WPS 3877

PROPENSITY SCORE MATCHING AND POLICY IMPACT ANALYSIS
A Demonstration in EViews

B. Essama-Nssah[*]
Poverty Reduction Group (PRMPR)
The World Bank
Washington, D.C.

Abstract

Effective development policymaking creates a need for reliable methods of assessing effectiveness. There should be, therefore, an intimate relationship between effective policymaking and impact analysis. The goal of a development intervention defines the metric by which to assess its impact, while impact evaluation can produce reliable information on which policymakers may base decisions to modify or cancel ineffective programs and thus make the most of limited resources. This paper reviews the logic of *propensity score matching* (PSM) and, using data on the National Support Work Demonstration, compares that approach with other evaluation methods such as double difference, instrumental variable and Heckman's method of selection bias correction. In addition, it demonstrates how to implement *nearest-neighbor* and *kernel-based methods*, and plot program incidence curves in EViews. In the end, the plausibility of an evaluation method hinges critically on the correctness of the socioeconomic model underlying program design and implementation, and on the quality and quantity of available data. In any case, PSM can act as an effective adjuvant to other methods.

Keywords: EViews, Double Difference, Impact Analysis, Instrumental Variables, Kernel Function, Matching, Propensity Score, Sample Selection.

JEL: C14, C81, C87, I38.

World Bank Policy Research Working Paper 3877, April 2006

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the view of the World Bank, its Executive Directors, or the countries they represent. Policy Research Working Papers are available online at *http://econ.worldbank.org*.

[*] The author would like to thank Glenn Sueyoshi (Quantitative Micro Software) and Jian Zhang (University of California at Davis) for very helpful suggestions. He is also grateful to Pedro Olinto (LCSPP) and Markus Goldstein (PRMPR) for insightful comments on an earlier draft of the paper. Finally, thanks are due to Martin Ravallion (DECRG) and Vijdan Korman (PRMPR) for information sharing and useful discussions.

TABLE OF CONTENTS

Tables

Figures

Boxes

Appendices

1. Introduction

Effective development policymaking creates a need for reliable methods of assessing whether an intervention had (or is having) the intended effect. *There should be therefore, an intimate relationship between effective policymaking and impact analysis.* The goal of an intervention defines the metric by which to assess its effectiveness. Effective methods of evaluation produce reliable information on what works and why, and policymakers may use such information to modify or cancel ineffective programs and thus make the most of limited resources (Grossman 1994).

The assessment of the impact of a program (or a development intervention) requires *a model of causal inference*. Holland (1986) specifies such a statistical model. He starts from the fundamental observation that the effect of a cause can be understood only in relation to another cause. This is the same idea underlying the economic principle of assessing the return to a resource employed in one activity relative to its opportunity cost (i.e. what it would have earned in the next best alternative use). Thus we can assess the effect of a development intervention only if we know what would have happened without such an intervention.

Consider a simple situation involving only two causes: program participation versus nonparticipation. A statistical causal inference model applicable to such a case involves the following elements: (1) a population of units upon which causes or interventions may act (e.g. individuals, households, districts, firms or regions), (2) an assumption that each unit is potentially exposable to the causes, (3) an observable variable[1] $\mathbf{d}$, indicating the cause to which a given unit is exposed (e.g. $\mathbf{d=1}$ for exposure, and zero otherwise), (4) a set of variables representing pre-exposure attributes for each unit (some attributes may be observable, call them $\mathbf{x}$, and some not, call these $\boldsymbol{\varepsilon}$); and (5) a variable, $\mathbf{y(d)}$, representing the potential response of unit to exposure. In fact, $\mathbf{y}$ represents two variables standing for two potential responses: $\mathbf{y_1}$ under exposure, and $\mathbf{y_0}$ if no exposure.

[1] Recall that a variable in this context may be viewed as a real-valued function defined over a unit. The value taken by such a variable for a given unit is an outcome of a measurement process applied to the unit (Holland 1986).

Within the above framework, the effect of exposure on unit **i** is measured relative to non-exposure on the basis of the response variable **y**. If we call this effect, $\mathbf{g_i}$, then $g_i = (y_{1i} - y_{0i})$. It is impossible to observe the value of the response variable for the same individual under two mutually exclusive states of nature (exposure and non-exposure). This *Fundamental Problem of Causal Inference* (Holland 1986) makes it impossible to observe the effect of exposure on unit **i**. This is why evaluation methods are considered as ways of dealing with this *missing data* problem. If the intervention is limited to a subset of the population, many of the methods suggest turning to non-exposed units (non-participants) in search of the missing information. They also specify circumstances under which the use of such information yields reliable estimates of the relevant effect.

The assumption of *unit homogeneity* (Holland 1986) characterizes a benchmark case where the effect on individual **i** could be reliably estimated. An individual response is a function of participation, observable and unobservable characteristics. Suppose we can find among non-participants an individual **j** with the same pre-exposure (observable and non-observable) attributes as participant **i**. Thus, under unit homogeneity, the outcome of this non-participant is a proxy for what would have happened to **i** hadn't she received the intervention. Hence, the effect of the intervention on **i** can be estimated as: $g_i = (y_{1i} - y_{0j})$. The assumption of unit homogeneity is thus analogous to the *ceteris paribus* assumption used in scientific enquiry. The assumption serves as a benchmark case against which to assess the implications of heterogeneity. In non-exposure state, one would generally expect response heterogeneity for participants and non-participants, particularly when eligible candidates are given the choice to participate or not[2].

The most common impact indicator of interest is the mean impact of treatment on the treated. It is also known as the *average treatment effect on the treated* (ATET). Here treatment means exposure to a cause or participation in a social program. Let $g = (y_1 - y_0)$, then the mean impact on the treated can be written as a conditional mean:

[2] Heckman and Smith (1995) cite the case where those who choose to join a social program do so because of the poor alternative they face outside the program. In such a case, non-participants would have better outcomes than participants had the latter not elected to participate. This response heterogeneity is also known as *selection bias*.

$ATET = E(g \mid x, d = 1) = E(y_1 \mid x, d = 1) - E(y_0 \mid x, d = 1)$. The missing data here relates to the counterfactual mean $E(y_0 \mid x, d = 1)$. One might be tempted to use the mean outcome for nonparticipants $E(y_0 \mid x, d = 0)$as a proxy for the above counterfactual mean. However, Heckman and Smith (1995) caution that subtracting the mean response for nonparticipants from the mean outcome of participants yields an estimate[3] which is equal to the average treatment effect on the treated (the parameter of interest) plus *selection bias*. Selection bias stems from the failure of the assumption of unit homogeneity. In general, nonparticipants differ from participants in the nonparticipation state. This heterogeneity may be due to observable or unobservable characteristics.

There are both experimental and nonexperimental ways of dealing with selection bias. In the case of social *experiments*, treatment is assigned randomly so that participation is statistically independent of potential outcomes. Thus, the control group is composed of individuals who would have participated but were denied access randomly. Heckman and Smith (1995) explain that the mean outcome of the control group provides an acceptable estimate of the counterfactual mean if randomization does not alter the pool of participants or their behavior, and if no close substitutes for the experimental program are readily available[4]. These authors further explain that randomization does not eliminate selection bias, but rather balances it between the two samples (participants and nonparticipants) so that it cancels out when computing the mean impact. This balancing act can be understood on the basis of the following consideration. Random assignment of treatment ensures that every eligible candidate has the same chance ex ante of being treated. *Therefore the distribution of both observed and unobserved characteristics prior to treatment is the same for both the treated and the control group*.

In the context of observational (or non-experimental) studies for causal effects, there is no direct estimate of the counterfactual mean analogous to the one based on randomization. This paper focuses on a popular class of impact estimators. These estimators rely on *propensity score matching* (PSM) which originated with Rosenbaum

[3] $E(y_1 \mid x, d = 1) - E(y_0 \mid x, d = 0) = ATET + [E(y_0 \mid x, d = 1) - E(y_0 \mid x, d = 0)]$

[4] There would be randomization bias if those who participate in an experiment differ from those who would have participated in the absence of randomization. Furthermore, substitution bias would occur if members of the control group can easily obtain elsewhere close substitutes for the treatment (Heckman and Smith 1995).

and Rubin (1983). The purpose of the paper is to review the logic of this matching method, and illustrate its implementation and the computation of related impact indicators in EViews. PSM is an algorithm that matches treated and nonparticipants on the basis of the conditional probability of participation (the propensity score), given the observable characteristics. If outcomes are independent of participation, conditional on observables, then the use of the matched comparison group would yield an unbiased estimate of the mean impact of treatment.

The paper is organized as follows. Section 2 reviews the structure of two types of matching algorithms. The nearest-neighbor method pairs a given participant with the member of the comparison group with the propensity score that is closest to that of the given participant. Kernel-based methods associate with the outcome of participant **i** a matched outcome computed as a kernel-weighted average of the outcomes of all non-participants. This section also provides a brief comparison of PSM with evaluation methods such as double difference (DD), instrumental variable (IV) and Heckman's method of selection bias correction. A more detailed description of these other methods is presented in Appendix C. Section 3 uses well-known and publicly available data sets to illustrate how to implement in EViews the methods described in section 2. Data on the treated come from the National Supported Work (NSW) Demonstration. The comparison group is drawn from the Population Survey of Income Dynamics (PSID) [Dehejia and Wahba 2002, 1999; Becker and Ichino 2002]. Concluding remarks are made in section 4. Appendix A shows how to estimate the propensity score using the log likelihood object (LOGL) while Appendix B provides the entire computer code for PSM.

2. Matching Methods

The Principle of Matching

Matching methods can be framed within the context of *nonparametric estimation* of the relation between an outcome variable for unit **i** ($\mathbf{y_i}$), a dummy variable indicating participation in the program ($\mathbf{d_i}$), and set of other characteristics ($\mathbf{x_i}$). These characteristics are also referred to as covariates and are assumed exogenous in the sense that they are not affected by the intervention. Such a relation can be stated as (Moffit

2004): $y_i = f(d_i, x_i)$. In the context of observational studies, the key assumptions underlying matching methods seek to mimic conditions similar to an experiment so that the assessment of the impact of the program can be based on a comparison of outcomes for a group of participants (i.e. those with **d_i=1**) with those drawn from a comparison group of non-participants (**d_i=0**).

To yield consistent estimates of program impact, matching methods rely on a fundamental assumption known as "*conditional independence*" or "*selection on observables*"[5]. This assumption can be formally stated as[6]:

$$(y_0, y_1) \perp d \mid x \tag{2.1}$$

The above expression states that potential outcomes are orthogonal to treatment status, given the observable covariates. In other terms, conditional on observable characteristics, participation is independent of potential outcomes. Assuming that there are no unobservable differences between the two groups after conditioning on **x_i**, any systematic differences in outcomes between participants and nonparticipants are due to participation If one is only interested in the mean impact for the treated, then the assumption of unconfoundedness can be weakened by focusing on potential outcomes in the nonparticipation state (Imbens 2004). This weaker version can be stated as follows.

$$y_0 \perp d \mid x \tag{2.2}$$

In other terms, the outcome in the counterfactual state is independent of participation, given the observable characteristics. Thus, conditional on the observables, outcomes for the non-treated (the comparison group) represent what the participants would have experienced had they not participated in the program. Obviously, this makes sense in the particular situation where selection into the program is based entirely on

[5] This assumption is also known as the *exogeneity* or *unconfoundedness* assumption or *ignorable treatment assignment* (Imbens 2004)

[6] The symbol $\perp$ represents orthogonality between two variables. Thus conditional independence may also be referred to as conditional orthogonality.

observable characteristics[7]. In order to solve the fundamental missing data problem, all we have to do is to find for each participant, one or more nonparticipants with the same values of observables. This is where matching comes in. In general, matching estimators of program effect impute the missing potential outcomes using only the outcomes of the matched individuals from the comparison group.

For matching to be feasible, there must be individuals in the comparison group with the same values of the covariates as the participant of interest. This requires an *overlap* in the distribution of observables between the treated and the comparison groups. The overlap assumption is usually stated as:

$$0 < \Pr\left(d = 1 \mid x\right) < 1 \tag{2.3}$$

A weaker version of the overlap assumption requires only the following (Imbens 2004).

$$p(x) = \Pr\left(d = 1 \mid x\right) < 1 \tag{2.4}$$

This implies the possible existence of a nonparticipant analogue for each participant. This is all that is required for the estimation of the mean impact on the treated (Smith and Todd 2005a). When this condition is not met, then it would be impossible to find matches for a fraction of program participants.

To fully appreciate the point of the overlap assumption, consider situations where, for some values of **x**, we have either $p(x) = 0$ or $p(x) = 1$. Individuals with such covariates are such that either they never receive treatment or they always receive it. If they always receive treatment, then they have no counterparts in the comparison group. On the other hand, if they never receive treatment, then they have no counterparts in the treated group. Thus, it would be impossible to use matching methods on such cases (Heckman, Ichimura and Todd 1998). In these circumstances, it is recommended to restrict matching and hence the estimation of the treatment effect on the region of common support. This implies using only nonparticipants whose propensity scores overlap with those of the participants.

[7] Hence the name "selection on observables" for this orthogonality assumption which implies that unobservables play no role in determining participation (Dehejia and Wahba 2002).

Assuming selection on observables, proper matching requires that we select from the sample of non-participants a comparison group in which the distribution of observed characteristics is as similar as possible to the distribution among the participants. In the case of an exact match, the only difference between a participant and her match is that the former received treatment while the latter did not. Hence we may refer to the unconfoundedness assumption as the assumption of *conditional homogeneity* and the overlap assumption as the *feasibility* assumption.

Imbens (2004) makes the following observations about the *plausibility* of the assumption of selection on observables in economic settings. The evaluation of any program ultimately entails the comparison of outcomes for participants and nonparticipants. The key issue then becomes the identification of units that best represent the treated unit had they not participated in the program. Matching analysis based on unconfoundedness is a useful initial step in any serious investigation of program effectiveness. Even in situations where agents do choose their treatment optimally, the assumption of selection on observables may still be valid if the difference in their behavior is driven by unobservables that are uncorrelated to the relevant outcomes. In particular, this might be the case if the objective of the decision maker is distinct from the outcome under consideration.

