Impact-Evaluation Guidelines
Technical Notes
No. IDB-TN-161
August 2010

A Primer for Applying Propensity-Score Matching

Carolyn Heinrich
Alessandro Maffioli
Gonzalo Vázquez

Office of Strategic Planning and Development Effectiveness, Inter-American Development Bank

A Primer for Applying Propensity-Score Matching

Impact-Evaluation Guidelines

Carolyn Heinrich
Alessandro Maffioli
Gonzalo Vázquez

Inter-American Development Bank
2010

Carolyn Heinrich. University of Wisconsin-Madison. cheinrich@lafollette.wisc.edu

Alessandro Maffioli. Inter-American Development Bank. alessandrom@iadb.org

A Primer for Applying Propensity-Score Matching

Abstract

Carolyn Heinrich* Alessandro Maffioli** Gonzalo Vázquez***

The use of microeconometric techniques to estimate the effects of development policies has become a common approach not only for scholars, but also for policy-makers engaged in designing, implementing and evaluating projects in different fields. Among these techniques, Propensity-Score Matching (PSM) is increasingly applied in the policy evaluation community. This technical note provides a guide to the key aspects of implementing PSM methodology for an audience of practitioners interested in understanding its applicability to specific evaluation problems. The note summarizes the basic conditions under which PSM can be used to estimate the impact of a program and the data required. It explains how the Conditional Independence Assumption, combined with the Overlap Condition, reduces selection bias when participation in a program is determined by observable characteristics. It also describes different matching algorithms and some tests to assess the quality of the matching. Case studies are used throughout to illustrate important concepts in impact evaluation and PSM. In the annexes, the note provides an outline of the main technical aspects and a list of statistical and econometric software for implementing PSM.

JEL Classification: C21. C40, H43, O12, O22

Keywords: Policy Evaluation, Microeconometrics, Propensity-Score Matching, Average Treatment Effect on the Treated, Development Effectiveness

* Professor of Public Affairs, La Follette School of Public Affairs, University of Wisconsin-Madison; cheinrich@lafollette.wisc.edu

** Economist – Lead Specialist, Inter-American Development Bank; alessandrom@iadb.org

*** Research Fellow, Inter-American Development Bank; gonzalovaz@iadb.org

Table of Contents

1. Introduction

Of fundamental interest in all program evaluation efforts is whether a particular intervention, as designed, is effective in accomplishing its primary objectives. A well-designed intervention (or "treatment") is typically based on theory or research evidence that articulates how the intervention's core mechanisms will work to achieve its goals and produce the desired outcomes. The main challenge of a credible impact evaluation is the construction of the counterfactual outcome, that is, what would have happened to participants in absence of treatment. Since this counterfactual outcome is never observed, it has to be estimated using statistical methods.

Experimental evaluation, in which assignment to treatment (or participation in the intervention) is random, has increasingly been encouraged and used in evaluating interventions because of its statistical advantages in identifying program impacts. Random assignment is used to assure that participation in the intervention is the only differentiating factor between units subject to the intervention and those excluded from it, so that the control group can be used to assess what would have happened to participants in the absence of the intervention.

Although random assignment is an extraordinarily valuable tool for evaluation, it is not always feasible to implement it. Not only is it costly to obtain cooperation of implementers of the intervention and study subjects, but a random assignment design must be developed and implemented prior to the start of the intervention. Considerable progress may be made, however, in understanding the effectiveness of interventions on core outcomes of interest through the application of rigorous nonexperimental evaluation methods. In addition to providing direct estimates of program effects on relevant outcomes, such methods can also address a variety of related and subsidiary questions, such as: are some interventions more effective for particular types of groups or units than others? What factors outside the control of the implementers influence outcomes, and how might the intervention be modified to account for them?

This evaluation guide focuses on a specific nonexperimental evaluation method known as *Propensity-score matching* (PSM). PSM uses information from a pool of units that do not participate in the intervention to identify what would have happened to participating units in the absence of the intervention. By comparing how outcomes differ for participants relative to observationally similar nonparticipants, it is possible to estimate the effects of the intervention. In recent years, facilitated in part by improvements in computing capacity and associated

algorithms, approaches that directly match participants with nonparticipants who have similar characteristics have replaced regression as one of the preferred methods for estimating intervention impacts using comparison group data.

The general idea of matching is straightforward. In absence of an experimental design, assignment to treatment is frequently nonrandom, and thus, units receiving treatment and those excluded from treatment may differ not only in their treatment status but also in other characteristics that affect both participation and the outcome of interest. To avoid the biases that this may generate, matching methods find a nontreated unit that is "similar" to a participating unit, allowing an estimate of the intervention's impact as the difference between a participant and the matched comparison case. Averaging across all participants, the method provides an estimate of the mean program impact for the participants.

One of the critical issues in implementing matching techniques is to define clearly (and justify) what "similar" means. Although it might be relatively simple to assign a comparison unit based on a single observable characteristic, in practice, if the matching process is to successfully mitigate potential bias, it has to be done considering a full range of covariates across which the treatment and comparison units might differ.

Propensity-score matching, one of the most important innovations in developing workable matching methods, allows this matching problem to be reduced to a single dimension. The propensity score is defined as the probability that a unit in the combined sample of treated and untreated units receives the treatment, given a set of observed variables. If all information relevant to participation and outcomes is observable to the researcher, the propensity score (or probability of participation) will produce valid matches for estimating the impact of an intervention. Therefore, rather than attempting to match on all values of the variables, cases can be compared on the basis of propensity scores alone.

The PSM technique has been applied in a very wide variety of fields in the program evaluation literature. For example, Heckman, Ichimura and Todd (1998), Lechner (1999), Dehejia and Wahba (2002), and Smith & Todd (2005) use PSM techniques to estimate the impact of labor market and training programs on income; Jalan and Ravallion (2003) evaluate antipoverty workfare programs; Galiani, Gertler and Schargrodsky (2005) study the effect of water supply on child mortality; Trujillo, Portillo and Vernon (2005) analyze the impact of health insurance on medical-care participation; Almus and Czarnitzki (2003) and Moser (2005)

evaluate the impact of research and development subsidies and patent laws on innovation; Lavy (2002) estimates the effect of teachers' performance incentives on pupil achievement; and Persson, Tabellini and Trebbi (2003) analyze the impact of electoral reform on corruption.

The main goal of this document is to provide a guideline for implementing the PSM estimator. Throughout this guide, we use case studies to illustrate important concepts in impact evaluation and PSM. The first case (Box 1) on estimating the impact of training programs for youth describes various types of evaluation questions that are frequently of interest in evaluations and explains why nonexperimental methods are often required to fully address them.

Section two provides some background on program-evaluation issues and introduces the idea of matching techniques. Section three describes in which context matching is a valid approach, considering theoretical assumptions and data-availability issues, and section four explains the basic mechanics of this technique. The main concerns in implementing matching estimators—namely, the estimation of the propensity score, the selection of a matching algorithm and the estimation of the treatment effect—are discussed in section five. Section six presents some tests to evaluate the validity of the assumptions and assess the quality of the matching. Finally, section seven discusses further issues like the calculation of standard errors and addressing some of the problems that may arise when implementing PSM techniques. Section eight concludes.

Box 1: Estimating the Impact of Training Programs for Youth

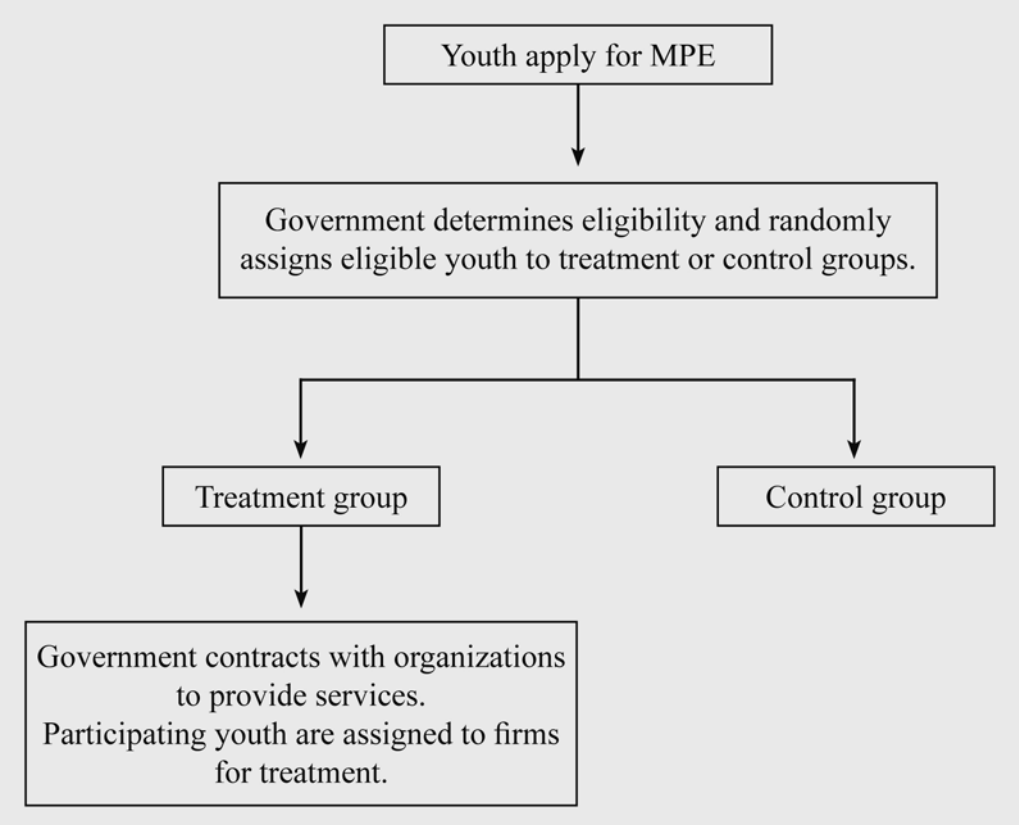
Introduction. In developing countries, it is not uncommon for youth to leave school before completing their secondary education. However, structural economic changes associated with increased economic openness have decreased labor demand in the lower-skilled, labor-intensive sectors, reducing employment among youth and widening wage gaps and economic inequality between those with higher and those with lower educational levels.

In Honduras, many poor youth have inadequate knowledge and skills to compete in the labor market. Approximately 25 percent of Honduran youth between 15-19 years of age neither attended school nor worked (the highest rate in Central America) in 2006, and those who did find a job in the labor market were often poorly prepared and/or trapped in poor-quality jobs

with low incomes. A pilot program, *Mi Primer Empleo* (MPE), was developed by the government of Honduras to promote youth employment by providing young people with life and work skills, specific job training, and an internship directly linked to the training.

Random assignment evaluation strategy. The government and pilot program funders worked to develop an impact-evaluation strategy. Because more eligible youth were expected to apply to the program than could be treated with the available funds, the government chose an **experimental**, **random assignment** strategy for enrolling eligible youth. Typical ethical concerns about excluding youth from services were mitigated by the fact that limited funds required rationing of services; random assignment would give every young applicant an equal chance of participation. Those randomly assigned to receive program services formed the *Mi Primer Empleo* treatment group, and those randomized out formed the control group (see figure B.1).

