4. Propensity Score Matching

Summary

Propensity score matching (PSM) constructs a statistical comparison group that is based on a model of the probability of participating in the treatment, using observed characteristics. Participants are then matched on the basis of this probability, or *propensity score*, to nonparticipants. The average treatment effect of the program is then calculated as the mean difference in outcomes across these two groups. The validity of PSM depends on two conditions: (a) conditional independence (namely, that unobserved factors do not affect participation) and (b) sizable common support or overlap in propensity scores across the participant and nonparticipant samples.

Different approaches are used to match participants and nonparticipants on the basis of the propensity score. They include nearest-neighbor (NN) matching, caliper and radius matching, stratification and interval matching, and kernel matching and local linear matching (LLM). Regression-based methods on the sample of participants and nonparticipants, using the propensity score as weights, can lead to more efficient estimates.

On its own, PSM is a useful approach when only observed characteristics are believed to affect program participation. Whether this belief is actually the case depends on the unique features of the program itself, in terms of targeting as well as individual takeup of the program. Assuming selection on observed characteristics is sufficiently strong to determine program participation, baseline data on a wide range of preprogram characteristics will allow the probability of participation based on observed characteristics to be specified more precisely. Some tests can be conducted to assess the degree of selection bias or participation on unobserved characteristics.

Learning Objectives

After completing this chapter, the reader will be able to discuss

- Calculation of the propensity score and underlying assumptions needed to apply PSM
- Different methods for matching participants and nonparticipants in the area of common support
- Drawbacks of PSM and methods to assess the degree of selection bias on unobserved characteristics
- Use of PSM in regression-based methods

PSM and Its Practical Uses

Given concerns with the implementation of randomized evaluations, the approach is still a perfect impact evaluation method in theory. Thus, when a treatment cannot be randomized, the next best thing to do is to try to mimic randomization—that is, try to have an observational analogue of a randomized experiment. With matching methods, one tries to develop a counterfactual or control group that is as similar to the treatment group as possible in terms of *observed* characteristics. The idea is to find, from a large group of nonparticipants, individuals who are *observationally similar* to participants in terms of characteristics not affected by the program (these can include preprogram characteristics, for example, because those clearly are not affected by subsequent program participation). Each participant is matched with an observationally similar nonparticipant, and then the average difference in outcomes across the two groups is compared to get the program treatment effect. If one assumes that differences in participation are based solely on differences in observed characteristics, and if enough nonparticipants are available to match with participants, the corresponding treatment effect can be measured even if treatment is not random.

The problem is to credibly identify groups that look alike. Identification is a problem because even if households are matched along a vector, X, of different characteristics, one would rarely find two households that are exactly similar to each other in terms of many characteristics. Because many possible characteristics exist, a common way of matching households is propensity score matching. In PSM, each participant is matched to a nonparticipant on the basis of a single propensity score, reflecting the probability of participating conditional on their different observed characteristics X (see Rosenbaum and Rubin 1983). PSM therefore avoids the "curse of dimensionality" associated with trying to match participants and nonparticipants on every possible characteristic when X is very large.

What Does PSM Do?

PSM constructs a statistical comparison group by modeling the probability of participating in the program on the basis of observed characteristics unaffected by the program. Participants are then matched on the basis of this probability, or propensity score, to nonparticipants, using different methods outlined later in the chapter. The average treatment effect of the program is then calculated as the mean difference in outcomes across these two groups. On its own, PSM is useful when only observed characteristics are believed to affect program participation. This assumption hinges on the rules governing the targeting of the program, as well as any factors driving self-selection of individuals or households into the program. Ideally, if available, pre-program baseline data on participants and nonparticipants can be used to calculate the propensity score and to match the two groups on the basis of the propensity score.

Selection on observed characteristics can also help in designing multiwave experiments. Hahn, Hirano, and Karlan (2008) show that available data on covariates for individuals targeted by an experiment, say in the first stage of a two-stage intervention, can be used to choose a treatment assignment rule for the second stage—conditioned on observed characteristics. This equates to choosing the propensity score in the second stage and allows more efficient estimation of causal effects.

PSM Method in Theory

The PSM approach tries to capture the effects of different observed covariates X on participation in a single propensity score or index. Then, outcomes of participating and nonparticipating households with similar propensity scores are compared to obtain the program effect. Households for which no match is found are dropped because no basis exists for comparison.