Diaz and Handa (2004) justify selection on observables in the context of PROGRESA[8] on the basis of the following features of the program. The inclusion of poor households in the program is based only on observable characteristics of households and the locality in which they reside. The program is mandatory and the rate of noncompliance with treatment is very low. Thus self-selection is not a major concern in this case and, matching provides a reliable approach to assess the impact of this program.

In practice matching may become more and more difficult, the larger the set of observable characteristics underpinning the matching exercise. Rosenbaum and Rubin (1983) show that the *dimensionality* of the matching problem can be significantly reduced by using the propensity score (the conditional probability of participation given the observed covariates). Thus instead of conditioning on an n-dimensional variable, units

[8] *Programa de Educacion, Salud y Alimentatcion* or the Education, Health and Nutrition Program of Mexico also known as *Oportunidades*.

are matched on a scalar variable. This simplification is due to the fact that conditional independence remains valid if we use the propensity score **p(x)** instead of the covariates **x**. Thus weak conditional independence [equation (2.2)] can be now expressed as:

$$y_0 \perp d \mid p(x) \qquad (2.5)$$

The rest of this section focuses on propensity score matching.

Propensity Score Matching Algorithms

Propensity score matching (PSM) pairs observations on the basis of the conditional probability of participation. Wooldridge (2002) motivates PSM with the following thought experiment analogous to unit homogeneity. Select a propensity score **p(x)** at random. Find two units from the population at large with the same score. Let one participate in the program and the other one not. We can use the outcome of the non-participant as a proxy for the outcome she would have experienced had she not joined the program.

Sianesi (2001) explains the basic steps involved in implementing PSM. Assume that we have data on the following: (1) a binary dummy variable identifying participants and non-participants, (2) the outcome to be evaluated, and (3) a set of covariates. First, estimate propensity scores on the covariates using *probit* or *logit* and retrieve their predicted values. Second, pair each participant **i** with some group of comparable non-participants (on the basis of propensity scores). Finally, estimate the counterfactual outcome of participant **i** as the weighted outcomes of her neighbors in the comparison group.

Formally, let **c(p$_i$)** be the set of the neighbors of **i** in the comparison group, then the matched outcome is defined by the following expression.

$$\hat{y}_i = \sum_{j \in c(p_i)} w_{ij} y_j ;\ w_{ij} \in [0, 1]; \quad \sum_{j \in c(p_i)} w_{ij} = 1 \qquad (2.6)$$

This matched outcome is our best guess of what participant **i** would have experienced had she not joined the program.

The specification of a matching algorithm is based on two key considerations. Each method requires the definition of *a measure of proximity* in order to identify nonparticipants who are acceptably close (in terms of the propensity score) to any given participant. This is the criterion that determines **c(p$_i$),** the set of the neighbors of **i** in the comparison group. Finally, we must select *a weighing function* that determines the weight to be assigned to each member of a neighborhood in the computation of the matched outcome according to (2.6). Here we focus on two classes of algorithms, *nearest-neighbor* and *kernel* matching. The nearest-neighbor method assigns a weight of one to the nearest nonparticipant and zero to others. If there are more than one individual in the neighborhood then the method assigns equal weight to each and a zero weight to people outside the neighborhood. The kernel method uses all the members of the comparison group within the common support, and the further away a comparison unit is from the treated one, the lower the weight it receives in the computation of the counterfactual outcome.

In general, matching estimators of the mean impact of treatment on the treated take the following form.

$$\theta_M = \sum_{i \in T} \omega_i \left(y_i - \sum_{j \in c(p_i)} w_{ij} y_j \right) = \sum_{i \in T} \omega_i g_i \qquad (2.7)$$

Where **T** stands for the set of treated, and **ω$_i$** can be interpreted more broadly as the evaluative weight assigned to participant **i**. Indeed, any evaluation entails the following four basic dimensions: (1) identification of the objects of value on the basis of the goal of the intervention; (2) the valuation of such objects through a measurement process (e.g. a survey instrument); (3) an overall (or aggregate) characterization of the social state in which these objects of value are observed; and (4) the ranking of alternative states. Evaluative weights underpin the overall characterization of a social state. The mean impact defined by (2.7) is an example of such a characterization. In standard applications, **ω$_i$** is taken to be the sampling weight associated with observation **i**.

There may be situations where one is interested in the distribution of the program impact. In such situations, the mean impact indicator is not that helpful. One can factor in distributional concerns in the evaluation by considering the incidence of the gains (Ravallion 2003, 2005). In the context of anti-poverty programs for instance, one may be interested in the incidence of welfare gains. This requires knowledge about the welfare impact conditional on pre-intervention welfare. Pre-intervention welfare can be estimated by subtracting welfare gains[9] (**g_i**) from post-intervention welfare for all participants. One could then compare both distributions of welfare.

Such a comparison could be made on the basis of a device analogous to the growth incidence curve (or GIC)[10]. The construction of the device is based on the fact that the distribution of outcomes is fully characterized by the mean outcome and the Lorenz representation of relative inequality. Let **$y_1(p)$** stand for the post intervention out come at percentile **p**, **$y_0(p)$** for the matched outcome in the non-treatment state. Let **$L_1(p)$** and **$L_0(p)$** stand for the Lorenz curves representing relative inequality in both states. The program would have a positive impact at percentile **p**, if the following is true.

$$\frac{y_1(p)}{y_0(p)} = \frac{\mu_1 L_1'(p)}{\mu_0 L_0'(p)} \geq 1 \tag{2.8}$$

Where **μ_s** and **$L_s'(p)$** stand respectively for the mean of the distribution and the first order derivative of the Lorenz function in state **s=0, 1**. Since the logarithm is a monotonic transformation, the above condition is equivalent to the following.

$$g(p) = \gamma + \Delta \ln L'(p) \geq 0 \tag{2.9}$$

[9] Ravallion (2001) illustrates the importance of accounting for opportunity cost in computing the gains for program participation. Using the example of cash transfer program designed to keep poor children in school, he explains that forgone income should be netted out of the cash transfer in order to avoid overestimating the income gains for the program. Indeed, children have to be in school in order to receive the cash transfer. Thus children who are working prior to joining the program would have to forgo their earnings.

[10] In the context of pro-poor growth analysis, the *growth incidence curve* depicts the rate of change in the welfare indicator at each percentile of the distribution due to economic growth (Ravallion and Chen 2003).

By analogy to the GIC, **g(p)** may be called *Program Incidence Curve* (PIC). Thus program incidence at percentile **p** is equal to the rate of change in average outcome between the two states plus a distribution adjustment factor equal to the rate of change of the slope of the Lorenz curve between the two states. The program is considered to have an unambiguous positive social impact if **g(p)≥0** for all **p**. Expression (2.9) also reveals that program incidence at percentile **p** is equal to the average treatment effect on the treated adjusted by a distributional factor based on the slope of the Lorenz curve.

If one is not interested in this decomposition, then a simpler way to proceed is to plot the ratio $\mathbf{y_1/y_0}$ as a function of **p**, the cumulative distribution of the participants ranked in increasing order of the counterfactual outcome. The program would have a positive impact at each percentile where this ratio is greater than one. We may call such a plot *Relative Program Incidence Curve* (RPIC). We show an example in section 3.

Nearest-Neighbor Matching

For each participant **i**, this method searches for the nonparticipant **j** with the closest propensity score. Based on this concept, the relevant neighborhood is defined by the following expression.

$$c(p_i) = \{ j \mid \min_j \| p_i - p_j \| \} \tag{2.10}$$

This formula can be used in matching with or without replacement. Matching with replacement creates the possibility of matching a given non-participant to more than one participant. With respect to the trade-off between bias and variance, replacement improves the quality of matches on average while increasing the variance of the impact estimator (Smith and Todd 2005a). In the case of matching without replacement, once a nonparticipant has found his match he drops out of consideration. Matching without replacement can lead to many poor matches in situations where there are many participants with high values of the propensity score and few nonparticipants with such values. This would lead such participants to be matched with nonparticipants who have quite different observable characteristics. Finally, the quality of the impact estimate based on nearest-neighbor matching without replacement depends on the order in which the observations come in the process.

One can try to avoid poor matches by implementing a variant of the nearest-neighbor approach known as *caliper matching*. This method selects the nearest neighbor within a caliper of width δ. The approach imposes a tolerance level on the distance between the propensity score of participant **i** and that of nonparticipant **j**. Formally, the corresponding neighborhood can be stated as follows (Sianesi 2001).

$$c(p_i) = \{ j \mid \delta > \| p_i - p_j \| = \min_j \| p_i - p_j \| \} \tag{2.11}$$

If there is no member of the comparison group within the caliper for the treated unit **i**, then the treated unit is left unmatched and dropped from the analysis. Thus caliper is a way of imposing the common support restriction. Naturally, there is uncertainty about the choice of a tolerance level.

A variant of caliper matching is known as *radius matching*. In this case, an estimate of the counterfactual is based on the outcomes of all members of the comparison group within the radius **r**, rather than the outcomes of the nearest neighbors within the radius (as in the case of caliper matching). The corresponding neighborhood is:

$$c(p_i) = \{ j \mid r > \| p_i - p_j \| \} \tag{2.12}$$

The nearest neighbor mean impact estimator can be written as:

$$\theta_{NN} = \frac{1}{n_t} \sum_{i \in T} \left(y_i - y_j \right) \tag{2.13}$$

Where $\mathbf{n_t}$ is the total number of treated units.

Kernel Matching

The idea behind kernel-based matching is to associate the outcome of participant **i** with a matched outcome computed as a kernel-weighted average of the outcomes of all non-participants. The weight assigned to non-participant **j** is in proportion to how close she is to participant **i**. These weights are computed as follows:

$$w_{ij} = \frac{K\left(\frac{p_i - p_j}{h}\right)}{\sum_{j \in \{d=0\}} K\left(\frac{p_i - p_j}{h}\right)} \tag{2.14}$$

where **h** stands for the bandwidth. Two kernel functions are commonly used in applied work: *Gaussian* and *Epanechnikov*. The former uses information on all non-participants and is defined by the following expression.

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2 / 2) \tag{2.15}$$

The Epanechnikov kernel is defined as follows.

$$K(u) = \frac{3}{4}(1 - u^2) \times I(| u | \leq 1) \tag{2.16}$$

where **I(.)** is the indicator function that takes the value of 1 if its argument is true, and zero otherwise. In other terms, this kernel uses a moving window within the group of non-participants selecting only those whose propensity score is within a fixed *bandwidth* of size **h** from $\mathbf{p_i}$, i.e. those for whom $||\mathbf{p_i}\text{-}\mathbf{p_j}||<\mathbf{h}$.

Other specifications of the kernel function include the *biweight* or quartic, the *triweigh*t and the *cosinus* functions. The quartic kernel is defined as follows.

$$K(u)=\frac{15}{16}\left(1-u^2\right)^2\times I(|u|\leq 1) \qquad (2.17)$$

Similarly, the triweight kernel is given by the following expression.

$$K(u)=\frac{35}{32}\left(1-u^2\right)^3\times I(|u|\leq 1) \qquad (2.18)$$

The cosinus kernel is equal to:

$$K(u)=\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)\times I(|u|\leq 1) \qquad (2.19)$$

Based on (2.14), kernel matching can be interpreted as a locally weighted regression of the outcome on a constant, where the weights are determined by some kernel function (Smith and Todd 2005a). This regression analogy can be extended as follows. Assume that expected outcome of participant **i** in non-participation state is a linear approximation written as:

$$\beta(p_j)\approx\beta_0+(p_j-p_i)\beta_1 \qquad (2.20)$$

Kernel-Weighted OLS minimizes the following weighted sum of squares.

$$S(\beta)=\sum_{j=1}^{n}K\left(\frac{p_j-p_i}{h}\right)\left[y_j-\beta_0-(p_j-p_i)\beta_1\right]^2 \qquad (2.21)$$

Hence, the outcome participant **i** would have achieved had she not participated in the program is equal to

$$\hat{y}(p_i)=\hat{\beta}_0 \qquad (2.22)$$

Generally speaking, kernel matching entails the following basic steps (Loader 2003): (i) Choose a particular participant and get her propensity score **p_i**; (ii) Assign weights to observations on non-participants according to their distance from **p_i**; (iii) Specify a local model of the expected outcome (e.g. 2.20); (iv) Estimate the underlying parameters according to (2.21); (v) Use the estimate of the intercept as an estimate of the expected outcome for participant **i** in non-participation state; (vi) Repeat for each participant. In fact, all matching algorithms discussed here can be viewed as extensions of the idea of a *moving average*. The basic idea behind the moving average involves sliding a "*window*" across the data and taking the average of the response variable (or outcome) for all observations in the window. This is essentially a matching algorithm.

A Brief Comparison with other Evaluation Methods

As noted earlier, the outcome of interest is essentially a function of observable and unobservable characteristics of the unit under consideration, and whether or not the unit participated in the program. To assess the effect of an intervention on the outcome, we need to control for all observable and unobservable influences except participation. Failure to control for any of these factors will bias the results. Evaluation methods can therefore be characterized in terms of how they deal with these potential sources of bias.

Randomization balances selection bias between participants and nonparticipants so that subtracting the average outcome of nonparticipants from that of participants yields an unbiased estimate of average program impact. In effect, randomization ensures that participants have the same distribution of pre-intervention attributes (both observable and unobservable) as nonparticipants. In non-experimental situations, PSM attempts to create conditions similar to an experiment by assuming conditional independence and controlling observable heterogeneity through matching. In effect, PSM is a nonparametric method that assumes away unobserved heterogeneity.

There is a parametric analogue of PSM based on the switching regression model where the switching mechanism is assumed exogenous. As we show in Appendix C, the outcome equation is traditionally written as follows.

$$y_i = x_i\beta + \theta d_i + [u_{0i} + (u_{1i} - u_{0i})d_i] \qquad (2.23)$$

where $\mathbf{u_{si}}$ stands for unobserved characteristics in state **s=0, 1** (for nonparticipation and participation respectively). Note that the specification of (2.23) reflects the assumption that program effect for a unit with characteristics $\mathbf{x_i}$ and the coefficients $\boldsymbol{\beta}$ defining the relationship between observables and the outcome are invariant to participation. The invariance of the treatment effect is known as the *common effect* assumption (Ravallion 2005) or the *homogeneous treatment effects* assumption (Blundell and Costa-Dias 2000). Conditional independence implies that the expected value the last term in brackets is equal to zero. Thus the application of OLS to the above equation would yield an unbiased and consistent estimate of program impact, $\hat{\theta}$. The basic difference between this regression analysis and PSM stems from the fact that regression analysis requires a specification of the relation between the outcome, the participation indicator and observed attributes, while PSM requires no such thing.