Figure B.1 *Mi Primer Empleo* – Assignment Mechanism



A primary advantage of an experimental approach to evaluation, in which assignment to participation in the intervention is random, is that it assures that participation in the intervention is the only factor that differs between those in the treatment group and those excluded from participating (the control group). In this regard, the control group serves as a perfect proxy for estimating the **counterfactual** outcome, that is, what would have happened to the treatment group in the absence of the intervention.

The average treatment effect (ATE) estimate generated by an evaluation that compares the average outcomes of *Mi Primer Empleo* participants with average outcomes of those in the control group of eligible youth will tell the government the average impact of the program on eligible youth who apply to the program (also known as the "intent to treat" estimate). This is one of the most commonly estimated impacts in random assignment evaluations. However, as in the case of many such experiments, **nonexperimental** evaluation methods would be needed to address a number of other questions of importance to the government about this program's effectiveness.

For example, not all youth who are assigned to receive treatment (i.e., to participate in *Mi Primer Empleo*) will show up. Some youth may get a job on their own, and others may decide they are not interested in the services offered. To estimate the **Average Treatment Effect on the Treated** (or ATT), analyses that compare participants to a similar group of eligible nonparticipants from the control group are necessary.

However, given that those who follow through in participating may very well be systematically different from those who are assigned to treatment but do not participate, it may not be appropriate to simply compare those randomized to treatment with those in the randomized-out control group. The voluntary nature of participation in many interventions introduces the potential for **selection bias**, where we only observe outcomes for a nonrandom subsample of all units assigned to treatment. This is an example where propensity-score matching (PSM) could be used to match participants with members of the control group who are similar in the same selective ways as those who receive services.

In addition, after being selected to participate in *Mi Primer Empleo*, youth were assigned to a service provider for structured job training in particular areas (e.g., agrobusiness, tourism, forestry, etc.). One of the government's objectives was to give priority to those

employment clusters it had identified as having particularly high potential for economic development, such as tourism, *maquiladora,* etc. If the government wanted to know if training in agrobusiness was more effective than training in tourism in connecting youth with jobs, it could not simply compare differences in outcomes between the subgroups of treatment youth assigned to the different training categories with those in the control group. Again, it would need to compare tourism trainees with similar youth in the control group. Furthermore, if the government wanted to know which agrobusiness training providers were most effective in serving the youth, they would need to use nonexperimental analyses to compare youth assigned to different agrobusiness training providers.

In the Honduras case, it was not possible to randomly assign participating youth to the different types of training and internship services, as the government and providers had to take into consideration the location of the providers relative to the youths' residences as well as the youths' specific interests in training. Thus, this case represents an example of an intervention where, even with random assignment, experimental analyses alone would not suffice to generate information important to the government in making training investment decisions.

2. Why Use Matching

The greatest challenge in evaluating any intervention or program is obtaining a credible estimate of the *counterfactual*: what would have happened to participating units if they had not participated? Without a credible answer to this question, it is not possible to determine whether the intervention actually influenced participant outcomes or is merely associated with successes (or failures) that would have occurred anyway. However, as its name implies, it is impossible to observe the counterfactual. Program evaluation faces a *missing data* problem, that the statistician Paul Holland called the *Fundamental Problem of Causal Inference*: it is impossible to observe the outcomes of the same unit in both treatment conditions at the same time (Holland, 1986).

One feasible solution to this problem is to estimate the counterfactual outcome based on a group of nonparticipants and calculate the impact of the intervention as the difference in mean outcomes between groups (see Box 2 for a brief discussion on the definition of the impact of a program). However, this approach is only valid under a very precise condition: the comparison group must be *statistically equivalent* to the treated group. In other words, the groups must be identical except for the fact that one of them received the treatment. Thus, the main concern is how to find a proper comparison group.

Box 2: Defining the Impact of a Program

The impact of a treatment for an individual i, noted δ_i, is defined as the difference between the potential outcome in case of treatment and the potential outcome in absence of treatment:

$$\delta_i = Y_{1i} - Y_{0i}$$

In general, an evaluation seeks to estimate the mean impact of the program, obtained by averaging the impact across all the individuals in the population. This parameter is known as **Average Treatment Effect** or ATE:

$$ATE = E(\delta) = E\left(Y_1 - Y_0\right)$$

where $E(.)$ represents the average (or *expected value*).

Another quantity of interest is the **Average Treatment Effect on the Treated**, or ATT, which measures the impact of the program on those individuals who participated:

$$ATT = E(Y_1 - Y_0 \mid D = 1)$$

Finally, the **Average Treatment Effect on the Untreated** (ATU) measures the impact that the program would have had on those who did not participate:

$$ATU = E(Y_1 - Y_0 \mid D = 0)$$

The problem is that all of these parameters are not observable, since they depend on counterfactual outcomes. For instance, using the fact that the average of a difference is the difference of the averages, the ATT can be rewritten as:

$$ATT = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1)$$

The second term, $E(Y_0 \mid D = 1)$, is the average outcome that the treated individuals would have obtained in absence of treatment, which is not observed. However, we do observe the term $E(Y_0 \mid D = 0)$, that is, the value of Y_0 for the untreated individuals. Thus, we can calculate:

$$\Delta = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 0)$$

What is the difference between Δ and the ATT? Adding and subtracting the term $E(Y_0 \mid D = 1)$:

$$\Delta = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1) + E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)$$
$$\Delta = ATT + E(Y_0 \mid D = 1) - E(Y_0 \mid D = 0)$$
$$\Delta = ATT + SB$$

The second term, *SB*, is the selection bias: the difference between the counterfactual for treated individuals and the observed outcome for the untreated individuals. If this term is equal to 0, then the ATT can be estimated by the difference between the mean observed outcomes for treated and untreated:

$$\widehat{ATE} = E(Y \mid D = 1) - E(Y \mid D = 0)$$

However, in many cases the selection bias term is not equal to 0 (see the teacher-training example below). In these cases, the difference in means will be a biased estimator of the ATT. The main goal of an evaluation is to ensure that the selection bias is equal to 0 in order to correctly estimate the parameter of interest.

If participants are randomly assigned to an intervention, the average difference in outcomes between the treated and control units is due to the impact of the treatment (or possibly sampling error). As long as the sample is sufficiently large, differences in mean group outcomes should reflect the average impact of the intervention. Box 3 provides a brief description of the experimental method.

Box 3: Experimental Design

In an experimental design, the assignment to treatment is determined by a purely random mechanism. For example, one could assign a number to each eligible individual and select the treated individuals by lottery.

The main advantage of random assignment is that it guarantees that the treatment status (D) is uncorrelated with any other variables, both observable and unobservable, and, as a result, the potential outcomes will be statistically independent of the treatment status. In technical notation:

$$(Y_1, Y_0) \perp D$$

This means that with random assignment, all the characteristics of the individuals are equally distributed between treated and untreated groups (i.e., the proportions are the same). On average, the groups will be identical, except for the fact that one of them received the treatment. This implies that:

$$E(Y_0 \mid D = 1) = E(Y_0 \mid D = 0)$$

which allows one to replace the left-hand side (unobservable) with the right-hand side, which is observable, to estimate the ATT. Thus, experimental design ensures that the selection bias term is 0, and therefore, the impact of the program can be estimated as a simple difference between the average outcomes between groups. The impact may also be estimated, with identical results, by running a linear regression of the outcome on the treatment status variable and a constant:

$$Y = \alpha + \beta D + \varepsilon$$

where β captures the impact of the program.

However, for many interventions, random assignment is not a feasible approach. In fact, it is frequently the case that assignment to treatment is intentionally nonrandom, such as the assignment of Honduran youth to particular training activities based on their interests and geographic location (see Box 1). In absence of an experimental design, the untreated group is unlikely to be a good comparison for the treated group because of *selection bias*. To further illustrate this problem, consider a teacher-training program where participation is voluntary. In this case, it is probable that the more motivated teachers will sign up for the program. However, it is also expected that more motivated teachers would have done better even in absence of the treatment, which means that their average value of Y_0 is different from the corresponding value of nonparticipating teachers. As a consequence, when comparing the differences in mean outcomes (e.g. student performance) of treated and untreated teachers, it is difficult to isolate the impact of the training from the effect of greater motivation of the treated teachers. In this context, the simple estimator based on the difference in means between treated and untreated groups is tainted by *self-selection bias*.

Selection bias may also arise from actions on the part of those implementing the intervention. For example, although one may account for explicit targeting criteria in designing an impact evaluation for an intervention, such as a health program oriented to low income families, if program administrators also selectively enroll families based on other, undocumented criteria such as the families' apparent willingness to cooperate with treatment, *administrative selection bias* (or *program placement bias*) will result. In this example, participating families would be more cooperative than nonparticipating families, which might correlate with other unobserved characteristics of the families.

It is also possible that, due to problems in the implementation process, an experimental evaluation design fails to produce a valid control group. For example, program administrators may circumvent procedures intended to ensure that assignment of units eligible for the intervention is random. This is not uncommon in situations when implementing staff are required to invite prospective subjects to participate in the intervention and then have to deny treatment after their expression of interest. In such cases, nonexperimental techniques may be required to adjust for the bias that arises.

In sum, in the context of nonexperimental designs (or flawed experimental designs), it is necessary to account and adjust for differences between treated and untreated groups in order to properly estimate the impact of the program.

We now introduce some notation to address these basic issues in more technical detail. We use Y_1 and Y_0 to denote the *potential outcomes* for a unit in presence and absence of the treatment, respectively. The *observed outcome* Y for an individual will be Y_1 if the individual is treated and Y_0 otherwise. We will use the binary variable D to indicate the treatment status of the observed units, namely, D=1 for those who participate and D=0 for those who do not participate. Then we can write the observed outcome as:

$$Y = (1 - D)Y_0 + DY_1$$

This expression should be interpreted as follows. When a given unit is treated, then D=1, and thus $(1-D)$=0. The observed outcome for this unit will be:

$$Y = 0 \cdot Y_0 + 1 \cdot Y_1 = Y_1$$

which means that the observed outcome (Y) for treated units is equal to the potential outcome in case of treatment (Y_1). In this case, the potential outcome in absence of treatment, Y_0, is not observed: since the unit was treated, it is impossible to know what would have happened to this unit in absence of treatment. For a treated unit, Y_0 is the *counterfactual*. Similarly, when the unit is not treated, $D=0$ and $(1-D)$=1, and thus $Y=Y_0$. In this case, the counterfactual is Y_1.

Evaluations employing random assignment methods assure that the treatment is independent of Y_0 and Y_1 and the factors influencing them. The average treatment effect for those subject to random assignment may be estimated as the simple difference in mean outcomes for those assigned to treatment and those assigned to the control group. Without random assignment, where treatment (D) may be correlated with factors influencing Y_0 and Y_1, participants may differ from nonparticipants in many ways besides the effect of the program, so the simple difference in outcomes between participants and nonparticipants will not necessarily identify the impact of the intervention.