PSM constructs a statistical comparison group that is based on a model of the probability of participating in the treatment T conditional on observed characteristics X, or the propensity score: $P(X) = \Pr(T=1|X)$. Rosenbaum and Rubin (1983) show that, under certain assumptions, matching on $P(X)$ is as good as matching on X. The necessary assumptions for identification of the program effect are (a) conditional independence and (b) presence of a common support. These assumptions are detailed in the following sections.

Also, as discussed in chapters 2 and 3, the treatment effect of the program using these methods can either be represented as the average treatment effect (ATE) or the treatment effect on the treated (TOT). Typically, researchers and evaluators can ensure only internal as opposed to external validity of the sample, so only the TOT can be estimated. Weaker assumptions of conditional independence as well as common support apply to estimating the TOT and are also discussed in this chapter.

Assumption of Conditional Independence

Conditional independence states that given a set of observable covariates X that are not affected by treatment, potential outcomes Y are independent of treatment assignment T. If Y_i^T represent outcomes for participants and Y_i^C outcomes for nonparticipants, conditional independence implies

$$(Y_i^T, Y_i^C) \perp T_i \mid X_i. \tag{4.1}$$

This assumption is also called *unconfoundedness* (Rosenbaum and Rubin 1983), and it implies that uptake of the program is based entirely on observed characteristics. To estimate the TOT as opposed to the ATE, a weaker assumption is needed:

$$Y_i^C \perp T_i \mid X_i. \tag{4.2}$$

Conditional independence is a strong assumption and is not a directly testable criterion; it depends on specific features of the program itself. If unobserved characteristics determine program participation, conditional independence will be violated, and PSM is not an appropriate method.[1] Chapters 5 to 9 discuss approaches when unobserved selection is present. Having a rich set of preprogram data will help support the conditional independence assumption by allowing one to control for as many observed characteristics as might be affecting program participation (assuming unobserved selection is limited). Alternatives when selection on unobserved characteristics exists, and thus conditional independence is violated, are discussed in the following chapters, including the instrumental variable and double-difference methods.

Assumption of Common Support

A second assumption is the *common support* or *overlap condition*: $0 < P(T_i = 1|X_i) < 1$. This condition ensures that treatment observations have comparison observations "nearby" in the propensity score distribution (Heckman, LaLonde, and Smith 1999). Specifically, the effectiveness of PSM also depends on having a large and roughly equal number of participant and nonparticipant observations so that a substantial region of common support can be found. For estimating the TOT, this assumption can be relaxed to $P(T_i = 1|X_i) < 1$.

Treatment units will therefore have to be similar to nontreatment units in terms of observed characteristics unaffected by participation; thus, some nontreatment units may have to be dropped to ensure comparability. However, sometimes a nonrandom subset of the treatment sample may have to be dropped if similar comparison units do not exist (Ravallion 2008). This situation is more problematic because it creates a possible sampling bias in the treatment effect. Examining the characteristics of dropped units may be useful in interpreting potential bias in the estimated treatment effects.

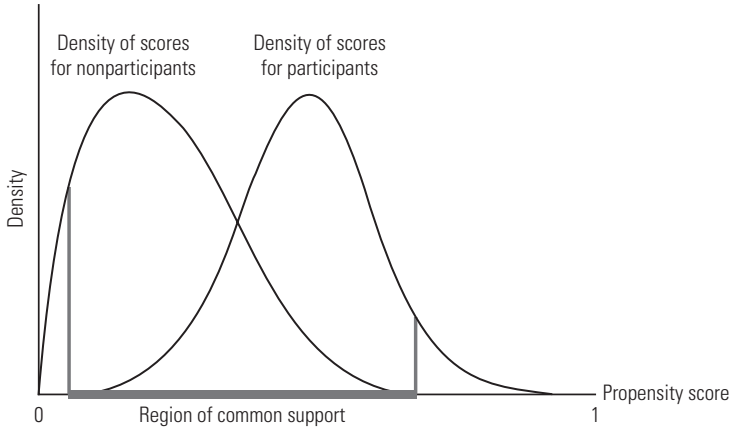
Heckman, Ichimura, and Todd (1997) encourage dropping treatment observations with weak common support. Only in the area of common support can inferences be made about causality, as reflected in figure 4.1. Figure 4.2 reflects a scenario where the common support is weak.