Suppose there is longitudinal or repeated cross-section information on outcomes and their determinants, and that unobserved influences enter the outcome equation additively and separately in the form of an individual-specific fixed effect, a common macroeconomic effect (the same for all individuals), and a temporal-individual-specific effect (Blundell and Costa-Dias 2000). Assume that the individual-specific and the macroeconomic components affect participation while the temporal-individual-specific effect is independent of participation and observed characteristics. PSM would not be valid in this case, as the conditional independence assumption no longer holds. The Double Difference (DD) or Difference in Differences (DiD) method offers a way to get rid of these troublesome unobservable characteristics. A DD estimate of program impact can be obtained in two steps. First, for each participant and comparison unit, take the difference in outcome before and after the intervention, then compute the difference between the average change for participants and nonparticipants. The first difference removes the offending unobserved heterogeneity and restores conditional independence, while the second produces the impact estimate.

Note that, once conditional independence has been established through first-differencing, one can use either regression analysis or PSM to control for observed heterogeneity. In fact, it has been observed that failure to make comparisons in the region of common support can contribute significant bias in DD estimates. Based on expression (2.7) the average treatment effect on the treated over the common support is now given by the following expression which combines PSM with DD to yield matched double difference (MDD) estimates.

$$\theta_{MDD} = \sum_{i\in T} \omega_i \left[(y_{ia} - y_{ib}) - \sum_{j \in c(p_i)} w_{ij} (y_{aj} - y_{jb}) \right] \tag{2.24}$$

Combining PSM with DD is not necessary if there is no observable heterogeneity left after differencing. In general, this combination tends to reduce the bias associated with other evaluation methods (Ravallion 2003).

When participation and outcomes are jointly determined, one can specify a selection model including one participation and one outcome equation. Then one can resort to the instrumental variable approach to try to sort out that part of program impact attributable to exogenous variation in participation. This parametric approach too relies on a sort of conditional independence assumption known as *the exclusion restriction*. This requires the instrumental variable to be independent of outcomes given participation (Ravallion 2003). Essentially, instrumental variable estimation (IVE) is a two-stage procedure. First estimate the participation equation as a nonlinear binary response model using probit or logit, just as in the first stage of PSM. Then use the predicted value for this stage as an instrument for the participation indicator in the outcome equation (2.23) and run OLS to estimate program impact. A robust identification strategy is to rely both on the nonlinearity of the first-stage estimation process and on the exclusion restriction.

This two-stage procedure suggests a regression-adjusted matching estimator that tends to produce asymptotically efficient estimates (Monteiro 2004). The regression is based on the equation for outcome in the matched comparison group, $y_{0i} = x_i\beta_0 + u_{0i}$. In this case, the following restriction is analogous to the conditional independence assumption: $\hat{u}_{0i} = \left(y_{0i} - x_i \overset{\wedge}{\beta}_0 \right) \perp d_i \mid p(z_i)$, where **z_i** is the set of participation determinants, at least one which must be excluded from the outcome equation. In this case, the impact estimator is

$$\theta_{MR} = \sum_{i\in T} \omega_i \left[(y_{1i} - x_i \overset{\wedge}{\beta}_0) - \sum_{j \in c(p_i)} w_{ij} (y_{0j} - x_j \overset{\wedge}{\beta}_0) \right] \tag{2.25}$$

Where $\overset{\wedge}{\beta}_0$ is the OLS estimate of the regression coefficients in outcome equation for the comparison group[11].

Finally, one can resort to the standard Heckman's selection-correction method to cope with heterogeneity bias (see Appendix C for details). Suppose we estimate the participation equation based on the probit model. The results of this probit analysis can be used to compute the following consistent estimates of the inverse Mills ratios $\overset{\wedge}{\lambda}_{0i} = \frac{\phi(z_i \overset{\wedge}{\gamma})}{1-\Phi(z_i \overset{\wedge}{\gamma})}, \overset{\wedge}{\lambda}_{1i} = \frac{\phi(z_i \overset{\wedge}{\gamma})}{\Phi(z_i \overset{\wedge}{\gamma})}$. Maintaining the case of *homogeneous impact* described above, a consistent two-stage estimate of **θ** can be obtained by running OLS regression of **y_i** on $x_i, d_i, \sigma_{u\varepsilon}[d_i \overset{\wedge}{\lambda}_{1i} + (1-d_i)\overset{\wedge}{\lambda}_{0i}$ using all of the observations. In other terms, the estimating equation is (Lalonde 1986):

$$y_i = x_i\beta + \theta d_i + \sigma_{u\varepsilon} \overset{\wedge}{\lambda}_i + v_i \tag{2.26}$$

Essentially, this two-stage estimator treats unobservable heterogeneity as a problem of an omitted variable, and solves this problem by including an estimate of the omitted variable as a regressor in the outcome equation along with the participation dummy and individual characteristics.

This comparison reveals how each method controls for observable and non-observable determinants of outcome besides participation. Randomization ensures that, in the pre-intervention state, the distribution of these determinants is the same for both participants and nonparticipants. PSM assumes conditional independence to get rid of unobservable heterogeneity and controls for observed heterogeneity through matching on the propensity score. The conventional regression method is a parametric analogue of PSM. The DD method is a two-step procedure that relies on differencing to control for unobservable heterogeneity stemming from fixed effects, and on averaging to control for observed heterogeneity. The IV method relies on regression analysis to control for observables and uses an instrumental variable to recover conditional independence. The Heckman approach is analogous to the IV method except that it interprets unobservable heterogeneity as an omitted variable problem. Thus instead of using an instrumental

[11] This approach extents easily to the case of PSM combined with DiD. See Monteiro (2004) for details.

variable for the endogenous dummy variable in the outcome equation, it adds an estimate of the omitted variable in the equation. It turns out that PSM can act as an effective adjuvant to all these methods.

With respect to the choice among various evaluation methods, this comparison suggests situations that are best suited to each method. Hence, no single method can be considered ideal in all circumstances, and one should consider a flexible application of available methods. In the end, the plausibility of an evaluation method hinges critically on the correctness of the socioeconomic model underlying program design and implementation. It also depends on the quality and quantity of data available. The specification of the underlying socioeconomic model must be grounded on a sound understanding of political and socioeconomic determinants of participation, and all relevant factors that influence outcome besides participation.

3. Numerical Implementation

This section illustrates how some of the algorithms and estimators described above might be implemented in EViews. We first describe the data used. Next, we explain the estimation of the propensity score. We also discuss the computer code and output for nearest-neighbor and kernel matching respectively. Finally, in the true spirit of impact evaluation, we present matched estimates along results from alternative methods.

Data

Our numerical implementation is based on two data sets from Dehejia and Wahba (1999)[12]. These authors explain that the available data sets are constructed from the data underlying Lalonde (1986). The data for the treated is contained in the file NSWRE74_TREATED.TXT. This is the male sub-sample of a 185 observations from the National Supported Work (NSW) Demonstration. This was a temporary employment program seeking to help disadvantaged workers[13] acquire basic job skills through work

[12] The actual data sets were downloaded from http://www.columbia.edu/%7Erd247/nswdata.html

[13] The program targeted women receiving aid for dependent children (AFDC), ex-drug addicts, ex-criminal offenders and high school drop outs of both sexes. It was run by the Manpower Demonstration Research

experience and counseling in a protected environment. Qualified applicants were assigned to either training or control group *randomly*. Members of the treatment group received job training for 9 to 18 months depending on their profile and the site. The non-experimental data for the comparison group are contained in the file PSID_CONTROLS.TXT. This file holds relevant information on 2490 individuals drawn from the Population Survey of Income Dynamics (PSID). This sample consists of all male heads of household less than 55 years old who were continuously observed in the PSID, and not classified as retired in 1975.

Table 3.1. Descriptive Statistics for the Underlying Data

Variable	**Treated**	**Comparison**
Age	25.82 (7.16)	34.85 (10.44)
Education	10.35 (2.01)	12.12 (3.08)
Black	0.84 (0.36)	0.25 (0.43)
Hispanic	0.06 (0.24)	0.03 (0.18)
Married	0.19 (0.39)	0.87 (0.34)
No High School Degree	0.71 (0.46)	0.31 (0.46)
Real Earnings in 1974	2095.57 (4886.62)	19428.75 (13406.88)
Real Earnings in 1975	1532.06 (3219.25)	19063.34 (13596.95)
Real Earnings in 1978	6349.14 (7867.40)	21553.92 (15555.35)
Zero Earnings in 1974	0.71 (0.46)	0.09 (0.28)
Zero Earnings in 1975	0.60 (0.49)	0.10 (0.30)
Sample Size	185	2490

Source: Author's calculations

For both groups, the variables used in the analysis include the following: (1) outcome of interest, represented by real earnings in 1978 (RE78); (2) participation indicator (TREAT =1 if participant, and 0 otherwise); (3) a set of pretreatment covariates including age (AGE), education (EDU), marital status (MARRIED=1 if married and 0

Corporation on ten sites across the United States including Atlanta, Chicago, Hartford, Jersey City, Newark, New York, Oakland, Philadelphia, San Francisco, and Wisconsin (Lalonde 1986).

otherwise), indicator of black race (BLACK=1 if black and 0 otherwise), indicator of Hispanic origin (HISP=1 if Hispanic and zero otherwise), indicator of high school degree holder (NODEGREE=1 if no high school degree, 0 otherwise), real earnings in 1975 (RE75), and real earnings in 1974 (RE74).

Table 3.1 contains summary statistics (mean and standard deviation, the latter is between parentheses) of variables in the data set and for both the treated and the comparison group. Focusing on the outcome variable, these descriptive statistics reveal that pretreatment mean earning levels are much lower for the treated than for the comparison group. In 1974, for instance, average earnings stand at about 2,096 for the treated compared to 19,429 for the comparison group. The variable "Zero Earnings in 1974" is a dummy variable that indicates observations for which real earnings in 1974 is zero. There is a similar variable for 1975. The mean of these dummy variables represents the proportion of the sample with zero earnings for the year in question. About 71 percent of the treated had no earnings in 1974 compared to 9 percent for the comparison group. Finally 71 percent of the treated have no high school degree compared to 31 percent for the comparison group. These descriptive statistics suggest that the program did indeed focus on disadvantaged workers.

The estimation of the propensity score requires that data for both participants and nonparticipants reside in the same workfile page. In preparing the data prior to estimation, we use the PAGEAPPEND command to combine information for both groups. The procedure works as follows. Both the source and the destination workfiles must be open. In our case, we bring the information on the treated into the file for the nonparticipants. The file NSW_TREATED is therefore the source file and PSID_CONTROLS is the destination file. We select the destination workfile as the default then invoke the PAGEAPPEND command according to the following syntax[14]: PAGEAPPEND SOURCEFILENAME. Specifically, with PSID_CONTROLS as default we use the command: PAGEAPPEND NSW_TREATED. Before execution, EViews issues the following warning:

[14] The general syntax of this command is: PAGEAPPEND(OPTIONS) WFNAME[\PAGENAME] [OBJECT_LIST]. WFNAME stands for the name of the source file. Optionally, one may specify the name of a page within the source file and the list of objects to be brought from the source. Other options include a sample restriction specifying which observations from the source page to be appended (see EViews 5.1 User's Guide for more details). The result of the command is to append the relevant content of the source page to the active page within the destination workfile.

"*Append will add 185 observations to the workfile and remove any structure. Do you want to continue? Yes or no*?" We choose "yes", and rename the new workfile page: "COMBINED". This page now holds a total of 2675 observations. It resides in the overall wokfile called "NSWDATA".

Estimating the Propensity Score

Underlying Model

The first step in propensity score matching is to estimate the probability of participation (receiving treatment) conditional on some covariates **x**. This can be based on a model of the probability that **d=1** given **x**, **Pr{d=1|z}**. This probability is also equal to the conditional expectation of the dummy variable **d**. Let **π(z)** stand for this conditional expectation, then we can write:

$$\pi(z) = E(d \mid z) = 1 \times \Pr\{d = 1 \mid z\} + 0 \times \Pr\{d = 0 \mid z\} = \Pr\{d = 1 \mid z\} \qquad (3.1)$$

By analogy to standard regression analysis, we can write a *participation model* as the sum of the conditional expectation and a random disturbance term, **u**. In other terms,

$$d_i = \pi(z_i) + \varepsilon_i \qquad (3.2)$$

It can be shown that **E(ε$_i$)=0**, and **var(ε$_i$)=π(z$_i$)[1-π(z$_i$)]=var(d$_i$)**.

To further specify the model, we assume that **d$_i$=1** only when the underlying function, **h(z$_i$, ε$_i$)**, of observable and unobservable characteristics is greater than zero, otherwise **d$_i$=0**. Let $h(z_i, \varepsilon_i) = z_i\gamma + \varepsilon_i$, and **F()** stand for the cumulative distribution of **ε$_i$**, then we have the following expression.

$$\pi(z_i) = \Pr\{\varepsilon_i > -z_i\gamma\} = 1 - \Pr\{\varepsilon_i \le -z_i\gamma\} = [1 - F(-z_i\gamma)] \qquad (3.3)$$

It is commonly assumed that the distribution function **F()** is symmetric as in the case of the normal or the logistic distributions. Thus, $\Pr\{\varepsilon_i > -z_i\gamma\} = \Pr\{\varepsilon_i < z_i\gamma\}$. This implies that **F(-z$_i$γ)=1-F(z$_i$γ)**, and $\pi(z_i) = F(z_i\gamma)$. We maintain this assumption of symmetry throughout. This expression reveals that, the marginal effect of covariate **z_k** on the propensity score is equal to the following.

$$m_k(z) = f(z\gamma)\gamma_k \tag{3.4}$$

where **f()** stands for the density function associated with distribution **F()**.