Matching methods are designed to ensure that impact estimates are based on outcome differences between comparable individuals. The basic idea behind matching can be traced to early statistical work, although the development of matching methods in the last 25 years (Rosenbaum and Rubin, 1983) and recent increases in computing power have facilitated their

wider implementation. The simplest form of matching pairs each participant to a comparison group member with the same values on observed characteristics (collected in a vector *X)*. If the number of variables in *X* is large, such an approach may not be feasible. The alternative applied in propensity-score matching is to compare cases that are "close" in terms of *X*, where participating units are matched with untreated units based on an estimate of the probability that the unit receives the treatment (the propensity score), as will be discussed in section five.

The matching estimator will not necessarily work in all circumstances; specific conditions have to be met to produce valid impact estimates. First, if the condition requiring one to find untreated units that are similar in all relevant characteristics to treated units is to be satisfied, it is clear that these characteristics must be observable to the researcher. In other words, PSM requires *selection on observables*; the inability of the researcher to measure one or more relevant characteristics that determine the selection process results in biased estimations of the impact of the program. Second, in order to assign a comparison unit to each treated unit, the probability of finding an untreated unit for each value of *X* must be positive. These issues will be addressed in the next section.

3. When to Use Matching: Assumptions and Data Requirements

In order to determine if matching is likely to effectively reduce selection bias, it is crucial to understand under what conditions it is most likely to work. This section discusses these issues, emphasizing the theoretical assumptions underlying the matching estimator and the data requirements for implementing it.

3.1 Assumptions

In an experimental design, randomization ensures that all the relevant characteristics, either observable or unobservable, of the studied units are *balanced* (this means, they are equally distributed) between treatment and control group and, because of this, the difference in mean outcomes correctly estimates the impact of the intervention. In the absence of randomization, however, the groups may differ not only in their treatment status, but also in their values of X. In this case, it is necessary to account for these differences (in econometric jargon, *to control for X* or *to condition on X*) to avoid potential biases.

To illustrate what this means, it is useful to think that controlling on X is done by stratifying the sample over X. For simplicity, consider the case where there is a single discrete variable X, such as level of educational attainment, that takes only a few values, for example, 1=no school, 2=elementary, 3=secondary and 4=postsecondary. Then, the observations are grouped according to their values of X: all the units (both treated and untreated) with X=1 are grouped together, as are units with X=2, X=3 and X=4, making a total of four groups of treated and untreated units matched by education levels. The mean outcomes for treated and untreated units are compared for each one of these strata, and subsequently, the ATT can be computed by averaging these differences over the four groups.

It is now clearer that two conditions must be satisfied to implement this estimator. First, the variables (X) on which the treated and untreated groups differ must be observable to the researcher. Although this may seem obvious for the simple example above, this assumption, known as the *conditional independence* or *unconfoundedness* assumption, becomes more subtle when a large number of variables may be potentially affecting the selection into the program.

Second, in order to calculate the difference in mean outcomes for each value of X, for each possible value of the vector of covariates X, there must be a positive probability of finding

both a treated and an untreated unit to ensure that each treated unit can be matched with an untreated unit. If some units in the treatment group have combinations of characteristics that cannot be matched by those of units in the comparison group, it is not possible to construct a counterfactual, and therefore, the impact for this subgroup cannot be accurately estimated. This is commonly known as the *common support* or *overlap condition.*

Box 4 presents a summary of these assumptions in more technical terminology.

Box 4: PSM Assumptions

Assumption 1 *(Conditional Independence Assumption or CIA): there is a set X of covariates, observable to the researcher, such that after controlling for these covariates, the potential outcomes are independent of the treatment status:*

$$(Y_1, Y_0) \perp D \mid X$$

This is simply the mathematical notation for the idea expressed in the previous paragraphs, stating: the potential outcomes are independent of the treatment status, *given X*. Or, in other words: **after controlling for *X*, the treatment assignment is "as good as random"**.

This property is also known as *unconfoundedness* or *selection on observables*. The CIA is crucial for correctly identifying the impact of the program, since it ensures that, although treated and untreated groups differ, these differences may be accounted for in order to reduce the selection bias. This allows the untreated units to be used to construct a counterfactual for the treatment group.

Assumption 2 *(Common Support Condition): for each value of X, there is a positive probability of being both treated and untreated:*

$$0 < P(D = 1 \mid X) < 1$$

This last equation implies that the probability of receiving treatment for each value of X lies between 0 and 1. By the rules of probability, this means that the probability of not receiving treatment lies between the same values*. Then, a simple way of interpreting this formula is the following: the proportion of treated and untreated individuals must be greater

than zero for every possible value of *X*. The second requirement is also known as *overlap condition*, because it ensures that there is sufficient overlap in the characteristics of the treated and untreated units to find adequate matches (or a *common support*).

When these two assumptions are satisfied, the treatment assignment is said to be *strongly ignorable* (Rosenbaum & Rubin, 1983)**

* This is because $P(D=0|X)=1-P(D=1|X)$

** In fact these conditions may be relaxed when the parameter of interest is the ATT. See Appendix 1.

3.2 Data Requirements

The data (variables) available for matching are critical to justifying the assumption that, once all relevant observed characteristics are controlled, comparison units have, on average, the same outcomes that treated units would have had in the absence of the intervention. Since in many cases the researcher does not know precisely the criteria that determine participation, it is common to control for all the variables that are suspected to influence selection into treatment (although controlling for too many variables could generate problems with the common support: see section 6.1). As a result, the researcher should have access to a large number of variables to be able to correctly characterize the propensity score.

Prior evaluation research has also shown that it is important for data for both the treatment and comparison units to be drawn from the same sources (i.e., the same data-collection instruments), so that the measures used (for control and outcome variables) are identical or similarly constructed. In cases where the data on treated units and comparison units derive from different sources, it is critical to attempt to ensure that the variables are constructed in the same way (e.g., under the same coding conventions). Any missing data should also be handled similarly for treated and untreated units. Although data errors are always a potential issue, the bias in impact estimates may be relatively small if data errors have the same structure for treated and comparison units. In contrast, if there are systematic differences in the way that errors are introduced, particularly for outcome measures, even small differences may induce substantial biases in impact estimates.

Finally, to obtain impact estimates that are generalizable to the population of interest, it is necessary for the pool of comparison units to have a sufficient number of observations with characteristics corresponding to those of the treated units. If the comparison pool is large enough, adequate matches may be possible even if the average unmatched characteristics are very different. If the variables in question are of substantial importance, however, it may be necessary to discard treated units whose characteristics cannot be matched in estimating impacts.

Because of its large data requirements (regarding both the number of variables and the sample size), the PSM methodology is often described as a "data-hungry method". When data are scarce, the appropriateness of this technique should be carefully analyzed.

4. Basic Mechanics of Matching

To understand the basic mechanics of matching methods, consider the following very simple example in which we want to calculate the ATT of a treatment D on the income level of the treated individuals. In this case X represents a single discrete variable, namely, education level. The database is presented in table 1.

Table 1 PSM – A Very Simple Example

i	*D*	Education	Income
1	0	2	60
2	0	3	80
3	0	5	90
4	0	12	200
5	1	5	100
6	1	3	80
7	1	4	90
8	1	2	70

When D=1, the observed outcome is equal to the potential outcome under treatment, Y_1, and when D=0, the observed outcome is the potential outcome in absence of treatment, Y_0. For each treated unit, the comparison unit is the untreated unit with most similar characteristics (or value of X). This special case of matching is called *nearest neighbor covariate matching*; the counterfactual outcome of the treated units is estimated by the observed outcome of the most similar untreated unit[1]. The results of the matching are shown in table 2.

[1] If two untreated units share the same value of X, we average the outcome between the two.

Table 2 A Very Simple Example (continued)

i	*D*	Education	Income	Match	Y_1	Y_0	Difference
1	0	2	60	-	-	-	-
2	0	3	80	-	-	-	-
3	0	5	90	-	-	-	-
4	0	12	200	-	-	-	-
5	1	5	100	[3]	100	90	10
6	1	3	80	[2]	80	80	0
7	1	4	90	[2.3]	90	85	5
8	1	2	70	[1]	70	60	10

Finally, the ATT is estimated by taking the differences between Y_1 (observed) and Y_0 (estimated), and averaging over all the treated units. In this case, the estimated ATT is (10+0+5+10) / 4 = 6.25.

Although the calculation of the matching estimator is very simple for this example, this case is far from the typical reality. The major practical problem arises when there are numerous differences between treated and untreated units to control for; this is the rule rather than the exception.

Box 5: The Curse of Dimensionality

An intuitive way of understanding the *problem of dimensionality,* to which matching estimators are subject, is the following. We saw that the idea of matching techniques is to match treated individuals with untreated units that are similar or close in terms of *X*. When *X* is a single variable, as in the example above, the meaning of the word "similar" is clear: if we take a treated and an untreated unit, the closer their values of *X*, the more similar the units are. Say *X* represents the income of an individual. Comparing the treated individual A with income *X*=1000 to two untreated individuals, B and C, with incomes 1100 and 2000 respectively, it is very easy to see that unit B is closer to A than C is.

However, imagine that we need to match observations on both income (*I*) and years of education (*E*). In this case the vector *X* contains the two variables: *X*=(*I,E*). Suppose that we want to pair individual A, who has *X*=(1000,5) with individuals B or C, who have *X*=(1100,12) and *X*=(2000,8), respectively. Which of the two untreated individuals, B or C, are closer to A? There is not an obvious answer to this question: individual B is closer to A in terms of income, but farther in terms of years of education. More generally, when working on multiple dimensions (that is, with many variables), the idea of "closeness" is not clearly defined.

The solution proposed by the statisticians Paul Rosenbaum and Donald Rubin to the dimensionality problem (discussed in Box 5) is to calculate the *propensity score*, which is the probability of receiving the treatment given *X*, noted $P(D=1 \mid X)$ or simply $p(X)$. Rosenbaum and Rubin (1983) proved a key result that forms the theoretical basis of PSM: when it is valid to match units based on the covariates *X*, it is equally valid to match on the propensity score. In other words, the probability of participation summarizes all the relevant information contained in the *X* variables. The major advantage realized from this is the reduction of *dimensionality*, as it allows for matching on a single variable (the propensity score) instead of on the entire set of covariates.

In effect, the propensity score is a balancing score for *X,* assuring that for a given value of the propensity score, the distribution of *X* will be the same for treated and comparison units. Appendix 1 further illustrates this key result.

Using this result, the procedure for estimating the impact of a program can be divided into three straightforward steps:

1. Estimate the propensity score
2. Choose a matching algorithm that will use the estimated propensity scores to match untreated units to treated units
3. Estimate the impact of the intervention with the matched sample and calculate standard errors.

These steps are addressed in detail in the next section.

5. How to Implement Propensity-Score matching (PSM)

5.1. Characterizing the Propensity Scores

The first step in PSM analysis is to estimate the propensity score. Normally, a logit or probit function is used for this purpose, given that treatment is typically dichotomous (i.e., D=1 for the treated and D=0 for untreated units). It is critical that a flexible functional form[2] be used and that all relevant covariates that relate to treatment status and outcomes are included in this model (to account for differences between treated and untreated units, as described in the preceding sections).

