear random, providing evidence for the validity of the conditional independence assumption. A common first step is to stratify the sample into equally-sized bins, based on the propensity score, and then test differences in both propensity scores and other independent variables between treated and untreated cases within bins, using independent samples t-tests (Dehejia and Wahba 1999, 2002).

Another very flexible way to assess balance is through estimation of the standardized bias (SB). This method, first described by Rosenbaum and Rubin (1985: 36), is equivalent to Cohen's d (Cohen 1988), a common measure of effect size. In order to assess balance, two SBs are to be computed: One before matching (unadjusted SB) and one after matching (adjusted SB). The degree to which the SB is attenuated by conditioning on the propensity score provides some indication of the degree to which the conditional independence assumption is satisfied. The formula for the SB is as follows:

$$\mathrm{SB} = 100 \times \frac{\bar{x}_i - \bar{x}_{i,j}}{\sqrt{\left(s_i^2 + s_j^2\right)/2}}$$

The means are indexed by i and i, j, signifying the mean of x for treated individuals less the mean for their matched, untreated counterparts. However, in the denominator, the variances are indexed by i and j, denoting the variance of x for all treated individuals and the variance of x for all untreated individuals, whether they are matched or not. Following convention (Cohen 1988; Rosenbaum and Rubin 1985), the rule of thumb used to judge whether the SB is substantively large – that is, whether the covariate in question is imbalanced – is $|\mathrm{SB}| \geq 20$.[10]

The SB is also a useful metric to assess balance across several matching methods by comparing the percent of potential confounders balanced under each method. Typically, the choice of a propensity score model and the assessment of balance is an iterative process. If treated and untreated cases are imbalanced, the initial propensity score model is modified to include more variables, or to include interaction or squared functions of key predictors until balance is achieved. Additionally, imbalanced covariates (if any) may be directly adjusted even after their inclusion in the treatment status model.

Estimation of Treatment Effects

The most common type of treatment effect is the average treatment effect (ATE), representing the average difference in potential outcomes across the entire population. Under other circumstances, however, the researcher may be concerned exclusively with the effect of a particular treatment only within that segment of the population who actually experienced the

alters the nature of the estimated treatment effect, particularly if a large number of cases are excluded. This can be dealt with by simply acknowledging that the estimated effect excludes certain kinds of cases, and these can be clearly described since the dropped cases are observed.

[10] Where substantive significance is as important as statistical significance, the standardized bias formula can also be used to estimate an effect size for the treatment effect estimate (see Cohen, 1988).

treatment. This is known as the average treatment on the treated (ATT). Alternately, one may be interested in estimating the average treatment on the untreated (ATU). It should be clear that ATE is simply a weighted average of ATT and ATU, as shown earlier. There are three conventional methods for estimating any of these treatment effect parameters: Regression, stratification, and matching.

The *regression method* involves controlling for the propensity score in a model that also includes the treatment status indicator (a dummy variable). The coefficient on this indicator provides an estimate of the ATE. This method represents a control function approach to treatment effect estimation. The regression model is specified as follows:

$$Y_i = \alpha + \beta T_i + \gamma P(x_i) + e_i$$

From this model, the ATE is, quite simply, β, the coefficient conforming to the treatment status indicator.[11] The regression method has the advantage of making it easy to adjust directly for covariates that remain imbalanced after conditioning on the propensity score. However, unlike other methods, this method relies on a specific functional form for assessing treatment effects. Also, to avoid extrapolating off-support, common support conditions should be imposed prior to employing the regression method.

The *stratification method* proceeds by dividing the sample into subclasses or strata of approximately equal size. This method groups several treated and untreated individuals within a common range of the propensity score. Since Cochran (1968) demonstrated that stratification into five subclasses removes approximately 90% of bias due to the stratifying variable (see also Rosenbaum and Rubin 1984), the accepted practice is to divide the sample into quintiles on the propensity score, although more strata may be used as sample size permits. A stratum-specific treatment effect is derived as the mean difference in the response variable between treated and untreated individuals in each stratum. The ATE is simply the weighted average of the stratum-specific treatment effects, where the weights are defined as the proportion of the sample in each stratum. Using S_i to index the propensity score stratum in which each individual is classified, ATE can be recovered from:

$$\text{ATE} = \sum_S \frac{N_s}{N} \left(\frac{1}{N_{1,s}} \sum_{i \in (T=1)} Y_i - \frac{1}{N_{0,s}} \sum_{i \in (T=0)} Y_i \right)$$

where N is the total number of individuals in the sample, N_s is the number of individuals in stratum s, $N_{1,s}$ is the number of treated individuals in stratum s, and $N_{0,s}$ is the number of untreated individuals in stratum s. By conditioning on the propensity score strata, the stratification method has the advantage of ensuring closer correspondence between treated and untreated cases within each subclass. Yet, even within strata, the cases may still be poorly matched.

The *matching method* involves selecting from the pool of all untreated individuals only those who closely resemble the treated individuals on the propensity score. This method is explicit about estimating the counterfactual response from a subset of the untreated sample.

---

[11] In practice, Wooldridge (2002) recommends augmenting the regression model in the following way:

$$Y_i = \alpha' + \beta' T_i + \gamma' P(x_i) + \delta' T_i \left[P(x_i) - \bar{P}(x_i) \right] + e_i'$$

where $\bar{P}(x_i)$ represents the mean propensity score for the target population and ATE is estimated the same way, but by using β' in place of β.