Maximum Likelihood Estimation

The relevant parameters of this model can be estimated by the maximum likelihood method. For a sample on size **n**, the likelihood function may be written as follows.

$$L = \prod_{i=1}^{n} \left(1 - F(z_i\gamma)\right)^{(1-d_i)} \left(F(z_i\gamma)\right)^{d_i} \tag{3.5}$$

The corresponding *log likelihood function* is

$$l(\gamma) = \sum_{i=1}^{n} \left[(1-d_i)\ln\left(1 - F(z_i\gamma)\right) + d_i \ln\left(F(z_i\gamma)\right)\right] \tag{3.6}$$

The first order conditions for maximizing the log likelihood with respect to the parameters may be written as:

$$\frac{\partial l(\gamma)}{\partial \gamma_k} = \sum_{i=1}^{n} \frac{[d_i - F(z_i\gamma)]}{F(z_i\gamma)[1 - F(z_i\gamma)]} f(z_i\gamma) z_{ik} = 0 \tag{3.7}$$

The first order derivative of the log likelihood function with respect to the constant term among the explanatory variables is known as the *generalized residual* (Greene 2000). It is defined as follows.

$$e_{gi} = \frac{[d_i - F(z_i\gamma)]}{F(z_i\gamma)[1 - F(z_i\gamma)]} f(z_i\gamma) \tag{3.8}$$

Thus, the first order conditions for maximization stated in (3.7) may be regarded as an *orthogonality condition* between the generalized residuals and the explanatory variables in **z**. The condition associated with the constant term (among the explanatory variables) implies that the sum of generalized residuals is equal to zero, which in turn implies that the sample average of estimated propensity scores must equal the proportion of observations for which $\mathbf{d_i=1}$.

To proceed with estimation of the relevant parameters, we must further specify the distribution function **F()**. EViews supports the following specifications: (1) *Logit*, based on the logistic distribution; (2) *Probit*, based on the cumulative distribution of the standard normal; and (3) *Gompit*, based on the Type-I extreme value distribution. We implement the logit method. In the case of binary dependent variable models (such as the one considered here), there are two ways of implementing maximum likelihood estimation in EViews. The simplest (which we use here) is to invoke the BINARY command. The more general approach is to use the log likelihood (LOGL) object. We explain this approach in Appendix A.

Table 3.2. Output of the Binary Command

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-7.474743	2.443511	-3.059018	0.0022
AGE	0.331690	0.120330	2.756509	0.0058
AGE2	-0.006367	0.001855	-3.431530	0.0006
EDU	0.849268	0.347706	2.442490	0.0146
EDU2	-0.050620	0.017249	-2.934625	0.0033
MARRIED	-1.885542	0.299331	-6.299189	0.0000
BLACK	1.135972	0.351785	3.229161	0.0012
HISP	1.969020	0.566859	3.473560	0.0005
RE74	-0.000106	3.53E-05	-3.003993	0.0027
RE75	-0.000217	4.14E-05	-5.235083	0.0000
RE742	2.39E-09	6.43E-10	3.716073	0.0002
RE752	1.36E-10	6.65E-10	0.204285	0.8381
BLACKU74	2.144130	0.426815	5.023557	0.0000
Mean dependent var	0.069159	S.D. dependent var		0.253772
S.E. of regression	0.146679	Akaike info criterion		0.162972
Sum squared resid	57.27243	Schwarz criterion		0.191605
Log likelihood	-204.9754	Hannan-Quinn criter.		0.173332
Restr. log likelihood	-672.6495	Avg. log likelihood		-0.076626
LR statistic (12 df)	935.3484	McFadden R-squared		0.695272
Probability(LR stat)	0.000000			

Source: Author’s calculations.

The basic syntax of the binary command is EQ_NAME.BINARY(OPTIONS) Y X_1 [X_2 X_3…]. EQ_NAME is the name of an equation object. The key applicable options involve the specification of the likelihood, the choice of the maximization algorithm, the method of computation of standard errors, the setting of the maximum number of iterations, the choice of a convergence criterion, and the setting of starting values for the coefficients. The normal distribution is the default likelihood function and quadratic hill climbing is the default maximization algorithm[15].

The specific command we use on our data is: EQUATION LOGITEQ.BINARY(D=L, M=1000, C=1E-10, SHOWOPTS) TREAT ZGRP. The first option (D=L) selects the logistic model, and the second sets the maximum number of iterations to one thousand. The third option sets the convergence criterion, while the last asks EViews to show these options with the results. The dependent variable is the treatment indicator (TREAT) while the

[15] Other algorithms include Newton-Raphson and Berndt-Hall-Hall-Hausman.

explanatory variables are contained in a group object called XGRP. Following the specification of Becker and Ichino (2002), these variables are: the constant term (C), AGE, age squared (AGE2), education (EDU), education squared (EDU2), MARRIED, BLACK, HISP, real earnings in 1974 (RE74), real earnings in 1975 (RE75), square of real earnings in 1974 (RE742), square of real earnings in 1975 (RE752) and an indicator for black participants who were not employed in 1974 (BLACKU74). The estimation results are presented in table 3.2.

These results show, for instance, that Blacks and Hispanics were more likely to participate in the NSW program, while marital status has quite a negative influence on the probability of participation. Based on the estimated propensity score, we determine that the region of common support is [0.00061066, 0.97552547].

Estimates of Program Impact

Table 3.3. Matched Estimates of the Treatment Impact

	PSM	MDD	RPSML	RPSMD
Nearest Neighbor	1667.64	1262.04	1103.88	283.63
Gaussian Kernel	1537.95	1933.70	1746.25	1033.28
Epanechnikov Kernel	1370.43	1480.26	1469.80	748.71

Source: Author's calculations

Table 3.3 presents matched estimates of average treatment effect on the treated based on different versions of propensity score matching. Column PSM contains results from standard propensity score matching, column MDD is based on matched double difference, while the last two combine regression analysis and propensity score matching. They differ only in the specification of the outcome equation in the pre-treatment state. Column RPSML is based on a modified specification proposed by Lalonde (1986). The included explanatory variables are: AGE, AGE2, EDU, BLACK, HISP, RE74, and RE75. The specification underlying the last column, RPSMD, comes from Dehejia and Wahba (2002). The relevant variable are: AGE, EDU, BLACK, HISP, NODEGREE, MARRIED, RE74, RE75, U74 and U75.

To make some sense of these results, we note from table (3.1) that average real earnings in 1978 were US$ 6,346.14 for the treated. We also have data on the experimental control group. Their average real earnings stood at US$ 4,554.80 in 1978. Thus ATET based on randomization is equal to US$ 1,794.34. Using this as a metric, we first note that all non-experimental methods represented in table 3.1 show a positive program impact. The results also clearly show that impact estimates vary with both matching algorithms and adjuvant tools. Given the data available, the nearest neighbor algorithm seems to perform best in the case of simple PSM. The Gaussian kernel algorithm seems to outperform the other algorithms across variants. In the particular case of regression adjusted matching based on the modified Lalonde (1986) specification, matching using the Gaussian kernel produces an estimate that is very close to the benchmark. A straight application of the DD method yields an estimate of average impact of US$ 2, 326.50. Combining kernel matching with DD seems to improve this estimate, particularly when using the Gaussian kernel. Also, the difference between the regression-adjusted results demonstrates the sensitivity of the results to the specification of the outcome equation. It is thus important to base such specification on a sound understanding of the theory and facts driving possible outcomes.

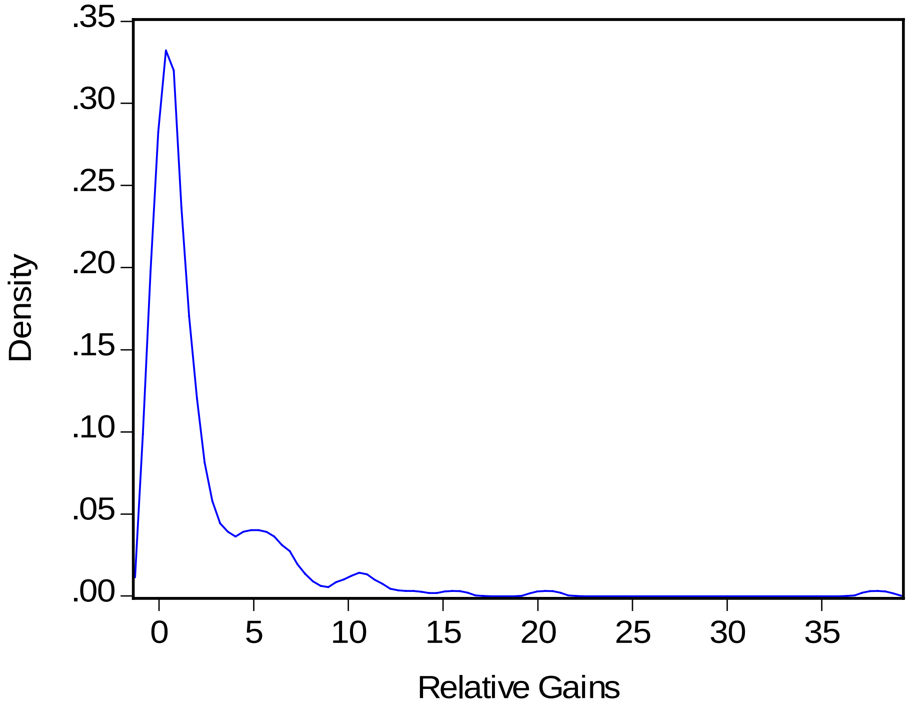
Table 3.4. Parametric Estimates of the Treatment Impact

	REGML	REGDW
OLS	217.94	4.16
IVE	881.44	1335.50
Heckman-1	989.91	796.74
Heckman-2	631.81	847.23

Source: Author's calculations

Table 3.4 presents parametric estimates of treatment impact based on regression analysis under a variety of assumptions about the correlation between participation and outcomes, and on effect invariance. Column REGML is based on the modified Lalonde (1986) specification while Dehejia and Wahba (2002) specification underlies column REGDW. The first three rows assume common effect, while the last does not. Row OLS (ordinary least squares) assumes conditional independence, while the last three do not. Row IVE (instrumental variable) corrects for unobserved heterogeneity using the propensity score as an instrument for the endogenous participation dummy. Heckman adds an estimate of the inverse Mills ratio to the relevant outcome equation to correct for unobserved heterogeneity. These parametric estimates show positive impact just like the nonparametric and semi-parametric methods of table 3.3. Given the experimental benchmark, these results show that, the nonparametric and semi-parametric methods generally outperform the parametric ones.

Figure 3.1. Distribution of Relative Impact Based on Gaussian Kernel PSM

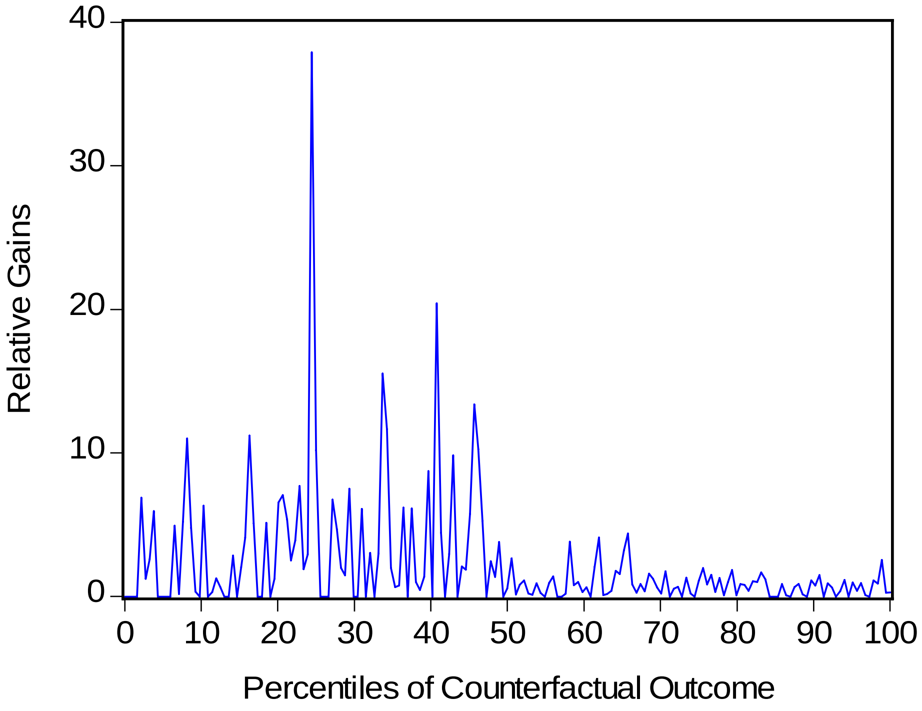


To look beyond average impact, we focus on the distribution of relative impact based on Gaussian kernel matching (the corresponding average is US$ 1,537.95 in table 3.3). This distribution is presented in figure 3.1. The figure is essentially a smoothed histogram of the sample distribution of the ratio **y_1/y_0** among the participants. The smoothing is based on the Epanechnikov kernel function using Silverman's method to determine the bandwidth. The EViews command to accomplish this is the following: SERIES_NAME.KDENSITY(K=E, S, 100, O=ARG)[16]. The data underlying figure 3.1 indicate

[16] The options have the following meanings: K=E for Epanechnikov, S for Silverman's formula $h=0.9\alpha n^{(-1/5)}$, where α=min{standard deviation, (interquartile range)/1.34}. The interquartile range is the difference

that relative impact varies from zero (for those participants who earned no income in 1978) to a maximum of 37.92, with an average of 2.27 and a standard deviation of 4.12.

Figure 3.2. Relative Program Incidence Based on Gaussian Kernel PSM



To identify who might have gained most out of the program, we look at the same distribution of impact in the form of the *relative program incidence curve* presented in figure 3.2. This curve is obtained as follows. Rank the participants in ascending order of the estimated counterfactual outcome. Compute the ratio $\mathbf{y_1/y_0}$ (relative impact), and the relative rank $\mathbf{p}$ of each participant. Finally, plot the relative impact as a function of $\mathbf{p}$. The relative incidence curve shows that, in general, the people who gain most from the program are among participants who would have been at the lower end of the counterfactual distribution. Indeed, the underlying data indicate that the mean relative impact is 3.85 for the 47 percent "poorest" participants, and only 0.87 for the 53 percent "richest".

between the 75th and 25th percentiles. The integer 100 specifies the number of points at which to evaluate the density function. Finally O=ARG specifies the matrix to contain the kernel density computation.