For a binary treatment variable, there is no strong advantage to using the logit vs. probit model, although both are typically preferred to a linear probability model, which is known to produce predictions outside the [0, 1] bounds of probabilities.

One of the key issues in characterizing the propensity score is the specification of the selection model, i.e., the identification of the variables that determine the participation. In most cases, there will be no comprehensive list of clearly relevant variables that will assure that the matched comparison group will provide an unbiased impact estimate. For each evaluation, it is important to consider what factors make the comparison units distinct from treated units. To the extent that these factors are associated with outcomes, controls for them are essential.

One obvious set of factors to include in PSM estimation are explicit criteria used in determining participation in the intervention, such as a project or program's eligibility or admission criteria. It is important to consider factors associated with both self-selection, such as a youth's distance from the location for applying for a training opportunity, as well as administrative selection, which may involve discretionary as well as overt criteria. Other institutional factors or implementation variables that might also influence take-up of treatment should, to the extent measurable, also be included. In cases such as the *Mi Primer Empleo* program (see Box 1), where there are more units qualified for treatment than there are treatment slots available, using the eligible units excluded from treatment as the comparison group may eliminate the need to control for some of the self- or administrative-selection factors that

[2] Flexible functional forms allow capturing possible nonlinearities of the participation model. To do this, in addition to the covariates, higher-order terms (like quadratic, cubic, etc) and / or interaction terms can be added to the model.

determine who or which units apply for treatment. Section 6.1 provides a guideline to assess the validity of the selection model.

An example of a participation model is presented in Box 6.

Box 6: Characterizing the Participation Model

To evaluate the impact of Ecuador's *Agricultural Services Modernization Program* (PROMSA), Maffioli, Valdivia and Vázquez (2009) implement a PSM estimator using farm-level data. The propensity score is estimated using various socio-demographic characteristics of farmers, namely: age, years of education, cohort dummies, gender and ethnic origin (a dummy equal to 1 if the farmer is indigenous), together with some higher order terms (like age and education squared) and interaction terms (age interacted with gender, etc). Table B.1 in the next page shows the results for the participation model.

Looking at the standard errors (shown in parentheses) we see that participation is clearly (nonlinearly) related to age, and that individuals without high school education have a higher probability of receiving treatment, although the number of years of education appears not to be significant. Furthermore, cohort and ethnic origin also affect the propensity score, while gender and the interaction terms do not seem to explain participation.

Table B.1 PROMSA participation model

Variable	Coefficient
Age	0.688
	(0.261)***
Age^2	-0.014
	(0.005)**
Age^3	0.00009
	(0.00003)**
Years of education	0.088
	(0.26)
Education^2	0.025
	(0.04)
Education^3	-0.001
	(0.001)
Primary education	-0.35
	(0.51)
High school or more	-1.457
	(0.452)***
Born between 1930-1939	2.706
	(1.105)**
Born between 1940-1949	3.609
	(1.442)**
Born between 1950-1959	3.911
	(1.552)**
Born between 1960-1969	3.793
	(1.545)**
Born between 1970-1979	4.739
	(1.555)***
Born between 1980-1989	5.045
	(1.767)***
Gender	0.098
	(0.75)
Indigenous	-0.29
	(0.131)**
Education*age	-0.001
	(0.002)
Education*gender	0.051
	(0.04)
Gender*age	-0.002
	(0.01)
Constant	-16.873
	(5.100)***
Observations	1179

Standard errors in parentheses

* Significant at 10%; ** significant at 5%; *** significant at 1%

Source: Maffioli et al (2009)

5.2. Choosing a Matching Algorithm

In choosing between different *matching algorithms*, that is, among alternative ways of using the propensity score to match comparison units with treated units, the following primary factors should be considered:

1. Matching with or without replacement
2. How to assess (or set the standard) for proximity, i.e., the closeness of the match
3. Whether and how to weight cases in the analysis
4. Number of comparison units matched to each treatment unit

Early matching estimators paired units in the treated group with those in the comparison group on a one-to-one basis. For each treated case, a case in the comparison group that was most similar to that case—in terms of propensity score, in the case of propensity-score matching—would be matched with it. In effect, the program impact for a particular treated case i was estimated as $Y_{1i} - Y_{0j(i)}$, where $Y_{0j(i)}$ is the outcome for the comparison case that is matched with the treated case i. The estimated program impact was obtained as the average of this measure over all treated cases. Such pairwise matching was usually performed using sampling from the comparison group without replacement, meaning that each comparison group member could be included as a matched case only once. However, matching without replacement may perform poorly when there is little overlap of the propensity scores or when the control group is small, since treated units are matched to observations that are not necessarily similar (see Dehejia and Wahba, 2002, for further details). This is why it is now more common for studies to use sampling with replacement, allowing for one comparison case to serve as the match for more than one treated case.

In addition, alternative approaches have recently been recognized as superior to pairwise matching. In contrast to matching one comparison group case with a given treated case, it has been found that estimates are more stable (and make better use of available data) if they consider all comparison cases that are sufficiently close to a given treated case. As indicated above, it is also important to include in the comparison only those cases that are sufficiently "close" to a given treated case. Although allowing a given case to be used in many comparisons may inflate sampling error, it is now generally accepted that the benefits of close matches outweigh these other costs.

The vast majority of studies using PSM measure the proximity of cases as the absolute difference in the propensity score. As Smith and Todd (2005) note, such an approach is not robust to "choice-based sampling," where the treated are oversampled relative to their frequency in the population of eligible individuals (Caliendo & Kopeinig, 2005). Matching on the log odds of the propensity score, defined as p/(1-p), assures that results are invariant to choice-based sampling (see Box 7).

Box 7: Matching with Choice-Based Sampling

With choice-based sampling, the number of treated and comparison cases does not reflect the likelihood that an individual with given characteristics participates in the program in the full universe, but rather is determined by various factors outside the control—and knowledge—of the researcher. Matching on the log odds of the propensity score has the advantage that it "spreads out" the density of scores at very low or very high propensity scores. Use of the log odds also allows for a consistent bandwidth to be used. In addition, since the logit is used to predict propensity score, the log odds are a linear combination of the independent variables, and a constant radius is expected to translate into the same metric at different propensity score levels.

Although the theory underlying propensity-score matching implies that as the sample size grows, matching on the propensity score also matches on all control variables, in any given application with a finite sample, there is no assurance that matches will be close enough to remove significant differences. In addition, since applications by necessity use a parametric structure to calculate the propensity score, inadequacies in the estimation method may cause further deviations. It is therefore necessary to compare the treated cases with the matched comparison cases. In general, if differences are too great, it may be necessary to alter the caliper used in the analysis or to modify the details of how the propensity score is estimated.

Below are descriptions of the most commonly employed matching algorithms.

Nearest neighbor matching is one of the most straightforward matching procedures. An individual from the comparison group is chosen as a match for a treated individual in terms of the closest propensity score (or the case most similar in terms of observed characteristics).

Variants of nearest neighbor matching include "with replacement" and "without replacement," where, in the former case, an untreated individual can be used more than once as a match and, in the latter case, is considered only once.

To avoid the risk of poor matches, **radius matching** specifies a "caliper" or maximum propensity score distance by which a match can be made. The basic idea of radius matching is that it uses not only the nearest neighbor within each caliper, but all of the comparison group members within the caliper. In other words, it uses as many comparison cases as are available within the caliper, but not those that are poor matches (based on the specified distance).

In many-to-one (radius) caliper matching with replacement, the estimator of program impact may be written as:

$$E(\Delta Y) = \frac{1}{N}\sum_{i=1}^{N}\left[Y_{1i} - \overline{Y}_{0\,j(i)}\right]$$

where $\overline{Y}_{0\,j(i)}$ is the average outcome for all comparison individuals who are matched with case i, Y_{1i} is the outcome for case i, and N is the number of treated cases. This approach does not limit the number of cases that are matched with a given participant, as long as those cases are "close" enough.

Kernel and local-linear matching are nonparametric matching estimators that compare the outcome of each treated person to a weighted average of the outcomes of all the untreated persons, with the highest weight being placed on those with scores closest to the treated individual. One major advantage of these approaches is the lower variance, which is achieved because more information is used. A drawback of these methods is that some of the observations used may be poor matches. Hence, the proper imposition of the common-support condition is of major importance for these approaches. When applying kernel matching, one also has to choose the kernel function and the bandwidth parameter.

Unfortunately, there is no clear rule for determining which algorithm is more appropriate in each context. However, a key issue that should be considered is that the selection of the matching algorithm implies a bias / efficiency trade-off. For instance, by using only one nearest neighbor we guarantee that we are using the most similar observation to construct the counterfactual. This minimizes the bias, since the characteristics between both units will be, in general, very similar. However, using this technique ignores a lot of information from the

sample, since many untreated units are not used for the estimation. Therefore, the reduction in the bias comes with an increase in the imprecision of the estimates caused by a higher variance, i.e., a decrease in efficiency. On the other hand, when using many neighbors, the estimator is more efficient since it exploits a larger quantity of information from the untreated pool, but at the price of increasing the bias by using poorer matches.

5.3. Estimating Intervention Impacts and Interpreting the Results

After propensity scores have been estimated and a matching algorithm has been chosen, the impact of the program is calculated by just averaging the differences in outcomes between each treated unit and its neighbor (or neighbors). Any of a number of statistical software programs can be used to perform the matching and to generate estimates of the impact of the intervention. The statistical-analysis software most frequently used for PSM is Stata. A program developed by Leuven and Sianesi (2003), **psmatch2**, can be installed in Stata and used to implement PSM. The options for PSM techniques include nearest neighbor, caliper matching (with and without replacement), radius matching, kernel matching, local-linear matching and Mahalanobis metric (covariate) matching. The **psmatch2** program also includes routines for common-support graphing (**psgraph**) and covariate imbalance testing (**pstest**).

Becker and Ichino (2002) also provide a program for PSM estimation in Stata that includes estimation routines for nearest neighbor, kernel, radius, and stratification matching and balancing tests can also be performed with this program.

Of course, it is not possible to interpret the results of the impact estimation without estimating the standard errors, which provide an indicator of the importance of sampling error in the estimates generated. Conventionally, standard errors of propensity score-matching estimates are obtained using bootstrap methods. In general, the bootstrap relies on sampling from the analysis sample with replacement, replicating the analysis multiple times. The estimated standard error is the standard deviation of the estimated-impact estimate across replications.

Bootstrapping methods for calculating standard errors are easily implemented in **psmatch2** or the Becker and Ichino PSM estimation program.

There are several important limitations, however, to bootstrap standard errors. As is the case for estimation procedures for analytical standard errors, theorems supporting the use of bootstrapping show that bootstrapping produces error estimates that are *asymptotically* unbiased.

For finite (small) samples, there is no certainty that estimates will be unbiased. An additional problem with bootstrapping is the intensive computer resources required to estimate them; with large samples, it is not always feasible to calculate bootstrap standard errors for all estimates (see e.g., Lechner, 2001).

Additional information on statistical software for PSM is shown in Appendix 2.

When interpreting the results, it is important to evaluate the robustness of the estimations by changing the matching algorithms or by altering the parameters of a given algorithm. Robustness checks help increase the reliability of the results by showing that the estimations do not depend crucially on the particular methodology chosen. An example is shown in Box 8.