With suitable matches between treated and untreated individuals, then, it is possible to estimate ATE as follows:

$$\text{ATE} = \frac{1}{N} \sum_{i \in (T=1)} \left[Y_i - \sum_{j \in (T=0)} \omega_{i,j} Y_j \right]$$

where i indexes treated individuals, j indexes untreated individuals, and $\omega_{i,j}$ is an arbitrary "weight" assigned to the response of each untreated individual that is subject to the constraint that $\sum \omega_{i,j} = 1 \forall i \in (T = 1)$. The latter criterion means that for each treated individual, the weights for all of his or her matched untreated subjects sum to unity. Note that, for untreated individuals who are poor matches, the weight may be zero. Importantly, all matching methods may be characterized as weighting functions, with a variety of algorithms available that define match suitability in different ways.

The simplest form of matching involves nearest available matching or *nearest neighbor matching*, in which the untreated case with the closest propensity score to a treated case is used as the counterfactual. In this case, $\omega_{i,j} = 1$ for all matched untreated respondents, a scenario known as single-nearest-neighbor or one-to-one matching. Treated individuals may also be matched to multiple nearest neighbors, known as many-to-one matching, in which case, $\omega_{i,j} = 1/J_i$ where J_i is the number of matched untreated cases for each treated individual i, with a maximum number of matches selected by the researcher (e.g., 2-to-1, 3-to-1, and so on). With multiple nearest neighbors, the mean of several untreated individuals serves as the counterfactual. There is a tradeoff between using single and multiple nearest neighbors (see Smith and Todd 2005). Using multiple nearest neighbors decreases variance at the cost of increasing potential bias because matches are less accurate. On the other hand, single nearest neighbor matching decreases bias but increases variance. A variation on nearest neighbor matching is *caliper matching*, which selects the specified number of untreated cases from within a maximum distance or tolerance known as a caliper. If no untreated subjects lie with the chosen caliper, the treated subject is unmatched.[12]

A more complicated weighting protocol is *kernel matching*, which provides a means of differential weighting of untreated cases by their distance from the treated subjects.[13] Kernel matching allows for finer distinctions in weighting compared to other methods. With this method, the researcher must decide which kernel function and bandwidth to use. Any finite probability distribution function may be used as a kernel, but the three most common kernels used in practice are the uniform, the Gaussian, and the Epanechnikov. A uniform kernel

[12] Nearest neighbor matching can be done with or without replacement. *Matching without replacement* means that once an untreated case has been matched to a treated case, it is removed from the candidates for matching. This may lead to poor matches when the distribution of propensity scores is quite different for the treated and untreated groups. Matching without replacement also requires that cases be randomly sorted prior to matching, as sort order can affect matches when there are cases with equal propensity scores. *Matching with replacement* allows an untreated individual to serve as the counterfactual for multiple treated individuals. This allows for better matches, but reduces the number of untreated cases used to create the treatment effect estimate, which increases the variance of the estimate (Smith and Todd 2005). As with the choice of the number of neighbors, one has to balance concerns of bias and efficiency.

[13] When there are many cases at the boundaries of the propensity score distribution, it may be useful to generalize kernel matching to include a linear term; this is called local linear matching. Its main advantage over kernel matching is that it yields more accurate estimates at boundary points in the distribution of propensity scores and it deals better with different data densities (Smith and Todd 2005).

assigns a weight of $1/J_i$ to each of the J_i matched untreated cases within a chosen radius. What makes this different from matching with multiple nearest neighbors or caliper matching is simply that all available matches within the specified radius are used, rather than the pre-specified number of matches.

A Gaussian kernel matches all untreated individuals to each treated individual, with weights assigned to untreated cases that are proportional to the well-known normal distribution. One can imagine a normal curve centered above each treated case on the propensity score continuum, with its kurtosis (i.e., the sharpness of the distribution's peak) determined by the size of the bandwidth. Each untreated case which lies beneath this curve (which, under a normal curve, is *all* untreated cases), receives a weight proportional to the probability density function of the normal curve evaluated at that point. In effect, this accords the greatest importance to counterfactual cases that are the closest on the propensity score metric. An Epanechnikov kernel matches all untreated individuals within a specified bandwidth of each treated individual, but unlike the Gaussian kernel, assigns zero weight to untreated cases that lie beyond the bandwidth.

Each of these propensity score methods described above has particular advantages and disadvantages, and is useful under different circumstances. If, for example, there are numerous treated and untreated cases throughout the propensity score distribution, then the nearest neighbor matching without replacement may be the best option. However, if the distributions are quite different, then kernel matching may be preferred. There are no foolproof rules for implementing propensity score matching. Rather, one must carefully consider each research situation, preferably implementing at least a few different matching protocols (if appropriate) in order to assess sensitivity of the estimates to the matching method.

AN EMPIRICAL ILLUSTRATION OF THE RELATIONSHIP BETWEEN ADOLESCENT EMPLOYMENT AND SUBSTANCE USE

To provide a step-by-step example of propensity score matching, we examine the relationship between youth employment and substance use. This is a question that has been of longstanding interest to developmental psychologists and criminologists.[14] Three decades of research leaves little doubt – with a few notable exceptions – that employment during adolescence is positively correlated with the use of a wide variety of illicit substances, including cigarettes, alcohol, marijuana, and hard drugs. In particular, research demonstrates that employment which is of "high intensity" – referring to a work commitment that is over 20 h per week, generally the median work intensity among adolescents – and which takes place during the school year, is most strongly correlated with substance use.