We now move to the implementation of the matching algorithms in EViews. The computer code for the entire program (PSCOREMATCH.PRG) is presented in Appendix B. It has three major components. The first estimates the propensity score and determines the region of common support. The second component relies on three subroutines to compute matched outcomes for the three algorithms of interest. The last component computes the average treatment effects on the treated as reported in table 3.2. We obtain the same results as those reported by Becker and Ichino (2002). These are reported in table 3.3 , column PSM.

Note that all the three subroutines we consider next are local as opposed to global. In EViews, global subroutines have the ability to create or alter global objects. Such objects stay in the workfile after the routine has run. All objects created by a local subroutine are local in the sense that they are meaningful only within the subroutine and disappear from the workfile once the routine has run. Thus, a subroutine may not use or update global objects directly from within the subroutine. However global objects corresponding to arguments may be used and updated by referring to the arguments. Such objects must be created outside the local subroutine and passed on to the subroutine as arguments. The use of local subroutines minimizes workfile clutter.

Nearest-Neighbor

Box 3.1 EViews Code for Nearest-Neighbor Matching

```
SUBROUTINE  LOCAL  NEIGHBOR(VECTOR VP, SERIES PS, SERIES Y, VECTOR
MO)
    SMPL @ALL
     !NT=@ROWS(VP)
     FOR !K=1 TO !NT
            SERIES  U{!K}= ABS(VP(!K) - PS)
            SCALAR M{!K}=@MIN(U{!K})
            SERIES   NDIJ{!K}=(U{!K}=M{!K})
            SCALAR  DNO{!K}=@SUM(NDIJ{!K})
            SERIES   YNWIJ{!K}  'To hold matched outcomes
            IF DNO{!K}   THEN
                 YNWIJ{!K}=(NDIJ{!K}/DNO{!K})*Y
            ENDIF
            MO(!K)=@SUM(YNWIJ{!K})
    NEXT
ENDSUB
```

Box 3.1 presents the code for a local subroutine called NEIGHBOR, designed to perform nearest-neighbor matching. The subroutine has four arguments referring to global objects: (1) the vector of propensity scores for the treated, (2) the series containing the propensity scores for the non-treated, (3) the series of outcomes for the non-treated, and (4) the vector to keep the matched outcomes. The core loop defining this subroutine works as follows. For each participant, define a series containing the distances between her propensity score and those of all nonparticipants. Find the minimum value. Create a dummy variable that is equal to one for each observation for which the distance is equal to the minimum and zero otherwise. Finally use this dummy variable to construct the relevant weight and apply equation (2.6) for the matched outcomes.

Kernel Matching

Box 3.2 EViews Code for Gaussian Kernel Matching

```
SUBROUTINE LOCAL GAUSS(SERIES CS, VECTOR VP, SERIES PS, SERIES Y,
VECTOR MO)
        !BW=0.06 .
        !NT=@ROWS(VP)
        SMPL @ALL IF CS
        FOR !K=1 TO !NT
                SERIES  U{!K}= ABS(VP(!K) - PS)/!BW
                SERIES KIJ{!K}=@DNORM(U{!K})
                SERIES YWIJ{!K}=(KIJ{!K}/@SUM(KIJ{!K}))*Y
                MO(!K)=@SUM(YWIJ{!K})
      NEXT
ENDSUB
```

Both boxes 3.2 and 3.3 present the code for the implementation of kernel matching. As suggested by the names, GAUSS implements matching using the Gaussian kernel while EPAN uses the Epanechnikov kernel. In addition to the four arguments of NEIGHBOR, these two subroutines have an extra argument to enforce the common support condition. Again, the fundamental logic is the same in these cases as the one underlying NEIGHBOR. Essentially, these two subroutines implement expression (2.6) where the weights are defined according to (2.14) and the kernel are defined respectively by (2.15) and (2.16). Note that in the case of the Epanechnikov kernel, there is no need to include the constant term since it cancels out in the definition of weights.

Box 3.3 EViews Code for Epanechnikov Kernel Matching

```
SUBROUTINE LOCAL EPAN(SERIES CS, VECTOR VP, SERIES PS, SERIES Y,
VECTOR MO)
     !BW=0.06
     !NT=@ROWS(VP)
     SMPL @ALL IF CS
     FOR !K=1 TO !NT
                 SERIES  U{!K}= ABS(VP(!K) - PS)/!BW
                 SERIES   ED{!K}= U{!K}<=1
                 SERIES   EKIJ{!K}=(1-U{!K}^2)*ED{!K}
                 SCALAR  DNO{!K}=@SUM(EKIJ{!K})
                 SERIES   YEWIJ{!K}
                 IF DNO{!K}   THEN
                        YEWIJ{!K}=(EKIJ{!K}/DNO{!K})*Y
                 ENDIF
                 MO(!K)=@SUM(YEWIJ{!K})
     NEXT
ENDSUB
```

4. Concluding Remarks

Effective development policymaking creates a need for reliable methods for assessing whether an intervention had (or is having) the intended effect. Such an assessment would be impossible without an estimate of what would have happened in the absence of the intervention. Evaluation methods generally rely on either a control or a comparison group to estimate this counterfactual. This paper reviews the logic of the propensity score matching method, compares matching to other methods and demonstrates numerical implementation in EViews using NSW data.

In general, individual characteristics (observable and unobservable) can confound any assessment of program impact. Failure to account for such heterogeneity will bias evaluation results. Evaluation methods can therefore be characterized in terms of how they control for these confounding effects in order to isolate the impact of the intervention. Randomization ensures that both participants and nonparticipants have the same distribution of these characteristics so that the comparison of average outcome between the two groups yields an unbiased estimate of program impact. Propensity score matching (PSM) attempts to create conditions similar to an experiment by assuming that unobservable heterogeneity plays no role in participation (conditional independence). PSM accounts for observable heterogeneity by pairing participants with nonparticipants on the basis of the conditional probability of participation, given observable characteristics. The feasibility of propensity score matching depends on the extent to which the distribution of propensity scores within the treatment group overlaps with that of the comparison group.

The specification of a matching algorithm hinges on two basic factors. The first involves the definition of a measure of proximity (in the space of propensity scores) in order to identify nonparticipants who may be considered similar enough to any given participant. The second entails a weighing function that determines the weight to be assigned to each member of a neighborhood in the computation of the counterfactual outcome. The choice among matching algorithms implies a trade-off between *bias* and *precision* in estimation.

The traditional regression analysis based on a switching model of the outcome is a parametric analogue of PSM when the switching mechanism is assumed to be exogenous. When this assumption fails and longitudinal data are available for both participants and nonparticipants, then one can use the DD method to control heterogeneity assuming that the unobservable part stems from a fixed effect. Alternatively, one can resort to IV or to Heckman's selection bias correction method to handle unobservable heterogeneity.

A comparison of all these methods suggests no one method fits all circumstances. Thus one should consider a flexible application of available evaluation methods. The numerical implementation of these methods on NSW data reveals that nonparametric methods tend to produce impact estimates that are closer to the experimental benchmark than the parametric approach. In the end, the plausibility of an evaluation method hinges critically on the correctness of the socioeconomic model underlying program design and implementation. It also depends on the quality and quantity of data available. The specification of the underlying socioeconomic model must be grounded on a sound understanding of political and socioeconomic determinants of participation, and all relevant factors that influence outcome besides participation.

Appendix A

Propensity Score Estimation with the Log Likelihood Object

The Log Likelihood Object (LOGL) is a flexible tool for estimating a broad class of statistical models by maximizing a log likelihood function with respect to parameters. This class of models includes, among others, multinomial logit, Heckman sample selection models and switching regression models. The basic structure of the object involves the description of the contribution of each observation in the sample to the log likelihood, and the selection of a method for computing the derivatives of the likelihood function with respect to the parameters. Once the object is specified, the ML command can be invoked to have EViews search for the parameter values that maximize the specified likelihood function using an iterative algorithm. In this appendix, we briefly review what is involved in both the specification of the object and the estimation of parameters. The review focuses on the program PSCORE_MLE.PRG designed to replicate the results from the BINARY command described in section 3. The entire program is presented at the end of the appendix.

Specification of the Log Likelihood Object

The specification of the log likelihood object follows the standard syntax of EViews and involves a set of declaration and assignment statements. These statements create the object and append expressions for the specification of the series that will contain the contribution of each parameter. Specification also entails the determination of the names for the parameters, the choice of the order of evaluation of the expressions (by observation, the default option, or by equation), and the method for computing the derivatives of the likelihood function with respect to the parameters.

The following command declares the log likelihood object.

```
LOGL  PSLGT
```

Once the object has been created, we use the APPEND command to include the necessary assignment statements in the object. Each likelihood specification must have a *control statement* providing the name of the series which will contain the likelihood contributions. The following statement accomplishes this.

```
PSLGT.APPEND @LOG LKLHD
```

Because of the large number of explanatory variables, we use the following loop to define the index $z_i\gamma$. Note that we also use the name GAMMA for the vector of coefficient instead of the default name C. The series index must be initialized outside the loop.

```
PSLGT.APPEND INDEX=0
FOR !J=1 TO !K
        PSLGT.APPEND  INDEX{!J} GMMA(!J)*ZGRP(!J)
        PSLGT.APPEND  INDEX=INDEX + INDEX{!J}
        PSLGT.APPEND  @TEMP INDEX{!J}
NEXT
```

The @TEMP statement causes EViews to delete from the workfile any series in the list once the specification has been evaluated. The next two statements define respectively the generalized residuals to be used in specifying analytic derivatives and the series containing the likelihood contributions. The expression for the contributions is based on equation (3.6).

```
PSLGT.APPEND GRES = (TREAT - @CLOGISTIC(INDEX))
PSLGT.APPEND LKLHD=TREAT*LOG(@CLOGISTIC(INDEX)) +(1-TREAT)*LOG(1 -
@CLOGISTIC(INDEX))
```

By default, EViews automatically computes numeric derivatives of the likelihood function. One has the option of specifying analytic expressions for some or all the relevant derivatives, using the @DERIV statement. The CHECKDERIV command allows one to check the validity of such expressions by comparing them with the numerically computed ones. The following loop specifies the analytic derivatives.

```
FOR !T=1 TO !K
        PSLGT.APPEND @DERIV GAMMA(!T) GRAD{!T}
        PSLGT.APPEND GRAD{!T}=GRES*ZGRP(!T)
NEXT
```

EViews always evaluates from top to bottom when executing the assignment statements in a LOGL object. Therefore, expressions which are used in subsequent calculations must be placed first. By default, EViews evaluates the specification by observation so that all of the assignment statements are evaluated for the first observation, then for the second, and so on across all the observations in the estimation sample. This is the correct order of evaluation for recursive models where the likelihood of an observation depends on previously observed values.

It is possible to change this default setting so that the specification is evaluated by equation. The first assignment statement is evaluated for all the observations, then the second, and so on for each of the assignment in the specification. To select a particular method of evaluation one can use either “@BYOBS” or “@BYEQN” after the first control statement creating the series that will contain individual contributions. Evaluation by equation is the correct order of evaluation for models where aggregate statistics from intermediate series are used as input to subsequent calculations.

Estimation

The choice of starting values is very important here because EViews uses an iterative procedure to find the maximum likelihood estimates of the parameters. If one has reasonable starting values, then they should be entered using the @PARAM command as illustrated below.

```
@PARAM BETA(1) 0.1 GAMMA(2) 0.1 GAMMA(3) 0.1 GAMMA(4) 1 ….
```

By default, Eviews uses values found in the coefficient vector prior to estimation. Thus, another way of proceeding is to use OLS (or some other estimation method) estimates as starting values. First estimate a linear specification of the model using the LS command, then invoke the ML command. This is the procedure we follow here.

```
EQUATION SVALEQ.LS TREAT ZGRP
COEF  BETA=SVALEQ.@COEFS
```

EViews uses the sample of observations specified prior to estimation. If there are missing values, the estimation procedure will stop after an error message has been issued.

Estimation is carried out when the ML command is issued as follows:

```
PSLGT.ML(SHOWOPTS, B, M=1000, C=1E-5)
```

The option SHOWOPTS causes EViews to show the prevailing options with the output. Option B selects the Berndt-Hall-Hall-Hausman maximization algorithm (the default is Marquardt). Option M sets the maximum number of iterations while C sets the convergence criterion.