Box 8: Checking the Robustness of the Results

Using the propensity score estimated as shown in Box 6, Maffioli et al. (2009) match treated farmers with similar untreated farmers to evaluate the impact of PROMSA on different outcomes like productivity, use of technology, associability, type of purchaser and access to formal credit, among others. The authors find evidence of impact on type of purchaser, associability and access to formal credit, but no effect is found on productivity and technology. To make sure that these findings are not driven by the selection of a particular strategy, coefficients are estimated using different matching algorithms. The results of this robustness check are shown in table B.2.

Table B.2 Robustness Checks

Outcome	Nearest Neighbor (1)	Nearest Neighbor (5)	Caliper (0.001)	Radius (0.001)	Normal Kernel
Productivity Index	-29.9515	44.0525	33.587	59.0923	53.4409
	(88.8228)	(38.5607)	(105.4436)	(43.7452)	(25.7168)**
	[60.9315]	[38.9558]	[73.7986]	[57.4125]	[38.9506]
Technology Index	0.1151	0.0362	0.1293	0.1093	-0.036
	(0.1266)	(0.0926)	(0.1371)	(0.1001)	(0.0790)
	[.177]	[.1711]	[.2105]	[.2095]	[.1568]
Associability	0.6393	0.6707	0.6571	0.6797	0.674
	(.0443)***	(.0315)***	(.0477)***	(.0325)***	(.0289)***
	[.0709]***	[.0812]***	[.0913]***	[.0758]***	[.0785]***
Type of purchaser	0.1378	0.1081	0.1494	0.1223	0.1136
	(.0463)***	(.0323)***	(.0489)***	(.0341)***	(.0273)***
	[.0623]**	[.0666]	[.0635]**	[.0601]**	[.0485]**
Access to formal credit	0.3441	0.3369	0.3782	0.3471	0.3461
	(.0385)***	(.0325)***	(.0409)***	(.0346)***	(.0303)***
	[.0808]***	[.1071]***	[.1091]***	[.0986]***	[.1111]***

Note: each column reports the matching estimator with a different matching algorithm (1) nearest neighbor matching using 1 nearest neighbor (2) nearest neighbor matching using 5 nearest neighbors (3) caliper mathing with a caliper of 0.001 (4) radius matching with a caliper of 0.001 (5) Kernel matching using Normal density function

Standard errors in parentheses, bootstraped clustered standard errors at *parroquia* level in brackets

* Significant at 10%; ** significant at 5%; *** significant at 1%

Source: Maffioli et al (2009)

The matching algorithms used are nearest neighbor (with one and five neighbors), caliper and radius (with a caliper of 0.001) and kernel (using a Normal density). Furthermore, standard errors are also estimated using bootstrap (in brackets).

The case for productivity index with Normal kernel (line 1 in the above table) is a good example of a nonrobust result. Although there seems to be a positive effect of the treatment under one specification, the significance fades after slight changes in the estimation method. The impact found on associability, type of purchaser and access to formal credit, on the other hand, does *not* appear to depend critically on the algorithm used, since both the value of the coefficients and its significance are very similar using different alternatives.

Box 9: Basic Steps in Implementing PSM

1. Characterize the propensity score:

- Define the selection model, using variables that:

$\Rightarrow$ Affect both the probability of participation and the outcome

$\Rightarrow$ Are not affected by the treatment

- Estimate the propensity score $P(D=1|X)$ using a binary choice model (probit or logit) and calculate the predicted probabilities

2. Choose an appropriate algorithm and perform the matching using the propensity score, considering:

- The key parameters of each algorithm (number of neighbors, caliper, bandwidth, etc.)
- The bias / efficiency trade-off

3. Estimate the results, evaluating:

- Standard error estimation and statistical significance
- Robustness checks.

6. Testing Assumptions and Specification Tests

As with any experimental or nonexperimental method used in impact evaluation, it is important to check the key assumptions that are made in the estimation and verify that the model specification is appropriate and that the results do not suffer from bias.

6.1. CIA: Guidelines and Tests for Model Specification

As explained earlier, the conditional independence assumption that we make in applying this method asserts that selection into the intervention is based only on observable characteristics of the eligible units, and that after conditioning on these variables influencing participation, the expected outcome in the absence of treatment does not depend on treatment status. Unfortunately, this assumption is not directly testable but still requires justification.

If one cannot argue that selection into treatment is completely random, knowledge of the selection process is essential, as omitting important variables in estimation may contribute to bias in the results. In effect, this requires justification of the specification of the (first-stage) propensity score model. There are some basic guidelines for model specification that can be checked or verified:

1. If explicit criteria are used in determining participation in the intervention, these variables should be included in estimating participation in treatment. The more transparent, precise and well-controlled the selection process, the more confidence one can have that all relevant variables are included. Frequently, knowledge of the institutional settings in which selection takes place is critical to model specification.
2. Measures included in the first-stage model should either be stable (constant) over time, deterministic with respect to time (e.g., age) or measured before participation so that they are not confounded with outcomes or the anticipation of treatment.
3. It is sometimes the case that one or more variables are particularly influential in determining participation, and one may want to "hard match" or "exact match" on such characteristics. Fundamentally, this requires performing the complete matching procedure separately for the subgroups defined by a given characteristic (e.g., separate estimation of the propensity score for men and women if the measure is gender).

4. Data for the treatment and comparison units should be drawn from the same sources (e.g., the same data-collection instruments) and the measures should be similarly constructed.
5. A variable that captures some randomness in the selection process (i.e., a randomization device such as a quota) is particularly useful as it assures that units with similar (or the same) characteristics can be observed in both the treated and untreated states.
6. Including irrelevant variables (that do not influence participation in the intervention) should be avoided so that they do not worsen the common support problem or unnecessarily increase the variance of the estimates.

There are also somewhat more formal specification tests that one can conduct to assess the validity of the propensity score model specification.

One very basic approach is to examine the statistical significance of the covariates and keep only those variables that are statistically significant and increase the predictive power of the model. One way to do this is to begin with a parsimonious model and add one covariate at a time, although even the most parsimonious model should include all variables explicitly known to influence selection (e.g., program-eligibility criteria).

A related approach, described by Caliendo and Kopeinig (2005) as the "hit-or-miss" method, involves choosing variables so as to maximize the within-sample correct prediction rate. For each observation, the estimated propensity score is compared to the sample proportion of units taking up the treatment, and the observations are classified based on whether the propensity score is larger or smaller than this proportion. For both of these above approaches, it is important to keep in mind that the primary goal is to balance the covariates between the treatment and comparison groups, more so than maximizing predictive power.

Another strategy for propensity score model specification involves beginning with a minimum model specification and adding blocks of potentially relevant variables, and then checking whether the goodness of fit of the model improves with each addition. Black and Smith (2004) use the root mean squared error criterion to assess goodness of fit with additions to the model, although one can also examine the standard errors of the variables and other criteria. A potential problem with this approach is that a smaller number of conditioning variables is less likely to contribute to problems in satisfying the common support condition, which in turn results in a narrower bandwidth that reduces bias. Thus, as Black and Smith note, one might be led to a more parsimonious specification that performs better by these criteria but leaves out variables

that are important (based on theory and other empirical evidence) in controlling for selection. Clearly, the results of this type of specification test need to be taken into account with the other guidelines above for specifying the propensity score model.

It is important to mention that all of these approaches need to be informed by sound theory to decide which variables are relevant to characterize the participation model.

6.2. Balancing Tests

The next step in assessing the quality of matching is to perform tests that check whether the propensity score adequately balances characteristics between the treatment and comparison group units. Formally, the objective of these tests is to verify that treatment is independent of unit characteristics after conditioning on observed characteristics (as estimated in the propensity score model):

$$D \perp X \mid p(X)$$

where X is the set of characteristics that are believed to satisfy the conditional independence assumption. In other words, after conditioning on $p(X)$, there should be no other variable that could be added to the conditioning set of the propensity score models that would improve the estimation, and after the application of matching, there should be no statistically significant differences between covariate means of the treatment and comparison units. It is important to note that only "after-matching" tests compare differences in time-invariant covariates (that are unaffected by treatment) for the resulting matched sample.

In examining the results of after-matching balancing tests, one looks to see that any differences in the covariate means between the two groups in the matched sample have been eliminated, which should increase the likelihood of unbiased treatment effects. If differences remain, refinements to the propensity score model specification should be made to improve the resulting balance, or a different matching approach should be considered. See Box 10 for an application of balancing tests.

Box 10: Checking the Balancing between Groups

Following the case of PROMSA from boxes 6 and 8, Maffioli et al. (2009) run t-tests of equality of means before and after the matching to evaluate if the PSM succeeds in balancing the characteristics between treated and untreated groups. Table B.3 shows the difference in means for different variables before the matching, that is, using the full sample.

Table B.3 Differences in Mean before Matching

Variable	Control	Treated	Difference
Age	46.09	48.19	2.099
	(11.24)	(10.61)	(.7214)***
Years of education	4.501	5.765	1.264
	(3.41)	(2.93)	(.2061)***
Primary education	0.8374	0.8	-0.0374
	(0.40)	(0.37)	(.0254)***
High school or more	0.108	0.1724	0.0643
	(0.38)	(0.31)	(.0221)***
Born between 1930-1939	0.0226	0.0446	0.022
	(0.21)	(0.15)	(.0110)**
Born between 1940-1949	0.0711	0.1099	0.0387
	(0.31)	(0.26)	(.0182)**
Born between 1950-1959	0.467	0.4776	0.0105
	(0.50)	(0.50)	(.0335)**
Born between 1960-1969	0.3171	0.2336	-0.0834
	(0.42)	(0.47)	(.0306)***
Born between 1970-1979	0.0658	0.103	0.0372
	(0.30)	(0.25)	(.0176)**
Born between 1980-1989	0.042	0.0206	-0.0214
	(0.14)	(0.20)	(.0126)*
Gender	0.8495	0.9078	0.0583
	(0.29)	(0.36)	(.0229)**
Indigenous	0.243	0.119	-0.124
	(0.32)	(0.43)	(.0271)***

Standard errors in parentheses

* Significant at 10%; ** significant at 5%; *** significant at 1%

Source: Maffioli et al (2009)

There is clear evidence of covariate imbalance between groups. To avoid the biases that this may generate, the authors use the propensity score presented in Box 6 to define a matched control sample. The results from the tests of equality of means for the matched sample are shown in Table B.4.