The data chosen for this example are from the National Longitudinal Survey of Youth 1997, which is a nationally representative survey of 8,984 youth born during the years 1980 through 1984 and living in the United States during the initial interview year in 1997. The data

[14] Apel et al. (2006, 2007, 2008); Bachman et al. (1981, 2003); Bachman and Schulenberg (1993); Gottfredson (1985); Greenberger et al. (1981); Johnson (2004); McMorris and Uggen (2000); Mihalic and Elliott (1997); Mortimer (2003); Mortimer et al. (1996); Paternoster et al. (2003); Ploeger (1997); Resnick et al. (1997); Safron et al. (2001); Staff and Uggen (2003); Steinberg and Dornbusch (1991); Steinberg et al. (1982, 1993); Tanner and Krahn (1991).

were collected to document the transition from school to work in a contemporary sample of adolescents. There is thus rich information on individual work histories, in addition to a variety of other indicators of health and wellbeing, including delinquency and substance use. For this illustration, we select the subsample of youth in the 12- to 14-year-old cohorts. Exclusion of those who worked in a formal ("paycheck") job prior to the first interview or with missing information at the second interview yields a sample of 4,667 youth. Propensity score matching is then used to estimate the average treatment effect (ATE) of high-intensity work during the school year on illicit substance use.

The key independent variable – the "treatment" – is a binary indicator for employment in a formal job of at least 15 h per week during the school year between the first and second interviews. A 15-h threshold is chosen because this represents the median number of hours worked per week during the school year among the workers in this sample. Just under one-quarter of the sample, 23.0% ($N = 1{,}073$), are employed during the school year, and 11.8% ($N = 550$) are employed at high intensity during the school year. The dependent variable – the "response" – is a binary indicator for the use of cigarettes, alcohol, marijuana, or other hard drugs between the first and second interviews. Almost half the sample, 49.4% ($N = 2{,}304$), reports such use.

The National Longitudinal Survey of Youth 1997 also has available a wide variety of background variables measured from the first interview that may be used to model the probability of high-intensity work at the second interview. Because this analysis is intended to be little more than an illustration of the propensity score approach, we limit our attention to 18 variables, listed in Table 26.1. These include measures of individual demographics (gender, race/ethnicity, age), family background (intact family, household size, mobility, income, socioeconomic status, disadvantage), family process (parental attachment, parental monitoring), school engagement (test scores, attachment to teachers, homework), and miscellaneous risk indicators (antisocial peers, delinquency, arrest), including prior substance use.

Estimation of the Propensity Score and Evaluation of Covariate Balance

The probability of intensive employment during the school year – the propensity score – is estimated from the logistic regression model, shown in Table 26.1. The pseudo R-square from the model is 0.187, meaning that the model explains 18.7% of the variation in intensive work.[15] The mean propensity score for treated youth is 0.25 (median = 0.25) and for untreated youth is 0.10 (median = 0.06). The model findings indicate that minorities are significantly less likely to work intensively, while youth who are older at the first interview are significantly more likely to be intensively employed. Youth who are from highly mobile households are also more likely to be intensively employed, as are youth from high-SES households (higher values on the low-SES index indicate lower SES). Youth who have higher test scores (as measured by the Peabody Individual Achievement Test) are more likely to work intensively, as are youth who have more antisocial peers and who use a wider variety of illicit substances.

The distributions of the propensity scores for treated and untreated youth are displayed in Fig. 26.1. As a measure of covariate balance, we rely on the standardized bias (SB)

[15] If we select the sample treatment probability as the classification threshold, 71.8 percent of the sample is correctly classified from the model shown in Table 26.1.

TABLE 26.1. Descriptive statistics, propensity score model, and balance diagnostics

Variable	Mean (S.D.)	Logit model of intensive work *b* (S.E.)	Balance diagnostics: standardized bias (SB)	
			Before matching	After matching
Demographics				
Male	50.7%	0.135 (0.105)	3.0	4.0
Minority	47.9%	−0.364 (0.117)**	−11.9	−1.5
Age	13.8 (0.9)	1.377 (0.077)***	104.5	−1.8
Family background				
Both biological parents	49.9%	−0.087 (0.140)	−13.2	5.7
Household size	4.6 (1.5)	0.040 (0.036)	−4.4	−2.0
Residential mobility	0.5 (0.4)	0.320 (0.157)*	−20.5	1.2
Family income ($10K)	4.5 (4.1)	−0.014 (0.018)	−2.4	1.4
Low SES index	0.9 (1.1)	−0.119 (0.063)+	−10.5	1.4
Disadvantage index	1.7 (1.5)	0.053 (0.054)	3.7	−0.0
Family processes				
Attachment to parents	4.0 (1.3)	−0.025 (0.041)	−17.1	−0.7
Monitoring by parents	1.7 (1.1)	−0.004 (0.052)	−13.9	1.2
School engagement				
P.I.A.T. percentile (÷10)	4.9 (3.4)	0.030 (0.017)+	1.6	−0.8
Attachment to teachers	5.1 (1.5)	−0.029 (0.034)	−22.8	0.2
Hours of homework	2.0 (3.1)	0.012 (0.021)	−3.2	0.7
Risk indicators				
Antisocial peer affiliation	1.2 (1.5)	0.072 (0.036)*	50.8	3.7
delinquency variety	0.9 (1.3)	0.032 (0.043)	30.5	−4.8
Arrested	5.4%	0.187 (0.193)	24.2	−3.3
Substance use variety	0.7 (1.0)	0.117 (0.056)*	45.3	−7.2

Note: $N = 4{,}667$. Estimates are unweighted. All variables are measured at the first (1997) interview. Means of binary variables are presented as percentages. Descriptive statistics and balance diagnostics are estimated from cases with valid data. Pseudo R-square for the logit model is 0.187, with a correct classification rate of 71.8%. The post-matching standardized bias is based on single-nearest-neighbor matching with no caliper.
$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (two-tailed tests).