The TOT Using PSM

If conditional independence holds, and if there is a sizable overlap in $P(X)$ across participants and nonparticipants, the PSM estimator for the TOT can be specified as the mean difference in Y over the common support, weighting the comparison units by the propensity score distribution of participants. A typical cross-section estimator can be specified as follows:

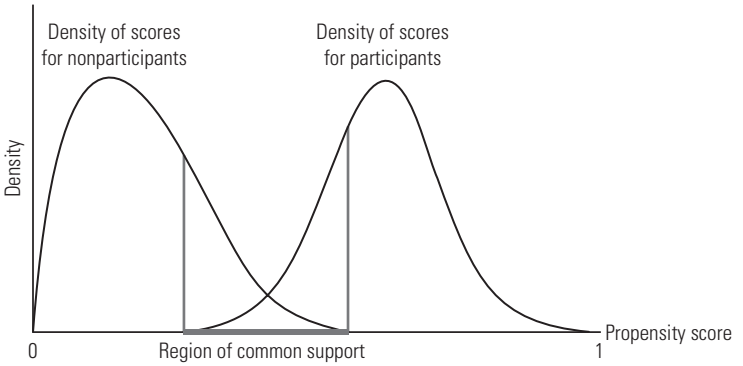
$$\text{TOT}_{\text{PSM}} = E_{P(X)\,|\,T=1}\{E[Y^T \mid T=1, P(X)] - E[Y^C \mid T=0, P(X)]\}. \qquad (4.3)$$

Figure 4.1 Example of Common Support



Source: Authors' representation.

Figure 4.2 Example of Poor Balancing and Weak Common Support



Source: Authors' representation.

More explicitly, with cross-section data and within the common support, the treatment effect can be written as follows (see Heckman, Ichimura, and Todd 1997; Smith and Todd 2005):

$$\Rightarrow \mathrm{TOT}_{\mathrm{PSM}} = \frac{1}{N_T}\left[\sum_{i \in T} Y_i^T - \sum_{j \in C} \omega(i,j) Y_j^C\right] \tag{4.4}$$

where N_T is the number of participants i and $\omega(i, j)$ is the weight used to aggregate outcomes for the matched nonparticipants j.[2]

Application of the PSM Method

To calculate the program treatment effect, one must first calculate the propensity score $P(X)$ on the basis all observed covariates X that jointly affect participation and the outcome of interest. The aim of matching is to find the closest comparison group from a sample of nonparticipants to the sample of program participants. "Closest" is measured in terms of observable characteristics not affected by program participation.

Step 1: Estimating a Model of Program Participation

First, the samples of participants and nonparticipants should be pooled, and then participation T should be estimated on all the observed covariates X in the data that are likely to determine participation. When one is interested only in comparing outcomes for those participating ($T = 1$) with those not participating ($T = 0$), this estimate can be constructed from a probit or logit model of program participation. Caliendo and Kopeinig (2008) also provide examples of estimations of the participation equation with a nonbinary treatment variable, based on work by Bryson, Dorsett, and Purdon (2002); Imbens (2000); and Lechner (2001). In this situation, one can use a multinomial probit (which is computationally intensive but based on weaker assumptions than the multinomial logit) or a series of binomial models.

After the participation equation is estimated, the predicted values of T from the participation equation can be derived. The predicted outcome represents the estimated probability of participation or propensity score. Every sampled participant and nonparticipant will have an estimated propensity score, $\hat{P}(X|T = 1) = \hat{P}(X)$. Note that the participation equation is not a determinants model, so estimation outputs such as t-statistics and the adjusted R^2 are not very informative and may be misleading. For this stage of PSM, causality is not of as much interest as the correlation of X with T.

As for the relevant covariates X, PSM will be biased if covariates that determine participation are not included in the participation equation for other reasons. These reasons could include, for example, poor-quality data or poor understanding of the local context in which the program is being introduced. As a result, limited guidance exists on how to select X variables using statistical tests, because the observed characteristics that are more likely to determine participation are likely to be data driven and context specific.[3] Heckman, Ichimura, and Todd (1997, 1998) show that the bias in PSM program estimates can be low, given three broad provisions. First, if possible, the same survey instrument or source of data should be used for participants and nonparticipants. Using the same data source helps ensure that the observed characteristics entering the logit or probit model of participation are measured similarly across the two groups and thereby reflect the same concepts. Second, a representative sample survey of eligible nonparticipants as well as participants can greatly improve the precision of the propensity score. Also, the larger the sample of eligible nonparticipants is, the more good matching will be facilitated. If the two samples come from different surveys,

then they should be highly comparable surveys (same questionnaire, same interviewers or interviewer training, same survey period, and so on). A related point is that participants and nonparticipants should be facing the same economic incentives that might drive choices such as program participation (see Ravallion 2008; such incentives might include access to similar markets, for example). One could account for this factor by choosing participants and nonparticipants from the same geographic area.