Table B.4 Differences in Mean after Matching

Variable	Control	Treated	Difference
Age	48.11	48.44	0.3275
	(10.56)	(9.91)	(0.840)
Years of education	5.662	5.783	0.1219
	(3.42)	(2.92)	(0.272)
Primary education	0.8432	0.7979	-0.0452
	(0.40)	(0.37)	(0.032)
High school or more	0.1358	0.1742	0.0383
	(0.38)	(0.31)	(0.030)
Born between 1930-1939	0.0209	0.0418	0.0209
	(0.20)	(0.14)	(0.015)
Born between 1940-1949	0.1219	0.1114	-0.0104
	(0.32)	(0.25)	(0.027)
Born between 1950-1959	0.4947	0.4808	-0.0139
	(0.50)	(0.50)	(0.042)
Born between 1960-1969	0.2508	0.2299	-0.0209
	(0.42)	(0.47)	(0.036)
Born between 1970-1979	0.0975	0.1045	0.0069
	(0.31)	(0.25)	(0.025)
Born between 1980-1989	0.0139	0.0209	0.0069
	(0.14)	(0.20)	(0.011)
Gender	0.8954	0.9059	0.0104
	(0.29)	(0.36)	(0.025)
Indigenous	0.1289	0.1219	-0.0069
	(0.33)	(0.42)	(0.028)

Standard errors in parentheses

* Significant at 10%; ** significant at 5%; *** significant at 1%

Source: Maffioli et al (2009)

Clearly, after matching, the differences are no longer statistically significant, suggesting that matching helps reduce the bias associated with observable characteristics.

One can also calculate the standardized difference, that is, the size of the difference in means of a conditioning variable (between the treatment and comparison units), scaled by (or as a percentage of) the square root of the average of their sample variances.[3] Still another option is to use an *F*-test or Hotelling test in a joint test for the equality of means between treatment and comparison units for all of the covariates (rather than testing for balance in each of the covariates separately). If the null hypothesis of joint equality of means in the matched sample is rejected, this implies that the propensity score model is inadequate to ensure balance.

If these tests indicate that balance has not been achieved, and there is no other (available) variable that could be added to the model, another approach to improving the propensity score model performance in balancing the covariates is to modify the form of the variables in the model. For example, if there are large mean differences in an important covariate in the model between the treatment and comparison groups, one can add the square of the variable and/or interactions with other variables in reformulating the model. The estimation of the propensity score, matching procedure and balancing test would then be repeated to check for improvement in the balancing performance. This process could be repeated until balance is achieved. It is important to keep in mind, however, that in some cases, balance on the matched samples may not be possible, regardless of the amount of adjustment efforts made.

It is also important to recognize that achieving balance for the full sample does not imply balance for the resulting matched sample that is used to estimate the treatment effect (which is checked in after-matching tests).

If the units in a sample are classified into subsamples, propensity scores will be more similar within subclasses than in the full sample, and covariates will tend to be better balanced within the subclasses (as individuals being compared are more similar to each other). If stratification on the propensity score is performed, the check for balance within each stratum is done after the initial estimation of the propensity score but before examining outcomes; that is, it is a before-matching specification test. If the test results show important within-stratum differences, then the propensity score model specification needs to be revised, or there may be insufficient overlap in the covariate distributions to allow for subgroups.

[3] Rosenbaum and Rubin (1985) suggest that a standardized difference of 20 or more should be viewed as large.

6.3. Verifying the Common Support Condition

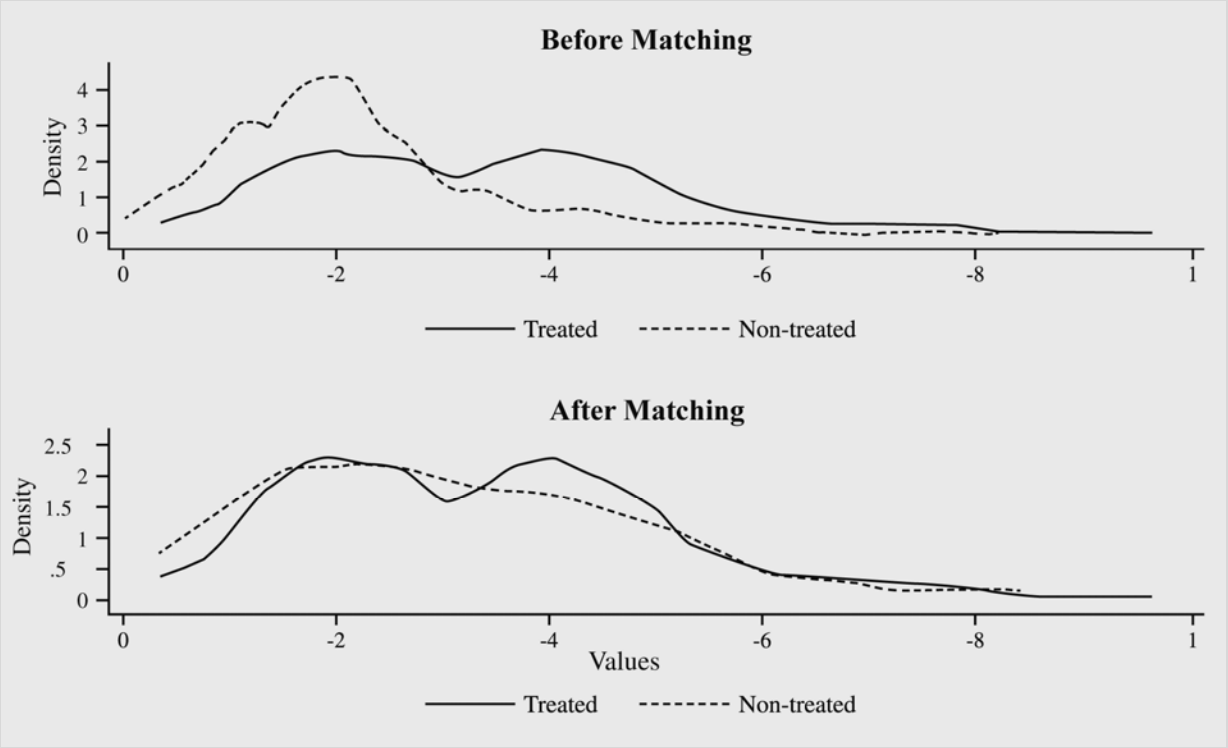
Another important step in investigating the validity or performance of the propensity score-matching estimation is to verify the common support or overlap condition. We assume that the probability of participation in an intervention, conditional on observed characteristics, lies between 0 and 1 (implying participation is not perfectly predicted, that is, $0 < P(D = 1 \mid X) < 1$). This assumption is critical to estimation, as it ensures that units with the same X values have a positive probability of being both participants and nonparticipants.

Checking the overlap or region of common support between treatment and comparison groups can be done with relatively straightforward strategies. One obvious approach is through visual inspection of the propensity score distributions for both the treatment and comparison groups. Simple histograms or density-distribution plots of propensity scores for the two groups, along with a comparison of the minimum and maximum propensity score values in each distribution, can typically give the researcher a good, initial reading of the extent to which there is overlap in the propensity scores of the treatment and comparison units. See Box 11 for an example.

Box 11: Visual Inspection of the Propensity Scores

In addition to the mean equality tests presented in box 10, it is useful to plot the distributions of the propensity scores for treated and untreated groups to visually check the overlap condition and to see if the matching is able to make the distributions more similar. The distributions of the propensity scores, before and after the matching, for the case of PROMSA are plotted in figure B.2.

Figure B.2 Propensity Score Distribution



Source: Maffioli et al (2009)

Visual inspection suggests that the densities of the propensity scores are more similar after matching. The plot also reveals a clear overlapping of the distributions.

To complement this informal analysis, more rigorous statistical tests like the Kolmogorov-Smirnov (KS) test (a way of testing the equality of two distributions) may be performed to confirm what visual inspection suggests. In this case, the KS test does not reject the null hypothesis of equality of distributions between groups after matching.

If there are clear and sizeable differences in the minima and maxima of the propensity score density distributions for the treatment and comparison groups, one strategy for addressing this problem is to delete all observations where propensity score is smaller than the minimum and larger than the maximum in the other group. However, it is not always necessary to discard observations that are very close to these bounds (e.g., that might fit within a specified caliper for matching).

One also needs to look for areas within the common support interval (defined by the minima and maxima) where there is only limited (or no) overlap between the two groups. This is sometimes more common in the tails of the density distribution, suggesting that units most (or least) likely to participate in the intervention are very different from the large majority of cases. One may also observe substantial distance from the cases at the very tail of the distribution to the cases with the next largest (or smallest) propensity scores.

A common strategy for addressing this problem is to trim the observations that fall outside the region of common support. It is important to note, however, that the subsequent estimation is only valid for the subpopulation within the common support. One can check the sensitivity of the propensity score estimation results to the exclusion of observations or their trimming from the sparser tails of the propensity score distributions.

7. Addressing Unobserved Heterogeneity: *diff-in-diff* matching

In some cases, the conditional independence assumption is clearly not met because units are selected into an intervention on the basis of unmeasured characteristics that are expected to influence outcomes. The example from section II, where the more motivated teachers self-select into the program, clearly illustrates the point. Since motivation is typically not observable to the researcher, it cannot be introduced in the model, and thus, the matching estimator will be unable to isolate the impact of the treatment from the effect of the motivation. In fact, in the case of self-selection, it is usually reasonable to think that unobserved variables (like ability, intelligence, motivation, risk aversion) may critically determine the participation model. Unfortunately, we know from previous sections that the usual matching estimator may be seriously biased in case of *selection-on-unobservables*.

However, if pretreatment data are available, the strong Conditional Independence Assumption may be relaxed. More precisely, under the assumption that unobserved variables are *time-invariant* (that is, their value does not change with time), the effect can be cancelled out by taking the difference in outcomes before and after the program.

The implementation of the *difference-in-differences* (or *diff-in-diff*) *matching estimator* is very similar to the cross-sectional version, except that outcome is measured in changes (between the pretreatment and post-treatment periods) instead of in levels. For treated cases, the dependent variable is the difference between outcomes in a period following participation and prior to participation, and for comparison cases, the outcome difference is calculated over the same periods. Even if participating units differ in important ways from those in the comparison group, so long as such differences are stable over time in their influence on outcomes, this specification can eliminate bias resulting from differences between participants and nonparticipants. Letting t and t' represent the pretreatment and post-treatment periods, respectively, the outcome for individual i will be:

$$\Delta Y_i = Y_{it'} - Y_{it}$$

Note how this specification allows us to relax the CIA (*Conditional Independence Assumption*): the counterfactual outcome of the treated individuals is allowed to differ from the observed outcome of the untreated, as long as their *trend* is the same. In technical terms:

$$E\left(Y_{0t'} - Y_{0t} \mid D = 1, X\right) = E\left(Y_{0t'} - Y_{0t} \mid D = 0, X\right) for X \in S$$

where S is defined as the overlapping support among the treatment and comparison groups. In other words, even if treated units differ in important ways from comparison units, as long as such differences are stable over time in their impact on outcomes, this specification can eliminate bias resulting from differences between treated and untreated units (i.e., it allows for unobserved heterogeneity).

The diff-in-diff matching estimator is simply implemented by calculating the propensity score on the baseline year and applying the steps described above to the differenced outcome.

It is worth noting that it may still be important to control for unit characteristics (X) that do not change over time. For example, if individuals with higher levels of education experience greater growth over time in earnings, it may be necessary to match individuals with the same levels of education. Only if the change in outcomes is not associated with a particular characteristic is it appropriate to omit that measure.