(Rosenbaum and Rubin 1985). Recall that variables with $|\text{SB}| \geq 20$ are considered imbalanced. By this criterion, seven variables are imbalanced prior to matching – age, mobility, teacher attachment, and all four of the risk indicators.[16] If a stricter criterion of ten is chosen instead, 12 variables are imbalanced prior to matching. After matching, however, no variable is imbalanced irrespective of whether 20 or 10 is chosen as the balance threshold. Notice also that, in all but one case (gender), the SB becomes smaller in absolute magnitude when matching is performed. The variable with the largest SB after matching is prior substance use ($\text{SB} = -7.2$).

[16] The sign of the standardized bias is informative. If positive, it signifies that treated youth (i.e., youth who work intensively during the school year) exhibit more of the characteristic being measured than untreated youth. Conversely, if negative, it means that treated youth have less of the measured quality than untreated youth.

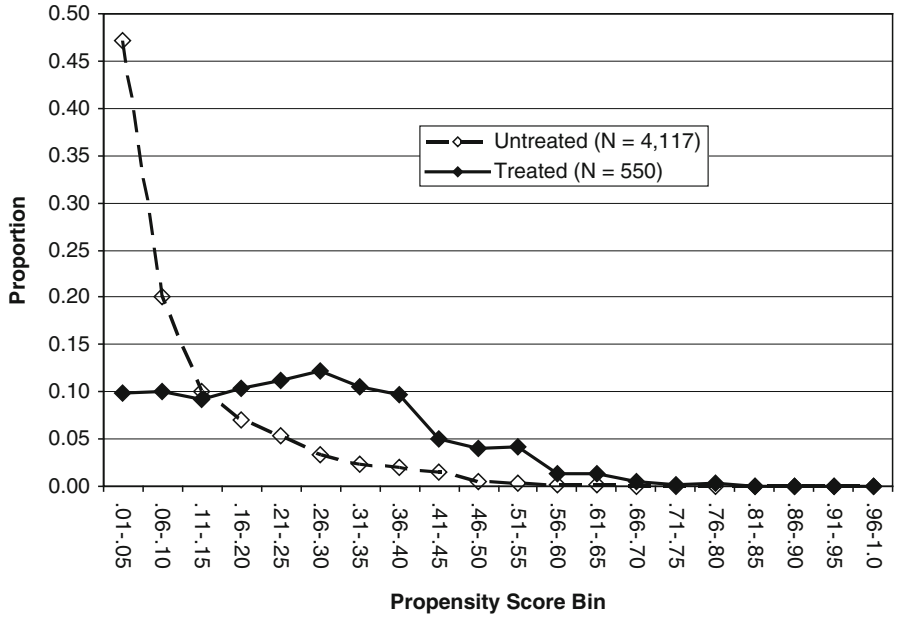

FIGURE 26.1. Propensity score distribution, by treatment status. Note: The sample includes 4,667 youth who were non-workers at the initial interview. "Treated" youth are those who report working at high intensity (more than 15 h per week) on average during the school year (September-May) between the first and second interviews. "Untreated" youth are those who report working at low intensity (15 or fewer hours per week) on average during the school year, or who report not working during the school year at all.

Estimation of the Average Treatment Effect

In Table 26.2, we provide estimates of the average treatment effect (ATE) from a variety of model specifications. In the first two rows (panel A), we perform standard regression adjustment, in which we regress substance use on intensive work. With no control variables, youth who work intensively during the school year are significantly and substantially more likely to engage in substance use. Specifically, their probability of substance use is 18.6 points higher than youth who work only moderately during the school year (15 or fewer hours per week) or who do not work during the school year at all. When the control variables from Table 26.1 are included, the ATE is attenuated but is still statistically significant, implying that intensively employed youth have a substance use likelihood that is 5.1 points higher. Thus, when traditional methods are used to establish the relationship between intensive work and substance use, the findings from previous research are replicated.[17] Specifically, high intensity during the school year appears to "cause" substance use.

In the remainder of Table 26.2, a variety of propensity score methods are employed – regression, stratification, and matching. Each method conditions on the propensity score in a slightly different way. First, in panel B, the regression method is used, in which substance use

[17] If a logistic regression model of substance use is estimated instead, the coefficient for intensive work with no control variables is 0.77 (odds ratio = 2.16), and with control variables is 0.28 (odds ratio = 1.33). Both coefficients are statistically significant at a five-percent level.