Nevertheless, including too many X variables in the participation equation should also be avoided; overspecification of the model can result in higher standard errors for the estimated propensity score $\hat{P}(X)$ and may also result in perfectly predicting participation for many households ($\hat{P}(X) = 1$). In the latter case, such observations would drop out of the common support (as discussed later). As mentioned previously, determining participation is less of an issue in the participating equation than obtaining a distribution of participation probabilities.

Step 2: Defining the Region of Common Support and Balancing Tests

Next, the region of common support needs to be defined where distributions of the propensity score for treatment and comparison group overlap. As mentioned earlier, some of the nonparticipant observations may have to be dropped because they fall outside the common support. Sampling bias may still occur, however, if the dropped nonparticipant observations are systematically different in terms of observed characteristics from the retained nonparticipant sample; these differences should be monitored carefully to help interpret the treatment effect.

Balancing tests can also be conducted to check whether, within each quantile of the propensity score distribution, the average propensity score and mean of X are the same. For PSM to work, the treatment and comparison groups must be balanced in that similar propensity scores are based on similar observed X. Although a treated group and its matched nontreated comparator might have the same propensity scores, they are not necessarily observationally similar if misspecification exists in the participation equation. The distributions of the treated group and the comparator must be similar, which is what balance implies. Formally, one needs to check if $\hat{P}(X|T=1) = \hat{P}(X|T=0)$.

Step 3: Matching Participants to Nonparticipants

Different matching criteria can be used to assign participants to non-participants on the basis of the propensity score. Doing so entails calculating a weight for each matched participant-nonparticipant set. As discussed below, the choice of a particular matching technique may therefore affect the resulting program estimate through the weights assigned:

- *Nearest-neighbor matching.* One of the most frequently used matching techniques is NN matching, where each treatment unit is matched to the comparison unit with the closest propensity score. One can also choose n nearest neighbors and do matching (usually $n=5$ is used). Matching can be done with or without

replacement. Matching with replacement, for example, means that the same nonparticipant can be used as a match for different participants.

- *Caliper or radius matching.* One problem with NN matching is that the difference in propensity scores for a participant and its closest nonparticipant neighbor may still be very high. This situation results in poor matches and can be avoided by imposing a threshold or "tolerance" on the maximum propensity score distance (*caliper*). This procedure therefore involves matching with replacement, only among propensity scores within a certain range. A higher number of dropped nonparticipants is likely, however, potentially increasing the chance of sampling bias.
- *Stratification or interval matching.* This procedure partitions the common support into different strata (or intervals) and calculates the program's impact within each interval. Specifically, within each interval, the program effect is the mean difference in outcomes between treated and control observations. A weighted average of these interval impact estimates yields the overall program impact, taking the share of participants in each interval as the weights.
- *Kernel and local linear matching.* One risk with the methods just described is that only a small subset of nonparticipants will ultimately satisfy the criteria to fall within the common support and thus construct the counterfactual outcome. Nonparametric matching estimators such as kernel matching and LLM use a weighted average of all nonparticipants to construct the counterfactual match for each participant. If P_i is the propensity score for participant i and P_j is the propensity score for nonparticipant j, and if the notation in equation 4.4 is followed, the weights for kernel matching are given by

$$\omega(i,j)_{KM} = \frac{K\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in C} K\left(\frac{P_k - P_i}{a_n}\right)}, \tag{4.5}$$

where $K(\cdot)$ is a kernel function and a_n is a bandwidth parameter. LLM, in contrast, estimates a nonparametric locally weighted (*lowess*) regression of the comparison group outcome in the neighborhood of each treatment observation (Heckman, Ichimura, and Todd 1997). Kernel matching is analogous to regression on a constant term, whereas LLM uses a constant and a slope term, so it is "linear." LLM can include a faster rate of convergence near boundary points (see Fan 1992, 1993). The LLM estimator has the same form as the kernel-matching estimator, except for the weighting function:

$$\omega(i,j)_{LLR} = \frac{K_{ij}\sum_{k \in C} K_{ik}(P_k - P_i)^2 - \left[K_{ij}(P_j - P_i)\right]\sum_{k \in C} K_{ik}(P_k - P_i)}{\sum_{j \in C} K_{ij}\sum_{k \in C} K_{ik}(P_k - P_i)^2 - \left(\sum_{k \in C} K_{ik}(P_k - P_i)\right)^2}. \tag{4.6}$$