Despite the benefits of difference-in-differences estimates, depending on the processes underlying the dynamics of program participation and outcomes, estimates may have biases that are not present in cross-sectional matching. If prior outcomes incorporate transitory shocks that differ for treatment and comparison units, since difference-in-differences estimation interprets such shocks as representing stable differences, estimates will contain a transitory component that does not represent the true program effect. More generally, the difference-in-differences estimates need to be understood as one of several estimates that rely on different assumptions.

Finally, another source of heterogeneity in effects may arise from different dosages of the treatment, which are neglected in models that record treatment (or participation) with a binary variable. If individuals or other units of analysis receive different levels of treatment that are influential in determining outcomes, an alternative matching technique, the generalized propensity score (GPS), can be applied to estimate the effects of different lengths of exposure to treatment on outcomes. Appendix 1 provides additional details on this matching estimation technique.

8. Conclusion

Propensity-score matching is one of the most commonly used techniques for dealing with biases associated with observable factors when evaluating the impact of a program. In this document, we have described the main issues to consider when implementing this methodology, which can be summarized in the following three steps:

1. The participation model must be characterized and the probability of participation predicted. A key objective of this step is to include variables that are likely to affect both the participation and the outcome of interest so that, conditional on these measured variables, there are no unmeasured factors affecting either participation or the relevant nonparticipation outcome. These covariates are used to estimate the propensity score with a probit or logit model.
2. Treated units are matched to similar untreated units based on the proximity of their propensity scores. At this point, a matching algorithm has to be chosen among the different alternatives (nearest neighbor, radius, kernel, etc.) considering data issues (such as sample sizes) and the bias/efficiency trade-off.
3. Once each treated unit has been matched with one or more untreated units, the impact of the program is estimated as a weighted average of the difference in outcomes between treated and untreated. These results need to be complemented with evidence of covariate balancing between groups and robustness checks.

Perhaps the most important issue to understand when implementing PSM is in which contexts it is more likely to work. As mentioned, PSM requires two main conditions to correctly estimate the impact of a program. The first, the Conditional Independence Assumption (or selection-on-observables condition), holds when assignment to treatment is determined only by observable characteristics. If participation is likely to be driven by factors that are not observable to the researcher, the matching estimator may be seriously biased. However, in presence of pretreatment information, a modified version, the difference-in-differences matching estimator, may be applied to correct for some of this bias, as long as the effect of unobserved factors is fixed over time.

The second assumption, known as the Common Support or Overlap Condition, requires the existence of a substantial overlap between the propensity scores of treated and untreated

units. If this assumption does not hold, it is impossible to construct a counterfactual to estimate the impact of the program.

It is crucial, therefore, to carefully evaluate if these two conditions are met before implementing the approach described in this document. At this point, a solid understanding of the program and a sound theoretical basis are essential to defining whether the PSM methodology is an appropriate technique to estimate the impact of interest.

References

Abadie, A., D. Drukker, J.L. Herr, and G. Imbens. 2004. "Implementing Matching Estimators for Average Treatment effects in Stata". *The Stata Journal* 4(3): 290-311.

Abadie, A., and G. Imbens. 2006. "Large Sample Properties of Matching Estimators for Average Treatment Effects". *Econometrica* 74(1): 235-267.

Abadie, A., and G. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators". *Econometrica* 76(6): 1537–1557.

Agüero, J., M. Carter, and I. Woolard. 2007. "The impact of unconditional cash transfers on nutrition: the South African Child Support Grant". Working Papers 39, *International Policy Centre for Inclusive Growth.*

Almus, M., and D. Czarnitzki. 2003. "The Effects of Public R&D Subsidies on Firms' Innovation Activities: The Case of Eastern Germany". *Journal of Business & Economic Statistics* 21(2): 226-236.

Becker, S., and A. Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Score". *The Stata Journal* 2(4): 358-377.

Black, D., and J. Smith. 2004. "How Robust is the Evidence on the Effects of the College Quality? Evidence from Matching". *Journal of Econometrics* 121(1): 99-124.

Caliendo, M., and S. Kopeinig. 2005. "Some Practical Guidance for the Implementation of Propensity-score matching". *Iza Discussion Paper* 1588. Institute for the Study of Labor (IZA).

Dehejia, R., and S. Wahba. 2002. "Propensity-score matching Methods for Nonexperimental Causal Studies". *The Review of Economic and Statistics* 84(1): 151-161.

Flores-Lagunes, A., A. Gonzalez, and T. Neumann. 2007. "Estimating the Effects of Length of Exposure to a Training Program: The Case of Job Corps". *IZA Discussion Papers* 2846, Institute for the Study of Labor (IZA).

Galiani, S., P. Gertler. and E. Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality". *Journal of Political Economy* 113(1): 83-120.

Heckman, J., H. Ichimura, and P. Todd. 1998. "Matching as an Econometric Evaluation Estimator". *The Review of Economic Studies* 65(2): 261-294.

Hirano, K., and G. Imbens. 2004. “The Propensity Score with Continuous Treatments”. Mimeographic document.

Holland, P. 1986. “Statistics and Causal Inference”. *Journal of the American Statistical Association* 81(396): 945-960.

Imai, K., and D. Van Dijk. 2004. “Causal Inference with General Treatment Regimes: Generalizing the Propensity Score”. *Journal of the American Statistical Association* 99(467): 854-866.

Imbens, G., and J. Wooldridge. 2009. “Recent Developments in the Econometrics of Impact Evaluation”. *Journal of Economic Literature* 47(1): 5-86.

Imbens, G.W. 2008. “Estimating Variances for Estimators of Average Treatment Effects”. Mimeographic document.

Jalan, J., and M. Ravallion. 2003. “Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching”. *Journal of Business & Economic Statistics* 21(1): 19-30.

Lavy, V. 2002. “Evaluating the Effect of Teachers' Group Performance Incentives on Pupil Achievement”. *The Journal of Political Economy* 110(6): 1286-1317.

Lechner, M. 1999. “The Effects of Enterprise-Related Training in East Germany on Individual Employment and Earnings”. *Annales d'Économie et de Statistique* 55/56: 97-128.

Lechner, M. 2001. “Program Heterogeneity and Propensity-score matching: An Application to the Evaluation of Active Labor Market Policies”. *The Review of Economics and Statistics* 84(2): 205-220.

Leuven, E., and B. Sianesi. 2003. "PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity-Score Matching, Common Support Graphing, and Covariate Imbalance Testing. *Statistical Software Components S432001 (revised May 2009)*. Newton, MA, United States: Boston College Department of Economics.

Maffioli, A., M. Valdivia, and G. Vázquez. 2009. “Impact of a Technology Transfer Program on Small Farmers: The Case of Ecuador’s PROMSA”. Mimeographic document.

Moser, P. 2005. “How Do Patent Laws Influence Innovation? Evidence from Nineteenth-Century World's Fairs”. *The American Economic Review* 95(4): 1214-1236.

Persson, T., G. Tabellini, and F. Trebbi. 2003. “Electoral Rules and Corruption”. *Journal of the European Economic Association* 1(4): 958-989.

Rosenbaum, P., and D. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects”. *Biometrika* 70(1): 41-55.

Rosenbaum, P., and D. Rubin. 1985. “Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score”. *The American Statistican* 39: 33-38.

Smith, J., and P. Todd. 2005. “Does matching overcome Lalonde’s critique of nonexperimental estimators?”. *Journal of Econometrics* 125(1-2): 305-353.

Trujillo, A., J. Portillo, and J. Vernon. 2005. “The Impact of Subsidized Health Insurance for the Poor: Evaluating the Colombian Experience Using Propensity-score matching”. *International Journal of Health Care Finance and Economics* 5(3): 211-239.

Appendix 1: *Some Technical Aspects of PSM*

The theoretical basis of the Propensity Score-Matching method lies in the results derived by Rosenbaum and Rubin (1983). The main idea can be summarized using the following theorems (see Rosenbaum and Rubin, 1983 for further details).

Theorem 1*: the Propensity Score* $p(X) = P(D = 1 \mid X)$ *is a balancing score,*

$$X \perp D \mid p(X)$$

Theorem 2*: if the Conditional Independence Assumption (CIA) holds, then the potential outcomes are independent of the treatment status, conditional on the propensity score* $p(X)$

$$(Y_1, Y_0) \perp D \mid X \Rightarrow (Y_1, Y_0) \perp D \mid p(X)$$

In other words, the first theorem states, roughly, that the characteristics between two groups with the same propensity score value will be balanced. The second theorem implies that conditioning on the propensity score is equivalent to conditioning on the full vector X, as long as this vector contains all the relevant information to satisfy the CIA. Then, rather than attempting to match on all values of X, cases can be compared on the basis of propensity scores alone.

The *average treatment effect on the treated* (ATT) can be estimated using these results. More precisely, recall that the ATT is:

$$ATT = E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1)$$

The CIA implies

$$E(Y_0 \mid D = 1, X) - E(Y_0 \mid D = 0, X)$$

And using theorem 2,

$$E(Y_0 \mid D = 0, X) - E(Y_0 \mid D = 0, p(X))$$

Combining these results, the Law of Iterated Expectations can be used to show that

$$\begin{aligned} ATT &= E(Y_1 \mid D = 1) - E(Y_0 \mid D = 1) \\ &= E_{p(x)|D=1}\left[E(Y_1 \mid D = 1, p(X)) - E(Y_0 \mid D = 0, p(X))\right] \end{aligned}$$

Note that, when estimating the ATT, the CIA assumption can be relaxed to:

$$Y_0 \perp D \mid X$$

since we need only to construct the counterfactual for the treated individuals. In practice, the matching estimator takes the form:

$$\widehat{ATE} = \frac{1}{N_T} \sum_{i \in I_1 \cap Sp} \left\{ Y_{1i} - \sum_{i \in I_1 \cap Sp} \omega_{ij} Y_{0J} \right\}$$

where N_T is the number of treated individuals, I_1 and I_0 are the sets containing treated and untreated individuals respectively, S_p is the common support and ω_{ij} is a weight assigned to each untreated individual based on the propensity score (its exact shape depends on the matching algorithm).

Standard Error Estimation

Abadie and Imbens (2008), Imbens and Wooldridge (2009), and Imbens (2008) recommend the use of an analytical formula for calculating standard errors that is asymptotically correct. Abadie and Imbens (2006) show that for matching estimates using a fixed number of matches, bootstrap standard errors are asymptotically biased. However, there is no work indicating whether bootstrap standard errors for radius-matching methods are consistent (see Imbens and Wooldridge, 2009).

The alternative approach recommended by Imbens and Woodridge (2009) and Imbens (2008) produces a conditional standard error, which provides an estimate of the variation in an impact estimate, conditional on the independent variables. Abadie and Imbens (2008) suggest an approach for estimating the unconditional standard error, which provides an estimate of the variation in the impact estimate that would result if the sample were chosen repeatedly from the full universe, with values on independent variables varying from sample to sample. The true value of the unconditional standard error must exceed the conditional standard error, but there is no certainty this will be the case for the estimates obtained in any one sample. Both approaches are somewhat involved and require the use of a matching algorithm that is as computer-intensive as that required to obtain program-impact estimates.