TABLE 26.2. Average treatment effect of intensive work on substance use

Model	ATE (S.E.)
A. Standard regression adjustment	
No control variables	0.186 (0.023)***
All control variables	0.051 (0.021)*
B. Propensity score model: regression	
No trimming	0.054 (0.026)*
Trim upper and lower 10%	0.074 (0.030)*
C. Propensity score model: stratification	
Five strata	0.057 (0.037)
Ten strata	0.028 (0.030)
D. Propensity score model: matching	
Nearest neighbor matching	
1 Nearest neighbor, no caliper	0.029 (0.045)
1 Nearest neighbor, caliper = 0.01	0.029 (0.045)
1 Nearest neighbor, caliper = 0.001	0.035 (0.036)
1 Nearest neighbor, caliper = 0.0001	0.024 (0.042)
3 Nearest neighbors, no caliper	0.032 (0.041)
3 Nearest neighbors, caliper = 0.01	0.032 (0.040)
3 Nearest neighbors, caliper = 0.001	0.050 (0.034)
3 Nearest neighbors, caliper = 0.0001	0.017 (0.041)
5 Nearest neighbors, no caliper	0.038 (0.040)
5 Nearest neighbors, caliper = 0.01	0.038 (0.040)
5 Nearest neighbors, caliper = 0.001	0.047 (0.036)
5 Nearest neighbors, caliper = 0.0001	0.016 (0.046)
Kernel matching	
Uniform kernel	0.029 (0.045)
Gaussian kernel	0.029 (0.045)
Epanechnikov kernel	0.029 (0.045)

Note: $N = 4{,}667$. Estimates are unweighted. For the propensity score regression and matching models, bootstrapped standard errors with 100 replications are provided.
$^{+}p < 0.10$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (two-tailed tests).

is regressed on intensive work, controlling only for the propensity score. In the first model, the entire sample is retained, while in the second model, only the middle 80% of the sample is retained. In both cases, the ATE of intensive work on substance use is positive and statistically significant. The results from propensity score regression thus harmonize with the results from standard regression adjustment.[18]

Next, in panel C, the sample is classified on the basis of the estimated propensity score into equal-sized intervals of five strata (quintiles) and ten strata (deciles). In both cases, the ATE is positive but no longer statistically significant. Notice also that, as the sample is classified into more strata – guaranteeing more homogeneous matches between the treated and

[18] Notice that the ATE from standard regression in panel A ($b = 0.051$) is very similar to the ATE from propensity score regression with no trimming in panel B ($b = 0.054$). The similarity is not coincidental. The discrepancy is only due to the fact that the propensity score was estimated from a logistic regression model at the first stage. Had a linear regression model been used instead, the two coefficients would be identical, although the standard errors would differ.

untreated in each stratum vis-à-vis the probability of intensive work – the ATE is halved. The results from propensity score stratification conflict with the foregoing models, and indicate that the positive correlation between intensive work and substance use is a selection artifact.

Finally, in panel D, a variety of matching protocols are employed.[19] We use single and multiple (three, five) nearest neighbors as matches within several calipers (none, 0.01, 0.001, 0.0001). We also employ kernel matching with three different kernels (uniform, Gaussian, Epanechnikov). In virtually all cases, the ATE is consistently in the neighborhood of 0.030, with a range of ATEs from 0.016 to 0.050. However, these estimates do not even approach statistical significance. The results from propensity score matching thus provide strong evidence that the relationship between intensive work and substance use is attributable to selection rather than causation.

To summarize, propensity score methods that ensure some degree of homogeneity between treated and untreated cases – stratification and matching, in particular – demonstrate that the positive correlation between intensive school-year employment and substance use is due to the fact that intensive workers differ in some fundamental way prior to their transition to work. Their higher substance use risk is, thus, due to self-selection rather than to the causal impact of their work involvement. In this application, regression-based methods – standard regression adjustment and propensity score regression – perform poorly because they fail to achieve a true "apples-to-apples" comparison between treated and untreated individuals.[20]

Other Examples of Propensity Score Matching in Criminology and Related Disciplines

The foregoing illustration represents a traditional application of propensity score methods to the study of individual-level treatment effects – in this case, of intensive school-year employment on substance use during adolescence. There are a number of other examples of this more traditional approach to treatment effect estimation in criminology, for example, arrest and wife battery (Berk and Newton 1985), court disposition and recidivism risk (Banks and Gottfredson 2003; Blechman et al. 2000; Caldwell et al. 2006; Krebs et al. 2009), youth employment and delinquency (Brame et al. 2004), exposure to community violence and perpetration of violence (Bingenheimer et al. 2005; Molnar et al. 2005), school sector and juvenile misbehavior (Mocan and Tekin 2006), marriage and crime (King et al. 2007), incarceration and crime (Sweeten and Apel 2007), and child maltreatment and adolescent weapon carrying (Leeb et al. 2007). In these studies, treatment status is modeled from as few as two covariates (Caldwell et al. 2006) to as many as 153 covariates (Bingenheimer et al. 2005). There are applications of regression (Banks and Gottfredson 2003; Berk and Newton 1985; Caldwell et al. 2006; Mocan and Tekin 2006; Molnar et al. 2005), stratification (Bingenheimer et al. 2005; Blechman et al. 2000; Brame et al. 2004; Leeb et al. 2007), and matching (King et al. 2007; Krebs et al. 2009; Sweeten and Apel 2007).

[19] We employ the user-written Stata protocol -psmatch2- to estimate average treatment effects from the matching models (see Leuven and Barbara 2003). To obtain the standard error of the ATE, we perform a bootstrap procedure with 100 replications.