- *Difference-in-difference matching.* With data on participant and control observations before and after program intervention, a difference-in-difference (DD) matching estimator can be constructed. The DD approach is discussed in greater detail in chapter 5; importantly, it allows for unobserved characteristics affecting program take-up, assuming that these unobserved traits do not vary over time. To present the DD estimator, revisit the setup for the cross-section PSM estimator given in equation 4.4. With panel data over two time periods $t = \{1,2\}$, the local linear DD estimator for the mean difference in outcomes Y_{it} across participants i and nonparticipants j in the common support is given by

$$\text{TOT}_{\text{PSM}}^{\text{DD}} = \frac{1}{N_T}\left[\sum_{i \in T}(Y_{i2}^T - Y_{i1}^T) - \sum_{j \in C}\omega(i,j)(Y_{j2}^C - Y_{j1}^C)\right]. \quad (4.7)$$

With only cross-sections over time rather than panel data (see Todd 2007), $\text{TOT}_{\text{PSM}}^{\text{DD}}$ can be written as

$$\text{TOT}_{\text{PSM}}^{\text{DD}} = \frac{1}{N_{T_2}}\left[\sum_{i \in T_2} Y_{i2}^T - \sum_{j \in C_2}\omega(i,j)\,Y_{j2}^C\right] - \frac{1}{N_{T_1}}\left[\sum_{i \in T_1} Y_{i1}^T - \sum_{j \in C_1}\omega(i,j)\,Y_{j1}^C\right]. \quad (4.8)$$

Here, Y_{it}^T and Y_{jt}^C, $t = \{1,2\}$ are the outcomes for different participant and nonparticipant observations in each time period t. The DD approach combines traditional PSM and DD approaches discussed in the next chapter. Observed as well as unobserved characteristics affecting participation can thus be accounted for if unobserved factors affecting participation are assumed to be constant over time. Taking the difference in outcomes over time should also difference out time-invariant unobserved characteristics and thus potential unobserved selection bias. Again, chapter 5 discusses this issue in detail. One can also use a regression-adjusted estimator (described in more detail later in this chapter as well as in chapter 5). This method assumes using a standard linear model for outcomes and for estimating the TOT (such as $Y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i$) and applying weights on the basis of the propensity score to the matched comparison group. It can also allow one to control for selection on unobserved characteristics, again assuming these characteristics do not vary over time.

A number of steps, therefore, can be used to match participants to nonparticipants. Comparing results across different matching methods can reveal whether the estimated program effect is robust. Box 4.1 describes some of these methods, from a study on the impact of a pilot farmer-field-school (FFS) program in Peru on farmers' knowledge of pest management practices related to potato cultivation (Godtland and

BOX 4.1 Case Study: Steps in Creating a Matched Sample of Nonparticipants to Evaluate a Farmer-Field-School Program

A farmer-field-school program was started in 1998 by scientists in collaboration with CARE-Peru. In their study of the program, Godtland and others (2004) applied three different steps for generating a common support of propensity scores to match nonparticipants to the participant sample. These steps, as described here, combined methods that have been formally discussed in the PSM literature and informal rules commonly applied in practice.

First, a propensity score cutoff point was chosen, above which all households were included in the comparison group. No formal rule exists for choosing this cutoff point, and Godtland and others used as a benchmark the average propensity score among participants of 0.6. Second, the comparison group was chosen, using a nearest-neighbor matching method, matching to each participant five nonparticipants with the closest value of the propensity score (within a proposed 0.01 bound). Matches not in this range were removed from the sample. As a third approach, the full sample of nonparticipants (within the common support) was used to construct a weighted match for each participant, applying a nonparametric kernel regression method proposed by Heckman, Ichimura, and Todd (1998).

To evaluate the comparability of the participant and matched nonparticipant samples across these three methods, Godtland and others (2004) conducted balancing tests to see whether the means of the observable variables for each group were significantly different. For the first and second methods, the balancing test was performed by dividing each comparison and treatment group into two strata, ordered by probability propensity scores. Within each stratum, a *t*-test of equality of means across participants and matched nonparticipants was conducted for each variable in the farmer participation equation. Godtland and others found that the null was not rejected for all but a few variables across the first two methods. For the third method, a test for the equality of means was conducted across the samples of participants and their weighted matches. The null was not rejected for all but two variables at the 10 percent level. Overall, their results found no systematic differences in observed characteristics across the participant and nonparticipant samples.

others 2004). Farmers self-selected into the program. The sample of nonparticipants was drawn from villages where the FFS program existed, villages without the FFS program but with other programs run by CARE-Peru, as well as control villages. The control villages were chosen to be similar to the FFS villages across such observable characteristics as climate, distance to district capitals, and infrastructure. Simple comparison of knowledge levels across participants and nonparticipants would yield biased estimates of the program effect, however, because the program was not randomized and farmers were self-selecting into the program potentially on the basis of observed characteristics. Nonparticipants would therefore need to be matched to participants over a set of common characteristics to ensure comparability across the two groups.