We now present the entire program and the resulting coefficient estimates.

```
'PSCORE_MLE.PRG  illustrates how to use the LOGL object in the estimation of
the propensity score.

'B. Essama-Nssah, PRMPR, The World Bank January 02, 2006
‘Revised March 02, 2006
MODE QUIET
DB PSMLE

'---------------------------------------------------------------------------------------------------------
'OPEN WORKFILE AND GET STARTING VALUES FROM OLS
WFOPEN NSWDATA 'Data from the National Supported Work Demonstration
GROUP ZGRP C AGE AGE2 EDU EDU2 MARRIED BLACK HISP RE74 RE75
RE742 RE752 BLACKU74
!K=XGRP.@COUNT 'Number of explanatory variables contained in group ZGRP
also equals number of coefficients: !K=@NCOEF
```

```
EQUATION SVALEQ.LS TREAT  ZGRP   'SVAL for Starting Values
COEF GAMMA =SVALEQ.@COEFS
'-----------------------------------------------------------------------------------------------------------
' SPECIFY LOG LIKELIHOOD OBJECT
LOGL   PSLGT   'Propensity Score based on Logit model
PSLGT.APPEND @LOGL LKLHD  'Series to hold contribution of each
observation to the log likelihood
'Initialize and create index function to use in computing generalized residuals
PSLGT.APPEND  INDEX=0
FOR !J=1 TO !K
      PSLGT.APPEND  INDEX{!J}=GAMMA(!J)*ZGRP(!J)
      PSLGT.APPEND  INDEX=INDEX+INDEX{!J}
      PSLGT.APPEND   @TEMP INDEX{!J}          'These components will be
deleted from the workfile
NEXT
PSLGT.APPEND GRES = (TREAT - @CLOGISTIC(INDEX))   'Generalized
residuals
'PSLGT.APPEND LKLHD = TREAT*LOG(@CLOGISTIC(INDEX))+(1-
TREAT)*LOG(1-@CLOGISTIC(INDEX))
'Try @recode command
PSLGT.APPEND LOGLK0= -LOG(1+EXP(INDEX))
PSLGT.APPEND LOGLK1= INDEX - LOG(1+EXP(INDEX) )
PSLGT.APPEND LKLHD=@RECODE(TREAT=0,LOGLK0, LOGLK1)
' Specify analytic derivatives
FOR !T=1 TO !K
       PSLGT.APPEND @DERIV GAMMA(!t) GRAD{!t}
       PSLGT.APPEND GRAD{!t} = GRES*ZGRP(!t)
NEXT

'-----------------------------------------------------------------------------------------------------------
' PERFORM MLE AND STORE RESULTS IN TABLES
PSLGT.ML(showopts, b, m=1000, c=1e-5) 'b=Berndt-Hall-Hall-Hausman
maximization algorithm (the default, Marquardt, did not work)
FREEZE(TABGRAD) PSLGT.CHECKDERIV
FREEZE(TABOUT)   PSLGT.OUTPUT
FREEZE(TABCOV) PSLGT.COEFCOV

'-----------------------------------------------------------------------------------------------------------
'COMPUTE PROPENSITY SCORE AND COMMON SUPPORT

SERIES PSHAT
SERIES COMSUP 'Region of common support
MODEL PSCORE
PSCORE.APPEND PSHAT=@CLOGISTIC(INDEX)
'PSCORE.APPEND PSHAT=1-@CLOGISTIC(-INDEX)'Alternative formulation
```

```
PSCORE.SCENARIO ACTUALS
PSCORE.SOLVE
SMPL @ALL IF TREAT=1 'Restrict the sample to the treated
          SCALAR MINPS=@MIN(PSHAT)
          SCALAR MAXPS= @MAX(PSHAT)
SMPL @ALL

COMSUP=(PSHAT>=MINPS AND PSHAT<=MAXPS)

'-------------------------------------------------------------------------------------------------------
'PREPARE REPORT
STORE INDEX  MINPS MAXPS PSCORE  PSLGT SVALEQ TABCOV
TABGRAD TABOUT
PAGECREATE(PAGE=REPORT) U 1
FETCH MINPS MAXPS PSCORE  PSLGT SVALEQ TABCOV TABGRAD
TABOUT

'----------------------------
'END OF PROGRAM
```

Table A1. Coefficient Estimates from PSCORE_MLE.PRG

	Coefficient	Std. Error	z-Statistic	Prob.
GAMMA(1)	-7.474766	3.035352	-2.462570	0.0138
GAMMA(2)	0.331690	0.156229	2.123102	0.0337
GAMMA(3)	-0.006367	0.002461	-2.586788	0.0097
GAMMA(4)	0.849266	0.427266	1.987673	0.0468
GAMMA(5)	-0.050620	0.021977	-2.303303	0.0213
GAMMA(6)	-1.885528	0.320700	-5.879406	0.0000
GAMMA(7)	1.135981	0.415008	2.737249	0.0062
GAMMA(8)	1.969024	0.627639	3.137191	0.0017
GAMMA(9)	-0.000106	5.15E-05	-2.057141	0.0397
GAMMA(10)	-0.000217	7.00E-05	-3.096242	0.0020
GAMMA(11)	2.39E-09	1.46E-09	1.636103	0.1018
GAMMA(12)	1.36E-10	2.91E-09	0.046684	0.9628
GAMMA(13)	2.144122	0.433813	4.942498	0.0000
Log likelihood	-204.9754	Akaike info criterion		0.162972
Avg. log likelihood	-0.076626	Schwarz criterion		0.191605
Number of Coefs.	13	Hannan-Quinn criter.		0.173332

Source: Author's calculations

Appendix B

Computer Code for Propensity Score Matching

'PSCOREMATCH.PRG illustrates the implementation of Propensity Score Matching (PSM) in EViews 5.1 using data, from Rajeev Dehejia's website, relating to the National Supported Work (NSW) Demonstration. The comparison group is a subsample from the Panel Study of Income Dynamics (PSID). See Becker Sascha O. and Ichino Andrea (2002) Estimation of the Average Treatment Effects Based on Propensity Scores, The Stata Journal, Vol.2, No.4: 358-377. The matching methods implemented here are also described and implemented in STATA by Becker and Ichino (2002). We use their results for validation.

'B. Essama-Nssah, PRMPR, The World Bank Group, January 04, 2006
‘Revised March 02, 2006
'***Subroutines must be placed at the very beginning of the program
'***Use local subroutines to prevent keeping nonessential intermediate results in the workfile

```
'-----------------------------------------------------------------------------------------------------
'DEFINE SUBROUTINES FOR THE PRODUCTION OF MATCHED OUTCOMES
'***Arguments:(1) CS=Common Support, (2) VP=Vector of Propensity Scores for
the Treated, (3) PS=Propensity Scores for the Nontreated, (4) Y=Outcomes for
the Nontreated, (5) MO=Matched Outcomes

'***Nearest-Neighbor
SUBROUTINE  LOCAL  NEIGHBOR(VECTOR VP, SERIES PS, SERIES Y,
VECTOR MO) 'No common support imposed
        SMPL @ALL
         !NT=@ROWS(VP) 'Total number of treated
         FOR !K=1 TO !NT
                SERIES  U{!K}= ABS(VP(!K) - PS)
                SCALAR M{!K}=@MIN(U{!K})
                SERIES   NDIJ{!K}=(U{!K}=M{!K})  'Indicator of a match
                SCALAR  DNO{!K}=@SUM(NDIJ{!K}) 'Number of matches
                SERIES   YNWIJ{!K}  'To hold matched outcomes
                IF DNO{!K}   THEN
                        YNWIJ{!K}=(NDIJ{!K}/DNO{!K})*Y
                ENDIF
                MO(!K)=@SUM(YNWIJ{!K})  'Matched outcomes for participant !k
      NEXT
ENDSUB
```

```
'***Gaussian Kernel
SUBROUTINE LOCAL GAUSS(SERIES CS, VECTOR VP, SERIES PS, SERIES
Y, VECTOR MO)
        !BW=0.06  'Bandwidth, same as the one used by Becker and Ichino (2002)
on the same data set.
        !NT=@ROWS(VP)
        SMPL @ALL IF CS  'Impose common support
        FOR !K=1 TO !NT
                SERIES  U{!K}= ABS(VP(!K) - PS)/!BW
                SERIES KIJ{!K}=@DNORM(U{!K})  'Gaussian kernel
                SERIES YWIJ{!K}=(KIJ{!K}/@SUM(KIJ{!K}))*Y
                MO(!K)=@SUM(YWIJ{!K}) 'Matched outcomes
      NEXT
ENDSUB
'***Epanechnikov Kernel
SUBROUTINE LOCAL EPAN(SERIES CS, VECTOR VP, SERIES PS, SERIES
Y, VECTOR MO)
       !BW=0.06  'Bandwidth
       !NT=@ROWS(VP)
       SMPL @ALL IF CS   'Common support
       FOR !K=1 TO !NT
                SERIES  U{!K}= ABS(VP(!K) - PS)/!BW
                SERIES   ED{!K}= U{!K}<=1  'Indicator function for the
Epanechnikov kernel
                SERIES   EKIJ{!K}=(1-U{!K}^2)*ED{!K}  'Epanechnikov kernel (no
need to include the constant factor 3/4)
                SCALAR  DNO{!K}=@SUM(EKIJ{!K})
                SERIES   YEWIJ{!K}
                IF DNO{!K}   THEN
                        YEWIJ{!K}=(EKIJ{!K}/DNO{!K})*Y
                ENDIF
                MO(!K)=@SUM(YEWIJ{!K})  'Matched outcomes
       NEXT
ENDSUB
'----------------------------------------------------------------------------------------------------
'START PROGRAM EXECUTION
MODE QUIET
DB PSM
WFOPEN NSWDATA  'Data from the National Supported Work Demonstration
GROUP ZGRP C AGE AGE2 EDU EDU2 MARRIED BLACK HISP RE74 RE75
RE742 RE752 BLACKU74
'----------------------------------------------------------------------------------------------------
'COMPUTE PROPENSITY SCORES
'Invoke the Binary and Fit Commands
```

```
EQUATION LOGITEQ.BINARY(d=L, m=1000, c=1e-10, showopts) TREAT
ZGRP 'ZGRP contains pre-treatment attributes
FREEZE(LOGITAB) LOGITEQ.OUTPUT
LOGITEQ.FIT PSHAT
STORE LOGITEQ LOGITAB
'------------------------------------------------------------------------------------------------------------
'DETERMINE REGION OF COMMON SUPPORT BASED ON THE SCORES OF
THE TREATED
SMPL @ALL IF TREAT=1 'Restrict the sample to the treated
          SCALAR MINPS=@MIN(PSHAT)
          SCALAR MAXPS= @MAX(PSHAT)
SMPL @ALL
SERIES COMSUP=(PSHAT>=MINPS AND PSHAT<=MAXPS)
'------------------------------------------------------------------------------------------------------------
'SEPARATE PARTICIPANTS FROM NONPARTICIPANTS
PAGECOPY(PAGE=TREATED, SMPL=@ALL IF TREAT=1) 'This corresponds
to observations in NSWRE74_TREATED
STOM(PSHAT,VPSP) 'VPSP is a vector of propensity scores for participants
STORE VPSP
PAGESELECT  COMBINED
PAGECOPY(PAGE=COMPARISON, SMPL=@ALL IF TREAT=0)
FETCH VPSP
!T=@ROWS(VPSP)   'Number of treated equals the size of the vector VPSP
'------------------------------------------------------------------------------------------------------------
'CREATE VECTORS YCN YCG AND YCE  WITH SIZE EQUAL TO NUMBER
OF TREATED
'*** The vectors will hold matched outcomes from nearest-neighbor, Gaussian
and Epanechnikov matching methods.
FOR %V YCN YCG YCE
          VECTOR(!T)  {%V}
NEXT
'------------------------------------------------------------------------------------------------------------
'COMPUTE NEAREST-NEIGHBOR-MATCHED OUTCOMES
CALL NEIGHBOR( VPSP, PSHAT, RE78, YCN)
'------------------------------------------------------------------------------------------------------------
'COMPUTE GAUSSIAN KERNEL-MATCHED OUTCOMES
CALL GAUSS(COMSUP, VPSP, PSHAT, RE78, YCG)
'------------------------------------------------------------------------------------------------------------
'COMPUTE EPANECHNIKOV KERNEL-MATCHED OUTCOMES
CALL EPAN(COMSUP, VPSP, PSHAT, RE78, YCE)
'------------------------------------------------------------------------------------------------------------
'STORE RESULTS TO BE USED LATER
STORE YCN YCG YCE
'------------------------------------------------------------------------------------------------------------
'COMPUTE  TREATMENT EFFECTS
PAGESELECT TREATED
```

```
FOR  %V  %S  %D YCN YNNB NNB YCG YGAUSS GSS YCE YNIKOV EPN
          FETCH {%V}
          MTOS({%V}, {%S})  'Convert vector into series
          SERIES {%D}=(RE78-{%S}) 'Individual gain=(outcome after treatment
minus estimated counterfactual)
NEXT
VECTOR(3) ATET
!T=1
FOR %D NNB GSS EPN
          ATET(!T)=@MEAN({%D})
          !T=!T+1
NEXT
FREEZE(IMPACT) ATET
SETLINE(IMPACT, 3)
!ROW=4
FOR %RLB NNBR GSSK EPNK
         SETCELL(IMPACT, !ROW, 1, %RLB, "L")
         !ROW=!ROW + 1
NEXT
'----------------------------------------------------------------------------------------------------------
'PLOT RELATIVE PROGRAM INCIDENCE CURVE BASED ON GAUSSIAN
ESTIMATES
SORT YGAUSS
SERIES RGAUSS=RE78/YGAUSS
SERIES PH=100*(@TREND/(@OBSRANGE -1))
GROUP PICG PH  RGAUSS    'PICG=Program Incidence Curve Based on
Gauss
FREEZE(RPICG) PICG.XY       'RPICG= Relative Program Incidence Curve
Based on Gauss
RPICG.LEGEND -DISPLAY
RPICG.ADDTEXT(L) Relative Gains
RPICG.ADDTEXT(B) Percentiles of Counterfactual Outcome
'Smooth the Histogram of Relative Impact
FREEZE(GKDENSE) RGAUSS.KDENSITY(K=E, S, 100, O=KRGSS)
'GKDENSE kdensity of RGAUSS, KRGSS results for kdens.
GKDENSE.LEGEND -DISPLAY
GKDENSE.ADDTEXT(L)  Density
GKDENSE.ADDTEXT(B) Relative Gains
STORE ATET GKDENSE  IMPACT  KRGSS  RPICG
'----------------------------------------------------------------------------------------------------------
'SAVE RESULTS AND CLOSE ORIGINAL DATA FILE
PAGESAVE NSWRESULTS
CLOSE NSWDATA
WFOPEN NSWRESULTS
'-----------------------------
'END OF PROGRAM
```

Appendix C
Coping with Unobservable Heterogeneity

The validity of using a comparison group in the estimation of the counterfactual outcome hinges crucially on the extent to which both participants and nonparticipants have similar characteristics prior to the intervention. There are two major sources of bias stemming from heterogeneity in either observable or non-observable characteristics. As noted by Ravallion (2005), PSM tries to recreate an observational analogue of a social experiment whereby each person has the same probability of participation. The fact that the propensity score is based only on observable characteristics clearly shows that the method assumes away unobservable heterogeneity that might arise from purposive placement into a program. This is in fact what the assumption of conditional independence entails. The success of this method in reducing overall bias in impact estimates depends on whether or not the bias due to observables moves in the same direction as that due to unobservable characteristics. In this appendix, we review a couple of ways of dealing with situations where this assumption may not hold. If one has data on outcomes and their determinants for both participants and nonparticipants before and after the intervention and believes that unobservable heterogeneity is time-invariant, then one can apply the *double difference method*, also known as "*difference in differences*". The second approach is to frame the issue within an endogenous switching model and apply appropriate econometric techniques such as *instrumental variable*, *Heckman's two-stage* or *maximum likelihood* estimation.