[20] As a further test of sensitivity, we estimated the ATE of intensive employment on substance use for subsamples with different substance use histories. For this test, we employed single-nearest-neighbor matching with no caliper, although the findings were not sensitive to this choice. Among the 2,740 youth who, at the initial interview, reported never having used illicit substances, ATE = 0.084 (S.E. = 0.060). Among the 1,927 youth who reported having used at least one type of illicit substance prior to the initial interview, ATE = −0.019 (S.E. = 0.046).

There have also been recent applications of the propensity score methodology in criminology that extend the method in appealing ways. One extension is the use of generalized boosted regression (GBR) in the construction of the propensity score, providing a more flexible model of treatment assignment compared to standard logistic regression, using higher-order polynomials and interaction terms. McCaffrey et al. (2004) applied the method to the study of the effect of a community-based, residential drug treatment program on alcohol and marijuana use among juvenile probationers (see Ridgeway 2006, for an application of GBR to racial bias in citation rates, search rates, and stop duration in police traffic stops). They found that GBR was superior to the standard logit model as determined by pre-treatment balance in the confounders, prediction error in the propensity scores, and the variance of treatment effect estimates.

A second extension is the application of the propensity score methodology to the study of aggregate-level treatment effects, such as at the level of neighborhoods. Tita and Ridgeway (2007) were interested in the impact of gang formation on 911 call volume (calls for service) among Pittsburgh census block groups. They employed a technique known as propensity score weighting that involves weighting untreated cases by the odds of treatment (as predicted by the treatment status model) such that the distribution of covariates is balanced between treated and untreated cases (see Hirano et al. 2003; Robins et al. 1992; Robins and Rotnitzky 1995; see McNiel 2007, for an application of weighting to the study of the effect of mental health courts on recidivism).

A third extension is the estimation of treatment effects when the treatment, outcome, and confounders are all time-dependent. Sampson et al. (2006) estimated the treatment effect of marriage on official criminal behavior through middle adulthood (ages 17–32) and later adulthood (ages 17–70) among men institutionalized in a Boston-area reform school as boys (see Glueck and Glueck 1950). They employed a variation on propensity score weighting known as inverse probability-of-treatment weighting (IPTW) that involves adjusting treatment effect estimates by a time-varying propensity score incorporating cumulative treatment exposure (see Robins 1999; Robins et al. 2000). Similarly, Nieuwbeerta et al. (2009) estimated the time-dependent treatment effect of incarceration on recidivism in a Dutch conviction cohort, but in the context of a discrete-time event history model. This approach is referred to as risk set matching (see Li et al. 2001; Lu 2005).

A fourth extension is the integration of propensity scores into the group-based trajectory methodology of Nagin (2005). In a series of papers, Haviland and Nagin (2005, 2007; Haviland et al. 2007) were interested in the effect of joining a gang on youth violence. Youth were first classified into trajectory groups on the basis of finite mixture models of prior violence throughout early adolescence. Within trajectory groups, propensity scores were then estimated using information from background covariates. Using this two-step approach, Haviland and Nagin demonstrated balance not only on potential confounders but also on prior realizations of the response variable.

DISCUSSION OF GOOD RESEARCH PRACTICE

Before closing this chapter, we would like to offer some of our own thoughts on what we regard as good research practice with the use of propensity scores.

(1) *Have a clear, unambiguous definition of treatment, preferably the first experience of treatment.* The propensity score methodology is best conceived as an observational

analog to a randomized experiment involving a novel treatment. Otherwise, the right-hand side of the treatment status model should account for state dependence effects, that is, the cumulative experience of the treatment state. Moreover, in these instances, it might be fruitful to match directly on individuals' cumulative treatment status in addition to the propensity score.

(2) *Be thoughtful about the specification of the treatment status model.* A concise, theoretically informed model of treatment status is not necessarily desirable, simply because criminological theory is not as explicit as other disciplines about relevant confounders in the treatment–response relationship. In the case of propensity score models, parsimony is not necessarily a virtue. Parsimonious models are only appropriate to the extent that balance can be demonstrated on confounders that are not included in the treatment status model. The researcher should make the case that any nonrandom sources of selection into treatment are fully captured by the treatment status model. Failure to do so, or to do so convincingly, renders the propensity score method no better than standard regression adjustment.

(3) *Include confounders that are temporally prior to treatment in the treatment status model.* The treatment status model should necessarily include predictors that are realized prior to treatment assignment. This should include pretreatment outcomes when they are available. Temporal priority of confounders vis-à-vis treatment implies that panel data are preferable to crosssectional data, because potential confounders at one time period can be used to model treatment status at the next time period. Modeling treatment status from crosssectional data is potentially problematic because some of the confounders in the prediction model could actually precede treatment status in time. A possible exception to this guideline is when the confounders refer to behavior prior to the interview, while treatment refers to behavior at the time of the interview, or even perhaps to expected behavior.

(4) *Demonstrate that the support condition is satisfied.* At a minimum, this should entail displaying the propensity score distributions for treated and untreated cases to ensure that there is sufficient overlap between them. This is especially important if the propensity score is used as a control variable in a regression model, in which case, results can be misleading if there are unmatched or poorly matched cases, as in the empirical example we provided earlier.

(5) *Demonstrate balance on confounders that are included in the treatment status model, as well as balance on confounders that are not included in the treatment status model, if available.* The latter criterion increases confidence that the conditional independence assumption has been satisfied. The most useful balance diagnostics include t-tests and estimates of the standardized bias (SB). Confounders in the treatment status model that remain imbalanced should be adjusted directly in the estimate of the treatment effect.