Calculating the Average Treatment Impact

As discussed previously, if conditional independence and a sizable overlap in propensity scores between participants and matched nonparticipants can be assumed, the PSM average treatment effect is equal to the mean difference in outcomes over the common support, weighting the comparison units by the propensity score distribution of participants. To understand the potential observed mechanisms driving the estimated program effect, one can examine the treatment impact across different observable characteristics, such as position in the sample distribution of income, age, and so on.

Estimating Standard Errors with PSM: Use of the Bootstrap

Compared to traditional regression methods, the estimated variance of the treatment effect in PSM should include the variance attributable to the derivation of the propensity score, the determination of the common support, and (if matching is done without replacement) the order in which treated individuals are matched (Caliendo and Kopeinig 2008). Failing to account for this additional variation beyond the normal sampling variation will cause the standard errors to be estimated incorrectly (see Heckman, Ichimura, and Todd 1998).

One solution is to use bootstrapping (Efron and Tibshirani 1993; Horowitz 2003), where repeated samples are drawn from the original sample, and properties of the estimates (such as standard error and bias) are reestimated with each sample. Each bootstrap sample estimate includes the first steps of the estimation that derive the propensity score, common support, and so on. Formal justification for bootstrap estimators is limited; however, because the estimators are asymptotically linear, bootstrapping will likely lead to valid standard errors and confidence intervals (Imbens 2004).

Critiquing the PSM Method

The main advantage (and drawback) of PSM relies on the degree to which observed characteristics drive program participation. If selection bias from unobserved characteristics is likely to be negligible, then PSM may provide a good comparison with randomized estimates. To the degree participation variables are incomplete, the PSM results can be suspect. This condition is, as mentioned earlier, not a directly testable criteria; it requires careful examination of the factors driving program participation (through surveys, for example).

Another advantage of PSM is that it does not necessarily require a baseline or panel survey, although in the resulting cross-section, the observed covariates entering the

logit model for the propensity score would have to satisfy the conditional independence assumption by reflecting observed characteristics *X* that are not affected by participation. A preprogram baseline is more helpful in this regard, because it covers observed *X* variables that are independent of treatment status. As discussed earlier, data on participants and nonparticipants over time can also help in accounting for some unobserved selection bias, by combining traditional PSM approaches with DD assumptions detailed in chapter 5.

PSM is also a semiparametric method, imposing fewer constraints on the functional form of the treatment model, as well as fewer assumptions about the distribution of the error term. Although observations are dropped to achieve the common support, PSM increases the likelihood of sensible comparisons across treated and matched control units, potentially lowering bias in the program impact. This outcome is true, however, only if the common support is large; sufficient data on nonparticipants are essential in ensuring a large enough sample from which to draw matches. Bias may also result from dropping nonparticipant observations that are systematically different from those retained; this problem can also be alleviated by collecting data on a large sample of nonparticipants, with enough variation to allow a representative sample. Otherwise, examining the characteristics of the dropped nonparticipant sample can refine the interpretation of the treatment effect.

Methods to address potential selection bias in PSM program estimates are described in a study conducted by Jalan and Ravallion (2003) in box 4.2. Their study estimates the net income gains of the Trabajar workfare program in Argentina (where participants must engage in work to receive benefits) during the country's economic crisis in 1997. The average income benefit to participants from the program is muddled by the fact that participants need not have been unemployed prior to joining Trabajar. Measurement of forgone income and, hence, construction of a proper counterfactual were therefore important in this study. Neither a randomized methodology nor a baseline survey was available, but Jalan and Ravallion were able to construct the counterfactual using survey data conducted about the same time covering a large sample of nonparticipants.