Generalizing the Propensity Score: Multiple Treatments and Dosage Effects

In cases where the treatment consists of multiple alternatives, a multinomial first-stage model may be estimated. The multinomial logit model imposes stronger assumptions than the multinomial probit model, although the multinomial probit is also more computationally intensive. Another option is to estimate a series of binomial models (see Lechner, 2001), where only two treatment options are considered at one time. This increases substantially the number of models to be estimated, and the results of each estimation apply only to the two selected groups. Lechner (2001) did not find significant differences in the performance of the multinomial probit approach and estimation of a series of models.

In many impact evaluations, treatment (or participation) is recorded with a binary variable equal to one if participation occurs and zero if no treatment is received. An important limitation of this approach to estimation is that it does not allow for exploration of the effects of differing levels of exposure to treatment, particularly in voluntary interventions where varying length of participation is a common feature. If individuals or other units of analysis receive different levels of treatment, then the average treatment effect estimated by conventional estimators is unlikely to capture the heterogeneity in effects arising from different dosages of the treatment. In other words, to the extent that exposure to treatment is influential in determining outcomes, the standard propensity score-matching approach to estimation may produce information that is of limited use to program/intervention designers.

In recent years, a new matching technique for evaluating interventions in multitreatment settings, the generalized propensity score (GPS), has been developed to estimate the causal effects of different lengths of exposure to treatment on outcomes. Similar to the unconfoundedness assumption made in PSM, the GPS approach assumes that selection into *levels* of the treatment is random, conditional on a set of rich observable characteristics. In this case, the *level* of participation is independent of the outcome that would occur in the absence of participation. If the model assumptions are satisfied, it is possible to use GPS to estimate the average treatment effects of receiving different levels of exposure to the treatment or intervention, thereby allowing for the construction of a "dose-response function" that shows how treatment exposure relates to outcomes.

The notation for GPS has been formalized by Hirano and Imbens (2004). They define $Y_i(t)$ as the set of potential outcomes of a treatment $t \in \mathcal{T}$, where $\mathcal{T}$ may be an interval of a

continuous treatment. For each unit *i*, we observe a vector of covariates X_i (that predict take-up of the treatment); the level of the treatment, T_i, that unit *i* actually receives, and the potential outcome associated with the level of treatment received, $Y_i = Y_i(t)$.

The (weak) unconfoundedness assumption states that, conditional on observed covariates, the level of treatment received (T_i) is independent of the potential outcome

$$Y_i(t) \perp T_i \mid X_i \qquad \forall t \in \mathcal{T}$$

In other words, we assume there is no systematic selection into levels of treatment based on unobservable characteristics.

The density function of the treatment conditional on pretreatment covariates is:

$$r(t,x) = f_{T|X}(t \mid x)$$

and the GPS is therefore defined as the conditional density of receiving a particular level of the treatment, $t = T$:

$$R = r(T,X)$$

Similar to the binary treatment (PSM) case, the GPS balances the covariates *within strata* defined by values of the GPS, so that the probability that $t=T$ does not depend on the value of X (and assignment to treatment *levels* is unconfounded).

As Hirano and Imbens (2004) show, this allows the estimation of the average dose-response function, $\mu(t) = E\left[Y_i(t)\right]$ using the GPS to remove selection bias.

After estimating the GPS, the next step is to estimate the conditional expectation of the outcome (Y_i) as a function of the treatment level, T, and the GPS, R:

$$\beta(t,r) = E[Y \mid T = t, R = r]$$

The regression function $\beta(t,r)$ represents the average potential outcome for the strata defined by $r(T,X) = R$, but it does not facilitate causal comparisons across different levels of treatment. That is, one cannot directly compare outcome values for different treatment levels to obtain the causal difference in the outcome of receiving one treatment level versus another.

A second step is required to estimate the dose-response function at each particular level of the treatment. This is implemented by averaging the conditional means $\beta(t,r)$ over the distribution of the GPS, $r(t,X)$, i.e., for each level of the treatment:

$$\mu(t) = E\left[\beta(t, r(t, X))\right]$$

Where $\mu(t)$ corresponds to the value of the dose-response function for treatment value *t*, and when compared to another treatment level, does have a causal interpretation.

Implementation of GPS estimation

Assuming a normal or lognormal distribution for the treatment, ordinary-least-squares (OLS) regression can be used to estimate the GPS, i.e., the conditional distribution of the treatment T_i given the covariates, X_i, that predict selection into levels of the treatment. It is also possible to assume other distributions and/or to estimate the GPS by other methods such as maximum likelihood.

Next, the conditional expectation of the outcome given the observed treatment level (T_i) and the estimated GPS (e.g., R_i) is modeled with a flexible linear specification (such as OLS). Hirano and Imbens (2004) recommend a quadratic approximation, such as the following:

$$E\left[Y_i \mid T_i, R_i\right] = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 R_i + \alpha_4 R_i^2 + \alpha_2 T_i . R_i$$

In the third step, the value of the dose-response function at treatment level *t* is estimated by averaging the regression function (from step two above) over the distribution of the GPS (holding constant the treatment level *t*). Bootstrap methods can be used to obtain standard errors that take into account estimation of the GPS and regression parameters α_0 to α_5.

As in PSM, it is also important to assess the balance of the covariates following GPS estimation. One approach suggested by Imai and van Dijk (2004) is to check the balance of each covariate by running a regression of each covariate on the log of the treatment and the GPS; if the covariate is balanced, then the treatment variable should have no predictive power, conditional on the GPS. A comparison of this coefficient to the corresponding coefficient of a regression that does not include the GPS can be used to assess how well the GPS performs in balancing.

In an alternative approach to assessing balance, Agüero, Carter and Woolard (2007) defined three different treatment terciles of the treatment variable and tested whether the mean value of the covariates were the same for the observations in the different treatment terciles. They then investigated whether the covariates were better balanced after conditioning on the estimated GPS. For each treatment tercile, they first calculated the estimated probability that

each observation might have received the median treatment level for the tercile. In other words, letting d_t denote the median treatment level received in tercile t, they calculated $r(d_t,X_i)$ for each observation. They then separated the observations into five quintiles, and for each GPS quintile block, they tested whether the means of the covariates for the observations that actually received low treatment were different from the means for those that did not receive low treatment. If the GPS successfully balanced the covariates, low and not-low treatment groups would look similar after conditioning on the GPS.

Checking for sufficient overlap in the distribution of covariates across different levels of the treatment is likewise important, although it is substantially more difficult in the multi-valued or continuous treatment case, given the many levels of treatment and multiple parameters of interest that may require different support conditions.

Flores-Lagunes, Gonzalez and Neumann (2007) developed an approach to informally assess overlap in the supports of different levels of treatment in their study of youth exposure to a job-training intervention. They began by dividing values of the treatment-exposure measure into five quintiles. For each quintile, they computed the value of the GPS for each unit (youth) at the median level of the treatment for the quintile. They then computed the value of the GPS at the same median level of treatment for all individuals that were not part of that particular quintile. They subsequently compared the supports of the values of the GPS for these groups (individuals in the quintile of focus with those of the other quintiles) by examining (superimposing) their histograms. This was repeated in turn for each quintile, generating plots for each of the comparisons.

Appendix 2: Software for Implementing PSM

Below we summarize information on the most commonly used software programs for propensity-score matching. This list of available programs and publications is adapted from a web page maintained by Elizabeth Stuart:

http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html

In addition, although there are currently no procedures or macros from the SAS Institute specifically designed to match observations using propensity scores, papers presented at an SAS Global Forum (and listed below) offer some macros that may be employed.

Stata

psmatch2 http://ideas.repec.org/c/boc/bocode/s432001.html

- Leuven, E., and B. Sianesi. 2003. "PSMATCH2: Stata Module to Perform Full Mahalanobis and Propensity-Score Matching, Common Support Graphing, and Covariate Imbalance Testing". *Statistical Software Components S432001(revised 02 May 2009)*. Newton, MA, United States: Boston College Department of Economics.
- Allows k:1 matching, kernel weighting, Mahalanobis matching
- Includes built-in diagnostics
- Includes procedures for estimating ATT or ATE

pscore http://www.lrz-muenchen.de/~sobecker/pscore.html

- Becker, S.O. and A. Ichino. 2002. "Estimation of Average Treatment Effects Based on Propensity Score". *The Stata Journal* 2(4): 358-377.
- k:1 matching, radius (caliper) matching, and stratification (subclassification)
- For estimating the ATT

match http://www.economics.harvard.edu/faculty/imbens/software_imbens

- Abadie, A., D. Drukker, J.L. Herr, and G. Imbens. 2004. "Implementing Matching Estimators for Average Treatment effects in Stata". *The Stata Journal* 4(3): 290-311.
- Primarily k:1 matching (with replacement)

- Allows estimation of ATT or ATE, including robust variance estimators

cem http://gking.harvard.edu/cem/

- Iacus, S.M., G. King, and G. Porro. 2008. "Matching for Causal Inference without Balance Checking".
- Implements coarsened exact matching

R

Matching http://sekhon.berkeley.edu/matching

- Sekhon, J. S. (in press). "Matching: Multivariate and Propensity-Score Matching with Balance Optimization". Forthcoming, *Journal of Statistical Software.*
- Uses automated procedure to select matches, based on univariate and multivariate balance diagnostics
- Primarily 1:M matching (where M is a positive integer), allows matching with or without replacement, caliper, exact
- Includes built-in effect and variance estimation procedures

MatchIt http://gking.harvard.edu/matchit

- Ho, D.E., K. Imai, G. King, and E.A. Stuart. (In press). "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference". Forthcoming, *Journal of Statistical Software.*
- Two-step process: does matching, then user does outcome analysis (integrated with Zelig package for R)
- Wide array of estimation procedures and matching methods available: nearest neighbor, Mahalanobis, caliper, exact, full, optimal, subclassification
- Built-in numeric and graphical diagnostics

cem http://gking.harvard.edu/cem/

- Iacus, S.M., G. King, and G. Porro. 2008. "Matching for Causal Inference Without Balance Checking".
- Implements coarsened exact matching
- Can also be implemented through MatchIt

optmatch http://cran.r-project.org/web/packages/optmatch/index.html

- Hansen, B.B., and M. Fredrickson. 2009. "Optmatch: Functions for Optimal Matching".
- Variable ratio, optimal, and full matching
- Can also be implemented through MatchIt

PSAgraphics http://cran.r-project.org/web/packages/PSAgraphics/index.html

- Helmreich, J.E., and R.M. Pruzek. 2009. "PSAgraphics: An R Package to Support Propensity Score Analysis". *Journal of Statistical Software* 29(6).
- Functions that primarily produce graphics to test balance within strata of categorical and quantitative covariates, represent estimated effect size by stratum, and various balance functions that provide measures of the balance achieved.

SAS

SAS usage note: http://support.sas.com/kb/30/971.html

Greedy matching (1:1 nearest neighbor)

- Parsons, Lori S. 2001. "Reducing Bias in a Propensity Score Matched-Pair Sample Using Greedy Matching Techniques". SAS SUGI 26, Paper 214-26.
- Parsons, Lori S. 2005. "Using SAS® Software to Perform a Case-Control Match on Propensity Score in an Observational Study". SAS SUGI 30, Paper 225-25.