(6) *Employ multiple propensity score methods as tests of robustness.* Generally speaking, we would advise against using propensity score regression, simply because our experience is that conclusions from these models rarely differ from standard regression adjustment. However, using propensity score stratification, this guideline means assessing the sensitivity of estimated treatment effects to the number of strata. If using propensity score matching, this means testing sensitivity to the specific matching protocol employed. Ideally, estimates from multiple methods will agree in substance. However, applied research settings are rarely ideal, and it is not uncommon for two

propensity score methods to lead to different inferences. In such instances, the relative merits of the two models should be closely evaluated before a decision is made about which set of results to report.

(7) *Think carefully about what type of treatment effect is relevant for the study.* For basic social scientific questions, the typical parameter of interest is the average treatment effect (ATE), which we estimated in our empirical example. However, in many policy applications, the average treatment effect on the treated or the untreated (ATT or ATU) may actually be more relevant. Consider, for example, an evaluation of a targeted after-school program for delinquency prevention. Program stakeholders and potential adopters of the program are clearly interested in ATT: "What effect did the program have on those individuals who participated in the program?" However, if one were interested in expanding the target population for the program, ATU may be the more relevant consideration. For example, consider the question of the effect of net-widening criminal justice policies such as the use of mandatory arrest: "What effect would the program have on those individuals who are not the usual targets of the intervention?"

CONCLUDING REMARKS

Propensity score matching has emerged as an important tool in the program evaluator's toolbox. Criminology is no exception to this trend, as indicated by the growing number of studies that employ the method. In fact, of the two dozen or so criminological studies reviewed earlier, all but one have been published since 2000. While propensity score matching is by no means a panacea to the pernicious problem posed by selection bias (what quasi-experimental method is?), it does have several advantages to recommend its use in causal effect estimation, as we have attempted to outline in this chapter. We would encourage researchers who are interested in applying the method to their specific question to think very carefully about the issues we have discussed here.

REFERENCES

Apel R, Bushway SD, Brame R, Haviland AM, Nagin DS, Paternoster R (2007) Unpacking the relationship between adolescent employment and antisocial behavior: a matched samples comparison. Criminology 45:67–97

Apel R, Bushway SD, Paternoster R, Brame R, Sweeten G (2008) Using state child labor laws to identify the causal effect of youth employment on deviant behavior and academic achievement. J Quant Criminol 24:337–362

Apel R, Paternoster R, Bushway SD, Brame R (2006) A job isn't just a job: the differential impact of formal versus informal work on adolescent problem behavior. Crime Delinq 52:333–369

Bachman JG, Johnston LD, O'Malley PM (1981) Smoking, drinking, and drug use among American high school students: correlates and trends, 1975–1979. Am J Public Health 71:59–69

Bachman JG, Safron DJ, Sy SR, Schulenberg JE (2003) Wishing to work: new perspectives on how adolescents' part-time work intensity is linked to educational disengagement, substance use, and other problem behaviors. Int J Behav Dev 27:301–315

Bachman JG Schulenberg JE (1993) How part-time work intensity relates to drug use, problem behavior, time use, and satisfaction among high school seniors: are these consequences or merely correlates? Dev Psychol 29:220–235

Banks D, Gottfredson DC (2003) The effects of drug treatment and supervision on time to rearrest among drug treatment court participants. J Drug Issues 33:385–412

Berk RA, Newton PJ (1985) Does arrest really deter wife battery? An effort to replicate the findings of the Minneapolis spouse abuse experiment. Am Sociol Rev 50:253–262

Bingenheimer JB, Brennan RT, Earls FJ (2005) Firearm violence exposure and serious violent behavior. Science 308:1323–1326

Blechman EA, Maurice A, Bueckner B, Helberg C (2000) Can mentoring or skill training reduce recidivism? Observational study with propensity analysis. Prev Sci 1:139–155

Brame R, Bushway SD, Paternoster R, Apel R (2004) Assessing the effect of adolescent employment on involvement in criminal activity. J Contemp Crim Justice 20:236–256

Caldwell M, Skeem J, Salekin R, Rybroek GV (2006) Treatment response of adolescent offenders with psychopathy features: a 2-year follow-up. Crim Justice Behav 33:571–596

Cameron AC, Trivedi PK (2005) Microeconometrics: methods and applications. Cambridge University Press, New York

Cochran WG (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 24:295–313

Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd edn. Lawrence Erlbaum, Hillsdale, NJ

Dehejia RH, Wahba S (1999) Causal effects in nonexperimental settings: reevaluating the evaluation of training programs. J Am Stat Assoc 94:1053–1062

Dehejia RH, Wahba S (2002) Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat 84:151–161

Glueck S, Glueck E (1950) Unraveling juvenile delinquency. The Commonwealth Fund, Cambridge, MA

Gottfredson DC (1985) Youth employment, crime, and schooling: a longitudinal study of a national sample. Dev Psychol 21:419–432

Greenberger E, Steinberg LD, Vaux A (1981) Adolescents who work: health and behavioral consequences of job stress. Dev Psychol 17:691–703

Haviland AM, Nagin DS (2005) Causal inferences with group based trajectory models. Psychometrika 70:1–22

Haviland AM, Nagin DS (2007) Using group-based trajectory modeling in conjunction with propensity scores to improve balance. J Exp Criminol 3:65–82

Haviland AM, Nagin DS, Rosenbaum PR (2007) Combining propensity score matching and group-based trajectory analysis in an observational study. Psychol Methods 12:247–267