PSM and Regression-Based Methods

Given that matching produces consistent estimates under weak conditions, a practical advantage of PSM over ordinary least squares (OLS) is that it reduces the number of dimensions on which to match participants and comparison units. Nevertheless, consistent OLS estimates of the ATE can be calculated under the assumption of conditional exogeneity. One approach suggested by Hirano, Imbens, and Ridder (2003) is to estimate a weighted least squares regression of the outcome on treatment *T* and other observed covariates *X* unaffected by participation, using the inverse of

BOX 4.2 Case Study: Use of PSM and Testing for Selection Bias

In their study of the Trabajar workfare program in Argentina, Jalan and Ravallion (2003) conducted a postintervention survey of both participants and nonparticipants. The context made it more likely that both groups came from a similar economic environment: 80 percent of Trabajar workers came from the poorest 20 percent of the population, and the study used a sample of about 2,800 Trabajar participants along with nonparticipants from a large national survey.

Kernel density estimation was used to match the sample of participants and nonparticipants over common values of the propensity scores, excluding nonparticipants for whom the estimated density was equal to zero, as well as 2 percent of the nonparticipant sample from the top and bottom of the distribution. Estimates of the average treatment effect based on the nearest neighbor, the nearest five neighbors, and a kernel-weighted matching were constructed, and average gains of about half the maximum monthly Trabajar wage of US$200 were realized.

Jalan and Ravallion (2003) also tested for potential remaining selection bias on unobserved characteristics by applying the Sargan-Wu-Hausman test. Specifically, on the sample of participants and matched nonparticipants, they ran an ordinary least squares regression of income on the propensity score, the residuals from the logit participation equation, as well as a set of additional control variables Z that exclude the instruments used to identify exogenous variation in income gains. In the study, the identifying instruments were provincial dummies, because the allocations from the program varied substantially across equally poor local areas but appeared to be correlated with the province that the areas belonged to. This test was used to detect selection bias in the nearest-neighbor estimates, where one participant was matched to one nonparticipant, which lent itself to a comparable regression-based approach.

If the coefficient on the residuals is significantly different from zero, selection bias may continue to pose a problem in estimating the program's impact. In the analysis, this coefficient was not statistically significant under the null hypothesis of no selection bias, and the coefficient on the propensity score was similar to the average impact in the nearest-neighbor matching estimate.

a nonparametric estimate of the propensity score. This approach leads to a fully efficient estimator, and the treatment effect is estimated by $Y_{it} = \alpha + \beta T_{i1} + \gamma X_{it} + \varepsilon_{it}$ with weights of 1 for participants and weights of $\hat{P}(X)/(1 - \hat{P}(X))$ for the control observations. T_{i1} is the treatment indicator, and the preceding specification attempts to account for latent differences across treatment and comparison units that would affect selection into the program as well as resulting outcomes. For an estimate of the ATE for the population, the weights would be $1/\hat{P}(X)$ for the participants and $1/(1 - \hat{P}(X))$ for the control units.

Box 4.3, based on a study conducted by Chen, Mu, and Ravallion (2008) on the effects of the World Bank–financed Southwest China Poverty Reduction Project, describes an application of this approach. It allows the consistency advantages of matching to be combined with the favorable variance properties of regression-based methods.

BOX 4.3 Case Study: Using Weighted Least Squares Regression in a Study of the Southwest China Poverty Reduction Project

The Southwest China Poverty Reduction Project (SWP) is a program spanning interventions across a range of agricultural and nonagricultural activities, as well as infrastructure development and social services. Disbursements for the program covered a 10-year period between 1995 and 2005, accompanied by surveys between 1996 and 2000 of about 2,000 households in targeted and nontargeted villages, as well as a follow-up survey of the same households in 2004 to 2005.

Time-varying selection bias might result in the treatment impact across participants and nonparticipants if initial differences across the two samples were substantially different. In addition to studying treatment effects based on direct propensity score matching, Chen, Mu, and Ravallion (2008) examined treatment effects constructed by OLS regressions weighted by the inverse of propensity score. As part of the analysis, they examined average treatment impacts over time and specifically used a fixed-effects specification for the weighted regression. Among the different outcomes they examined, Chen, Mu, and Ravallion found that the initial gains to project areas for such outcomes as income, consumption, and schooling diminish over the longer term (through 2004–05). For example, the SWP impact on income using the propensity score weighted estimates in the trimmed sample fell from about US$180 in 2000 (*t*-ratio: 2.54) to about US$40 in 2004 to 2005 (*t*-ratio: 0.45). Also, school enrollment of children 6 to 14 years of age improved significantly (by about 7.5 percentage points) in 2000 but fell over time to about 3 percent—although this effect was not significant—by 2004 to 2005. This outcome may have resulted from the lapse in tuition subsidies with overall program disbursements.