Double Difference

This approach compares outcome changes over time for the participants with those for the nonparticipants. The changes are computed over time relative to a pre-intervention baseline. To see clearly what is involved, consider the following general expression for the outcome. $y_{it} = y_{1it}d_i + y_{0it}(1-d_i) = y_{0it} + (y_{1it} - y_{0it})d_i$

where $y_{sit} = \beta_s(x_i) + u_{sit}$; $s = 0, 1$ for the comparison group and participants respectively. The subscript **t** stands for time. The above outcome equation can now be written as:

$$y_{it} = \beta_{0t}(x_i) + [\beta_{1t}(x_i) - \beta_{0t}(x_i)]d_i + u_{it} \qquad \text{(C1)}$$

where $u_{it} = (\eta_i + \nu_t + \xi_{it})$. Thus the random disturbance has three components, an individual-specific fixed effect, a common macroeconomic effect (the same for all individuals) and a temporary individual-specific effect (Blundell and Costa-Dias 2000). The key assumption underlying the DD method states that participation is independent only of the temporary individual-specific effect so that $E[u_{it} \mid x_i, d_i] = E[\eta_i \mid x_i, d_i] + \nu_t$. The assumption that unobserved heterogeneity is separable and time invariant implies that it can be controlled for by taking differences in outcomes over time. We can write the expected value of these differences as follows.

$$E[\Delta_t y \mid x_i, d_i] = [\beta_{0a}(x_i) - \beta_{0b}(x_i)] + \{[\beta_{1a}(x_i) - \beta_{1b}(x_i)] - [\beta_{0a}(x_i) - \beta_{0b}(x_i)]\}d_i \qquad \text{(C2)}$$

In the above expression, **t=a, b** represents after and before. The coefficient of **d$_i$** measures program impact and is equal to $\theta(x_i) = \left[E(\Delta_t y \mid x_i, d_i = 1) - E(\Delta_t y \mid x_i, d_i = 0)\right]$. This is the basic idea behind the DD method.

Practically, the double difference method involves the following basic steps. Given relevant data, for each participant and comparison unit, first calculate the difference between the values of the outcome indicator after and before the intervention and take the average within each group. Then compute the difference between these two averages to get an estimate of program impact. Averaging may be thought of as a way on controlling for observed heterogeneity. However, it has been observed that failure to make comparisons in a region of common support can contribute significant bias in the DD estimates. Thus one may use PSM prior to double differencing in order to ensure strong similarity between the comparison and the treated groups. This procedure leads to the matched DD estimator presented in the text (equation 2.24).

The data base underlying the double difference approach also allows one to estimate program impact using regression analysis to control for changes in observable characteristics over time. For each observation in the two surveys (baseline and post intervention), assuming homogenous treatment effects and linear outcome equations, the change in outcome over time takes the following form:

$$\Delta_t y = (x_{ia} - x_{ib})\beta + \theta d_i + (\xi_{ia} - \xi_{ib}) \qquad \text{(C3)}$$

Now that the "troublesome" disturbances (η_i and ν_t) have disappeared through differencing, we can safely apply ordinary least squares estimation to (C3) and obtain an unbiased and consistent estimate of program impact. Blundell and Costa-Dias (2000) note that, in the particular case of a job training program, it may be that enrolment in the program is more likely if a temporary dip in earnings occurs right before the program starts (Ashenfelter's dip). One would expect a faster growth in earning even without participation in the program. Thus the DD method may overestimate the impact of treatment. The method will also break down if the treated and the comparison groups react differently to the common macroeconomic shock.

The above formulation suggests that all observable characteristics that remain invariant over time do not contribute in explaining changes in outcomes. Yet there are situations where changes in outcome over time are determined by initial conditions. For instance, in a program designed to improve schooling among targeted groups, it is reasonable to think that parental education and area of residence affect gains is schooling by children (Ravallion 2001). To handle these situations, the above equation may be recast in the following form:

$$\Delta_t y = \beta_a x_{ia} + \beta_b x_{ib} + \theta d_i + (\xi_{ia} - \xi_{ib}) \qquad \text{(C4)}$$

Now observable characteristics can still affect change in outcomes over time even if those characteristics do not change themselves.

The implementation of the double difference method assumes that the follow up survey covers the same individuals or households. This can be difficult in practice as some units in the baseline survey may drop out for some reason (e.g. they may have moved to an unknown address, or they just no longer wish to be involved). If this attrition happens randomly, then the follow-up survey may still be representative of the baseline population. If not, then attrition bias will corrupt the DD method (Baker 2000).

Selection Bias Correction Methods

The Model

When the exogeneity assumption underlying both the PSM and the DD methods fails, one can resort to methods of estimating endogenous switching regression models. The outcome equation can now take the following general form:

$$y_i = \beta_0(x_i) + [\beta_1(x_i) - \beta_0(x_i)]d_i + u_{0i} + (u_{1i} - u_{0i})d_i \qquad \text{(C5)}$$

This equation, formulated as a regime switching model, clearly reflects the fact that outcome is a function of participation, observed and unobserved characteristics.

Instrumental Variable Estimation (IVE)

The estimation of program impact depends on structural assumptions made about response heterogeneity (Heckman 2001). If it is believed that participants and non-participants differ only in observed characteristics, then $\mathbf{u_{1i}=u_{0i}}$, and program impact is measured by: $g(x_i) = [\beta_1(x_i) - \beta_0(x_i)]$. If it is further assumed that the effect is constant across individuals so that program impact reduces to $g(x_i) = [\beta_1(x_i) - \beta_0(x_i)] = \theta$. This assumption implies that $\boldsymbol{\beta_0(x_i)}$ and $\boldsymbol{\beta_1(x_i)}$ are parallel curves differing only in the level Blundell and Costa-Dias (2002) refers to his case as homogenous effect. The outcome equation now becomes.

$$y_i = \beta_0(x_i) + \theta d_i + [u_{0i} + (u_{1i} - u_{0i})d_i] \qquad \text{(C6)}$$

The last term captures the influence of differences in unobservables that affect the outcome of participants relative to non-participants. When the expected value of this term is equal to zero, conditional independence obtains and the PSM method of impact estimation remains valid. Alternatively, one can apply OLS to the *exogenous switching regression* based on (C6) and get an unbiased estimate of average impact via the coefficient of $\mathbf{d_i}$. If the expected value of the disturbance term in (C6) is not zero, *participation and outcomes are jointly determined*. One can resort to the instrumental variable approach to try to sort out that part of program impact attributable to exogenous variation in participation. This requires a separate model of participation including, among other explanatory variables, one variable reflecting some observable exogenous variation in program participation. Such a variable must be correlated with participation but not with the error term in the outcome model. The instrumental variable must not be included in the outcome equation. Heckman and Smith (1995) note that "*randomization acts as an instrumental variable by creating variation in the receipt of treatment among participants*". In many non-experimental situations one can turn to geography, politics or discontinuities created by program design in search of instrumental variables (Ravallion 2005). The particular procedure described in the text is analogous to Two-Stage Least Squares.

Heckman's Two-Stage Procedure

Endogenous switching models offer a more general framework for coping with unobservable heterogeneity. We can reformulate the model in equation (C5) as follows (Maddala 1983). Let $\mathbf{y_{1i}}$ be the outcome if unit **i** participates in the program ($\mathbf{d_i=1}$), and $\mathbf{y_{0i}}$ the outcome associated with nonparticipation ($\mathbf{d_i=0}$). For the participants, we write:

$$y_{1i} = x_i\beta_1 + u_{1i} \qquad \text{(C7)}$$

For the nonparticipants,

$$y_{0i} = x_i\beta_0 + u_{0i} \qquad \text{(C8)}$$

Furthermore, suppose that we observe $\mathbf{d_i=1}$ when a latent index **h** is strictly greater than zero. The latent index is defined by the following equation.

$$h_i = z_i\gamma + \varepsilon_i \qquad \text{(C9)}$$

For a participant with characteristics $\mathbf{x_i}$ and $\mathbf{z_i}$, the counterfactual outcome is equal to the conditional expectation $\mathbf{E(y_{0i}|d_i=1)}$. This expectation is equal to.

$$E(y_{0i} \mid d_i = 1) = E(x_i\beta_0 + u_{0i} \mid \varepsilon_i > -z_i\gamma) = x_i\beta_0 + E(u_{0i} \mid \varepsilon_i > -z_i\gamma) \qquad \text{(C10)}$$

Assume that $\boldsymbol{\varepsilon}$ and $\mathbf{u_0}$ follow a bivariate normal distribution with marginal means equal to zero and covariance $\boldsymbol{\sigma_{0\varepsilon}}$. Let the corresponding standard deviations be $\boldsymbol{\sigma_0}$ and $\boldsymbol{\sigma_\varepsilon=1}$. Then, the conditional expectation of $\mathbf{u_0}$ given ε (or the regression of $\mathbf{u_0}$ on ε) can be simply written as $E(u_{0i} \mid \varepsilon_i) = \sigma_{0\varepsilon}\varepsilon_i$. All we have to do now is to compute the expected value of this random variable in the relevant region. We first note the general idea that if **x** is a continuous random variable, then the density of **x** truncated at **c** such that **x>c** is equal to: $f(x \mid x > c) = \dfrac{f(x)}{\Pr\{x > c\}}$; $f(x \mid x \le c) = 0$. In our particular case, we have the following: $f(\varepsilon_i \mid \varepsilon_i > -z_i\gamma) = \dfrac{\phi(\varepsilon_i)}{1-\Phi(-z_i\gamma)} = \dfrac{\phi(\varepsilon_i)}{\Phi(z_i\gamma)}$ where **φ()** and **Φ()** stand respectively for the density and the cumulative distribution functions of the standard normal variate. The conditional expectation of **u₀** given participation can be written as follows: $E(u_{0i} \mid \varepsilon_i > -z_i\gamma) = E(\sigma_{0\varepsilon}\varepsilon_i \mid \varepsilon_i > -z_i\gamma) = \dfrac{\sigma_{0\varepsilon}}{\Phi(z_i\gamma)}\int_{-z_i\gamma}^{\infty} \varepsilon_i \phi(\varepsilon_i) d\varepsilon_i = \sigma_{0\varepsilon}\dfrac{\phi(z_i\gamma)}{\Phi(z_i\gamma)}$. The final result stems from the fact that $\dfrac{d\phi(x)}{dx} = -x\phi(x)$ (Greene 2000).

The above considerations imply that the counterfactual outcome is equal to

$$E(y_{0i} \mid d_i = 1) = x_i\beta_0 + \sigma_{0\varepsilon}\frac{\phi(z_i\gamma)}{\Phi(z_i\gamma)} \tag{C11}$$

Similarly,

$$E(y_{1i} \mid d_i = 1) = x_i\beta_1 + \sigma_{1\varepsilon}\frac{\phi(z_i\gamma)}{\Phi(z_i\gamma)} \tag{C12}$$

and,

$$E(y_{0i} \mid d_i = 0) = x_i\beta_0 - \sigma_{0\varepsilon}\frac{\phi(z_i\gamma)}{1-\Phi(z_i\gamma)} \tag{C13}$$

The expected impact for participant **i** can now be calculated as follows:

$$E(g_i) = E(y_{1i} \mid d_i = 1) - E(y_{0i} \mid d_i = 1) = x_i(\beta_1 - \beta_0) + (\sigma_{1\varepsilon} - \sigma_{0\varepsilon})\frac{\phi(z_i\gamma)}{\Phi(z_i\gamma)} \tag{C14}$$

The terms $\lambda_{0i} = \dfrac{\phi(z_i\gamma)}{1-\Phi(z_i\gamma)}$, $\lambda_{1i} = \dfrac{\phi(z_i\gamma)}{\Phi(z_i\gamma)}$ are known as inverse Mills ratios (or hazard rates in reliability theory, Heckman 1976). As we will see shortly, these ratios play a key role in a two stage procedure designed to find consistent estimates of the underlying structural parameters.

Expression (C14) clearly shows the way the expected impact depends on both observable characteristics through $x_i(\beta_1 - \beta_0)$, and unobservable heterogeneity through the term $(\sigma_{1\varepsilon} - \sigma_{0\varepsilon})\dfrac{\phi(z_i\gamma)}{\Phi(z_i\gamma)}$. This latter term would vanish under randomization, for outcomes would be independent of participation.

There is a two-stage estimation method that one can use to obtain consistent estimates of the structural parameters that enter the computation of impact (Maddala 1983). The first stage involves probit analysis to obtain a consistent estimate of ($\gamma/\sigma_\varepsilon$). Thus the coefficients of the determinants of participation are estimable only up to a factor of proportionality. This is what justifies the normalization of the variance of ε to one.