Heckman JJ, Joseph Hotz V (1989) Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training. J Am Stat Assoc 84:862–874

Hirano K, Imbens GW, Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71:1161–1189

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

Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econ Stat 86:4–29

Johnson MK (2004) Further evidence on adolescent employment and substance use: differences by race and ethnicity. J Health Soc Behav 45:187–197

King RD, Massoglia M, MacMillan R (2007) The context of marriage and crime: gender, the propensity to marry, and offending in early adulthood. Criminology 45:33–65

Krebs CP, Strom KJ, Koetse WH, Lattimore PK (2009) The impact of residential and nonresidential drug treatment on recidivism among drug-involved probationers. Crime Delinq 55:442–471

Leeb RT, Barker LE, Strine TW (2007) The effect of childhood physical and sexual abuse on adolescent weapon carrying. J Adolesc Health 40:551–558

Leuven E, Barbara S (2003) PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Available online: http://ideas.repec.org/c/boc/bocode/s432001.html

Li YP, Propert KJ, Rosenbaum PR (2001) Balanced risk set matching. J Am Stat Assoc 96:870–882

Lu B (2005) Propensity score matching with time-dependent covariates. Biometrics 61:721–728

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

McMorris BJ Uggen C (2000) Alcohol and employment in the transition to adulthood. J Health Soc Behav 41: 276–294

McNeil DE, Binder RL (2007) Effectiveness of a mental health court in reducing criminal recidivism and violence. Am J Psychiatry 164:1395–1403

Mihalic SW Elliott DS (1997) Short- and long-term consequences of adolescent work. Youth Soc 28:464–498

Mocan NH, Tekin E (2006) Catholic schools and bad behavior: a propensity score matching analysis. J Econom Anal Policy 5:1–34

Molnar BE, Browne A, Cerda M, Buka SL (2005) Violent behavior by girls reporting violent victimization: a prospective study. Arch Pediatr Adolesc Med 159:731–739

Morgan SL, Harding DJ (2006) Matching estimators of causal effects: prospects and pitfalls in theory and practice. Sociol Methods Res 35:3–60

Mortimer JT (2003) Working and growing up in America. Harvard University Press, Cambridge, MA

Mortimer JT, Finch MD, Ryu S, Shanahan MJ, Call KT (1996) The effects of work intensity on adolescent mental health, achievement, and behavioral adjustment: new evidence from a prospective study. Child Dev 67: 1243–1261

Nagin DS (2005) Group-based modeling of development. Harvard University Press, Cambridge, MA

Nieuwbeerta P, Nagin DS, Blokland AAJ (2009) Assessing the impact of first-time imprisonment on offenders' subsequent criminal career development: A matched samples comparison. J Quant Criminol 25:227–257

Paternoster R, Bushway S, Brame R, Apel R (2003) The effect of teenage employment on delinquency and problem behaviors. Soc Forces 82:297–335

Ploeger M (1997) Youth employment and delinquency: reconsidering a problematic relationship. Criminology 35:659–675

Resnick MD, Bearman PS, Blum RW, Bauman KE, Harris KM, Jo J, Tabor J, Beuhring T, Sieving RE, Shew M, Ireland M, Bearinger LH, Richard Udry J (1997) Protecting adolescents from harm: findings from the national longitudinal study of adolescent health. J Am Med Assoc 278:823–832

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

Robins JM (1999) Association, causation, and marginal structural models. Synthese 121:151–179

Robins JM, Rotnitzky A (1995) Semiparametric efficiency in multivariate regression models with missing data. J Am Stat Assoc 90:122–129

Robins JM, Mark SD, Newey WK (1992) Estimating exposure effects by modeling the expectation of exposure conditional on confounders. Biometrics 48:479–495

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

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

Rosenbaum PR, Rubin DB (1984) Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 79:516–524

Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat 39:33–38

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

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

Safron DJ, Schulenberg JE Bachman JG (2001) Part-time work and hurried adolescence: the links among work intensity, social activities, health behaviors, and substance use. J Health Soc Behav 42:425–449

Sampson RJ, Laub JH, Wimer C (2006) Does marriage reduce crime? A counterfactual approach to within-individual causal effects. Criminology 44:465–508

Smith JA, Todd PE (2005) Does matching overcome lalonde's critique of nonexperimental estimators? J Econom 125:305–353

Staff J Uggen C (2003) The fruits of good work: early work experiences and adolescent deviance. J Res Crime Delinq 40:263–290

Steinberg L Dornbusch S (1991) Negative correlates of part-time work in adolescence: replication and elaboration. Dev Psychol 17:304–313

Steinberg L, Fegley S, Dornbusch S (1993) Negative impact of part-time work on adolescent adjustment: evidence from a longitudinal study. Dev Psychol 29:171–180

Steinberg LD, Greenberger E, Garduque L, Ruggiero M, Vaux A (1982) Effects of working on adolescent development. Dev Psychol 18:385–395

Sweeten G, Apel R (2007) Incapacitation: revisiting an old question with a new method and new data. J Quant Criminol 23:303–326

Tanner J, Krahn H (1991) Part-time work and deviance among high-school seniors. Can J Sociol 16:281–302

Tita G, Ridgeway G (2007) The impact of gang formation on local patterns of crime. J Res Crime Delinq 44:208–237

Widom CS (1989) The cycle of violence. Science 244:160–166

Wooldridge JM (2002) Econometric analysis of cross section and panel data. MIT Press, Cambridge, MA