The methods described here, however, assume that a matched comparison unit prior to program implementation provides the counterfactual of what would have happened over time to mean outcomes for participants in the absence of treatment. If spillovers exist, the intervention changes outcomes for nonparticipants and creates an additional source of bias. Chen, Mu, and Ravallion (2008) tested for spillovers by examining non-SWP projects in nontargeted villages and found positive spillover effects on the control villages through the displacement of non-SWP spending; however, they found these spillovers were unlikely to bias the treatment effects substantially.

Notes

1. If unobserved variables indeed affect both participation and outcomes, this situation yields what is called a "hidden bias" (Rosenbaum 2002). Although the conditional independence assumption, or unconfoundedness, cannot be verified, the sensitivity of the estimated results of the PSM method can be checked with respect to deviations from this identifying assumption. In other words, even if the extent of selection or hidden bias cannot be estimated, the degree to which the PSM results are sensitive to this assumption of unconfoundedness can be tested. Box 4.2 addresses this issue.
2. As described further in the chapter, various weighting schemes are available to calculate the weighted outcomes of the matched comparators.
3. See Dehejia (2005) for some suggestions on selection of covariates.

References

Bryson, Alex, Richard Dorsett, and Susan Purdon. 2002. "The Use of Propensity Score Matching in the Evaluation of Active Labour Market Policies." Working Paper 4, Department for Work and Pensions, London.

Caliendo, Marco, and Sabine Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys* 22 (1): 31–72.

Chen, Shaohua, Ren Mu, and Martin Ravallion. 2008. "Are There Lasting Impacts of Aid to Poor Areas? Evidence for Rural China." Policy Research Working Paper 4084, World Bank, Washington, DC.

Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: A Reply to Smith and Todd." *Journal of Econometrics* 125 (1–2): 355–64.

Efron, Bradley, and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap.* Boca Raton, FL: Chapman & Hall.

Fan, Jianqing. 1992. "Design-Adaptive Nonparametric Regression." *Journal of the American Statistical Association* 87 (420): 998–1004.

———. 1993. "Local Linear Regression Smoothers and Their Minimax Efficiencies." *Annals of Statistics* 21 (1): 196–216.

Godtland, Erin, Elisabeth Sadoulet, Alain de Janvry, Rinku Murgai, and Oscar Ortiz. 2004. "The Impact of Farmer-Field-Schools on Knowledge and Productivity: A Study of Potato Farmers in the Peruvian Andes." *Economic Development and Cultural Change* 52 (1): 129–58.

Hahn, Jinyong, Keisuke Hirano, and Dean Karlan. 2008. "Adaptive Experimental Design Using the Propensity Score." Working Paper 969, Economic Growth Center, Yale University, New Haven, CT.

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme." *Review of Economic Studies* 64 (4): 605–54.

———. 1998. "Matching as an Econometric Evaluation Estimator." *Review of Economic Studies* 65 (2): 261–94.

Heckman, James J., Robert LaLonde, and Jeffrey Smith. 1999. "The Economics and Econometrics of Active Labor Market Programs." In *Handbook of Labor Economics*, vol. 3, ed. Orley Ashenfelter and David Card, 1865–2097. Amsterdam: North-Holland.

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (4): 1161–89.

Horowitz, Joel. 2003. "The Bootstrap in Econometrics." *Statistical Science* 18 (2): 211–18.

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

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

Jalan, Jyotsna, and Martin Ravallion. 2003. "Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching." *Journal of Business and Economic Statistics* 21 (1): 19–30.

Lechner, Michael. 2001. "Identification and Estimation of Causal Effects of Multiple Treatments under the Conditional Independence Assumption." In *Econometric Evaluation of Labor Market Policies*, ed. Michael Lechner and Friedhelm Pfeiffer, 43–58. Heidelberg and New York: Physica-Verlag.

Ravallion, Martin. 2008. "Evaluating Anti-Poverty Programs." In *Handbook of Development Economics*, vol. 4, ed. T. Paul Schultz and John Strauss, 3787–846. Amsterdam: North-Holland.

Rosenbaum, Paul R. 2002. *Observational Studies.* New York and Berlin: Springer-Verlag.

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

Smith, Jeffrey, and Petra Todd. 2005. "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–53.

Todd, Petra. 2007. "Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated." In *Handbook of Development Economics*, vol. 4, ed. T. Paul Schultz and John Strauss, 3847–94. Amsterdam: North-Holland.