The results of the probit analysis lead to the following consistent estimates of the inverse Mills ratios $\hat{\lambda}_{0i} = \dfrac{\phi(z_i\hat{\gamma})}{1-\Phi(z_i\hat{\gamma})}, \hat{\lambda}_{1i} = \dfrac{\phi(z_i\hat{\gamma})}{\Phi(z_i\hat{\gamma})}$. These estimates can then be used in the following two regression equations.

$$y_{1i} = x_i\beta_1 + \sigma_{1\varepsilon}\hat{\lambda}_{1i} + v_{1i}, \forall d_i = 1 \qquad \text{(C15)}$$

and

$$y_{0i} = x_i\beta_0 - \sigma_{0\varepsilon}\hat{\lambda}_{0i} + v_{0i}, \forall d_i = 0 \qquad \text{(C16)}$$

An application of **OLS** to the above equations produces consistent estimates of $\boldsymbol{\beta_1}$, $\boldsymbol{\sigma_{1\varepsilon}}$, $\boldsymbol{\beta_0}$, $\boldsymbol{\sigma_{0\varepsilon}}$. This approach is consistent with Heckman (1976) interpretation of selection models within the framework of an *omitted variable problem*. The proposed solution to this problem is to include an estimate of the omitted variable as a regressor in the outcome equation. We can thus compute an estimate of program impact as follows.

$$\hat{g}_i = x_i\left(\hat{\beta}_1 - \hat{\beta}_0\right) + (\hat{\sigma}_{1\varepsilon} - \hat{\sigma}_{0\varepsilon})\hat{\lambda}_{1i} \qquad \text{(C17)}$$

If one is not willing to assume normality for the random error in the participation or selection equation, it is possible to use the logit model at the first stage instead of the probit model. The following transformation can then be used to estimate the inverse Mills ratios (Lee 1983). Let $\mathbf{q_i}$ stand for the quantiles associated with the predicted probabilities from the first stage estimation. Define these through the inverse cumulative standard normal distribution as $q_i = \Phi^{-1}(\hat{p}_i)$. The inverse Mills ratios are estimated as follows $\hat{\lambda}_{0i} = \frac{\phi(q_i)}{1-\Phi(q_i)} = \frac{\phi(q_i)}{1-\hat{p}_i}$, $\hat{\lambda}_{1i} = \frac{\phi(q_i)}{\hat{p}_i}$.

In the case of homogeneous impact (with common regression coefficients) described by equation (C6), a consistent two-stage estimate of θ can be obtained by running the OLS regression of **yi** on $x_i, d_i, \sigma_{u\varepsilon}[d_i \hat{\lambda}_{1i} + (1-d_i)\hat{\lambda}_{0i}]$ using all of the observations. In other terms, the estimating equation is (Lalonde 1986):

$$y_i = x_i\beta + \theta d_i + \sigma_{u\varepsilon}\hat{\lambda}_i + v_i \qquad \text{(C18)}$$

Maximum Likelihood Estimation (MLE)

The two-stage estimator is consistent but not asymptotically efficient. One may therefore wish to apply full information maximum likelihood to the endogenous switching model to gain efficiency. Note that observing **y1i** or **y0i** for the outcome variable **yi** is conditional on participation. Therefore, the contribution of each observation to the likelihood function is based on conditional probabilities. In particular, the contribution to the log likelihood can be written as.

$$l_i(\cdot) = d_i \ln \Pr(y_i \mid d_i = 1) + (1-d_i)\ln \Pr(y_i \mid d_i = 0) \qquad \text{(C19)}$$

In the case of participants, we have the following conditional probability:

$$\Pr(y_{1i} \mid d_i = 1) = \frac{\int_{-z_i\gamma}^{\infty} f(u_{1i},\varepsilon_i)d\varepsilon_i}{\Pr(\varepsilon_i > -z_i\gamma)} = \frac{\int_{-z_i\gamma}^{\infty} f(u_{1i},\varepsilon_i)d\varepsilon_i}{\Phi(z_i\gamma)} \qquad \text{(C20)}$$

The joint density function of **u1i** and **εi** can be factorized as $f(u_{1i},\varepsilon_i) = f(u_{1i}) \times f(\varepsilon_i \mid u_{1i})$. The normality assumption implies that $f(u_{1i},\varepsilon_i) = \frac{1}{\sigma_1}\phi(t_1)\times\phi(-t_{1\varepsilon})$ where $t_1 = \frac{(y_{1i} - x_i\beta_1)}{\sigma_1}$ and $t_{1\varepsilon} = \frac{z_i\gamma + \rho_{1\varepsilon}(y_{1i} - x_i\beta_1)/\sigma_1}{\sqrt{1-\rho_{1\varepsilon}^2}}$.

The result stems from the fact that, if **y** and **x** are jointly and normally distributed with parameters **μy**, **μx**, **σy**, **σx** and **ρxy**, then the conditional density function of **y** given **x**

[f(y|x)] is equal to the density of a normal distribution with mean $E(y \mid x) = \mu_y + \rho_{xy} \dfrac{\sigma_y}{\sigma_x}(x - \mu_x)$ and variance **(1-ρ^2)σ^2_y**.

Therefore $\Pr(y_{1i} \mid d_i = 1) = \dfrac{\dfrac{1}{\sigma_1}\phi(t_1)\int_{-z_i\gamma}^{\infty}\phi(-t_{1\varepsilon})d\varepsilon_i}{\Phi(z_i\gamma)}$. In other terms, the probability of observing **y$_{1i}$** conditional on participation can be written as follows.

$$\Pr(y_{1i} \mid d_i = 1) = \frac{\frac{1}{\sigma_1}\phi(t_1)[1 - \Phi(-t_{1\varepsilon})]}{\Phi(z_i\gamma)} = \frac{\frac{1}{\sigma_1}\phi(t_1) \times \Phi(t_{1\varepsilon})}{\Phi(z_i\gamma)} \qquad \text{(C21)}$$

Similarly, $\Pr(y_{0i} \mid d_i = 0) = \dfrac{\dfrac{1}{\sigma_0}\phi(t_0)\int_{-\infty}^{-z_i\gamma}\phi(-t_{0\varepsilon})d\varepsilon_i}{1 - \Phi(z_i\gamma)}$. Thus,

$$\Pr(y_{0i} \mid d_i = 0) = \frac{\frac{1}{\sigma_0}\phi(t_0)[1 - \Phi(t_{0\varepsilon})]}{1 - \Phi(z_i\gamma)} \qquad \text{(C22)}$$

The contribution of each observation to the log likelihood function can now be specified as follows: $l_i(\cdot) = d_i l_{1i}(\cdot) + (1 - d_i) l_{0i}(\cdot)$ where

$$l_{1i}(\cdot) = [-\ln \sigma_1 + \ln \phi(t_1) + \ln \Phi(t_{1\varepsilon}) - \ln \Phi(z_i\gamma)] \qquad \text{(C23)}$$

and

$$l_{0i}(\cdot) = [-\ln \sigma_0 + \ln \phi(t_0) + \ln[1 - \Phi(t_{0\varepsilon})] - \ln[1 - \Phi(z_i\gamma)]] \qquad \text{(C24)}$$

References

Baker, Judy L. 2000. *Evaluating the Impact of the Development Projects on Poverty: A Handbook for Practitioners*. Washington, D.C.: The World Bank (Directions in Development).

Becker, Sascha O. and Ichino, Andrea. 2002. Estimation of Average Treatment Effects Based on Propensity Score Matching. *The Stata Journal*, Vol.2, No. 4:358-377.

Blundell, Richard and Costa-Dias, Monica. 2002. Alternative Approaches to Evaluation in Empirical Microeconomics. Working Paper CWP10/02 Institute for Fiscal Studies, Department of Economics, University College London.

Blundell, Richard and Costa-Dias, Monica. 2000. Evaluation Methods for Non-Experimental Data. *Fiscal Studies*, Vol. 21, No. 4: 427-468.

Caliendo, Marco and Hujer, Reinhard. 2005. The Microeconometric Estimation of Treatment Effects—An Overview. IZA Discussion Paper No. 1653.

Caliendo, Marco and Kopeinig, Sabine. 2005. Some Practical Guidance for the Implementation of Propensity Score Matching. IZA Discussion Paper No. 1588.

Carneiro, Pedro, Hansen, Karsten T., and Heckman James J. 2002. Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies. National Bureau of Economic Research (NBER) Working Paper No. 8840.

Chen, Shaohua and Ravallion Martin. 2003. Hidden Impact? Ex-Post Evaluation of an Anti-Poverty Program. World Bank Policy Research Working Paper No. 3049.

Cleveland, William S. 1979. Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, Volume 74, Number 368:829-836.

Cleveland, William S., and Loader, Clive. 1996. Smoothing by Local Regression: Principles and Methods. In W. Hardle and M. G. Schimek (eds), *Statistical Theory and Computational Aspects of Smoothing*. Heidelberg: Physica Verlag.

Dabalen, Andrew, Paternostro, Stefano and Gaëlle, Pierre. 2004. The Returns to Participation in the Nonfarm Sector in Rural Rwanda. World Bank Policy Research Working Paper No. 3462.

Dehejia, Rajeev H. 2005. Practical Propensity Score Matching: A Reply to Smith and Todd. *Journal of Econometrics*, Volume 125, No. 1-2: 355-364.

Dehejia, Rajeev H, and Sadek Wahba. 2002. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *The Review of Economics and Statistics*, Vol.84, No.1: 151-161.

Dehejia, Rajeev H, and Wahba, Sadek. 1999. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association*, Vol. 94,No.448: 1053-1062.

Dehejia, Rajeev H, and Wahba, Sadek. 1997. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs, in Rajeev Dehejia, Econometric Methods for Program Evaluation, Ph. D. Dissertation, Harvard University (Chapter 1).

Diaz, Juan-Jose and Handa, Sudhanshu. 2004. Estimating the Selection Bias of Matching Estimators using Experimental Data from PROGRESA. Department of Economics , University of Maryland, and Department of Public Policy, University of North Caroline at Chapel Hill.

Essama-Nssah, B. 2004. Empowerment and Poverty-Focused Evaluation. *Development Southern Africa* Vol.21, No.3:509-530.

Greene, William H. 2000. *Econometric Analysis*. Upper Saddle River (New Jersey): Prentice Hall.

Grossman, Jean Baldwin. 1994. Evaluating Social Policies: Principles and U.S. Experience. *The World Bank Research Observer*, Vol 9, No.2: 159-180.

Hall, H. Brownwyn. 2002. Notes on Sample Selection Model (Copyrighted PDF downloaded from http://elsa.berkeley.edu/~bhhall/e244/sampsel.pdf).

Heckman, James. 2001. Accounting for Heterogeneity, Diversity and General Equilibrium in Evaluating Social Programs. *The Economic Journal*, 111 (November): F654-F699.

Heckman, James J., Ichimura, Hidehiko, and Todd Petra. 1998. Matching as an Econometric Evaluation Estimator. *Review of Economics Studies* 65:261-294.

Heckman, James J., Ichimura, Hidehiko, Smith, Jeffrey and Todd, Petra. 1998. Characterizing Selection Bias Using Experimental Data. *Econometrica*, Vol.66, No.5: 1017-1098.

Heckman, James J. and Smith, Jeffrey A. 1995. Assessing the Case for Social Experiments. *Journal of Economic Perspectives*, Vol.9, No. 2:85-110.

Heckman James J. 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, Vol. 5, No.4: 475-492.

Holland, Paul W. 1986. Statistics and Causal Inference. *Journal of the American Statistical Association*, Vol. 81, No. 396: 945-960.

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

Jalan, Jyotsna and Ravallion, Martin. 2003. Estimating the Benefit Incidence of an Antipoverty Program by Propensity-Score Matching. *Journal of Business and Economic Statistics*, Vol. 21, No. 1:19-30.

Lalonde, Robert J. 1986. Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review*, Vol. 76, No. 4:604-620.

Lee, Lung-Fei. 1983. Generalized Econometric Models with Selectivity. *Econometrica*, Vol. 51, No. 2:507-512.

Loader, Catherine. 2005. Regression, Likelihood, Smoothing. Cleveland: Department of Statistics, Case Western Reserve University. (Presentation at the Neuroinformatics Workshop, Wood Hole, MA, August 17, 2005).

Loader, Catherine. 2004. Smoothing: Local Regression Techniques. In James Gentle, Wolgang Hardle, Yoichi Mori (eds): *Handbook of Computational Statistics*. Heidelberg: Springer-Verlag.

Lokshin, Michael, and Yemtsov, Ruslan. 2003. Evaluating the Impact of Infrastructure Rehabilitation Projects on Household Welfare in Rural Georgia. World Bank Policy Research Working Paper 3155.

Lokshin, Michael and Sajaia, Zurab. 2004. Maximum Likelihood Estimation of Endogenous Switching Regression Models. *The Stata Journal*, Vol. 4, No. 3: 282-289.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Moffit, Robert A. 2004. Introduction to the Symposium on the Econometrics of Matching. *The Review of Economics and Statistics*, **86**(1): 1-3.

Monteiro, Natalia Pimenta. 2004. Using Propensity Matching Estimators to Evaluate the Impact of Privatization on Wages. NIPE Working Paper No. 12/2004 University of Minho.

Ravallion, Martin. 2005. Evaluating Anti-Poverty Programs. World Bank Policy Research Paper No. 3625. Washington D.C: The World Bank.

Ravallion, Martin. 2003. Assessing the Poverty Impact of an Assigned Program. In François Bourguignon and Luiz Pereira da Silva (eds) *The Impact of Economic Policies on Poverty and Income Distribution: Evaluation Techniques and Tools.* New York: Oxford University Press.

Ravallion, Martin. 2001. The Mystery of the Vanishing Benefits: An Introduction to Impact Evaluation. *The World Bank Economic Review*, Vol. 15, No. 1: 115-140.

Ravallion, Martin, and Chen, Shaohua. 2005. Hidden Impact? Household Saving in Response to a Poor-Area Development Project. Journal of Public Economics 89: 2183-2204

Ravallion, Martin, and Lokshin, Michael. 2004. Gainers and Losers from Trade Reform in Morocco. World Bank Policy Research Working Paper No. 3368. Washington D.C. The World Bank.

Ravallion, Martin, and Chen, Shaohua. 2003. Measuring Pro-Poor Growth. *Economics Letters* 78: 93-99.

Rosenbaum, Paul and Rubin, Donald. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, Vol.70, No.1:41-55.

Sianesi, Barbara. 2001. Implementing Propensity Score Matching Estimators with STATA. Presentation at the UK Stata Users Group, VII Meeting. London (May).

Silverman, B.W. 1986. Density *Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Smith, Jeffrey A. and Todd, Petra E. 2005a. Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators? *Journal of Econometrics*, Volume 125, No. 1-2: 305-353.

Smith, Jeffrey A. and Todd, Petra E. 2005b. Rejoinder. *Journal of Econometrics*, Volume 125, No. 1-2: 365-375.

Smith, Jeffrey A. and Todd, Petra E. 2001. Reconciling Conflicting Evidence on the Performance of Propensity-Score Matching Methods. *The American Economic Review*, Vol. 91, No.2: 112-118.

Wooldridge, Jeffrey M. 2002. *Econometric Analysis of Cross Section and Panel Data. Cambridge* (Massachusetts): The MIT Press.