Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated[1]

Petra E. Todd
University of Pennsylvania

March 19, 2006

[1] This chapter is under preparation for the *Handbook of Agricultural Economics*, Volume 4, North Holland, edited by Robert E. Evenson and T. Paul Schultz. This chapter was presented at a Conference at the Rockefeller Center in Bellagio, Italy, organized by the editors. I am grateful to T. Paul Schultz, Andrew Foster and other conference participants for very helpful comments.

Contents

1 Introduction

This chapter considers econometric methods for evaluating effects of social programs when the programs are nonrandomly placed and/or the program participants are nonrandomly selected. For example, family planning programs are often targeted at high fertility regions and individuals typically self-select into the programs. Similarly, health, education and nutrition interventions are often targeted at high poverty areas and eligibility for such programs may be restricted to individuals who meet some criteria. The focus of this chapter is on estimation of the effects of program interventions with nonexperimental data. Some of the estimation methods can also be adapted to experimental settings, to address related problems of nonrandom program attrition or dropout.

Two questions that are often of interest in evaluating effects of program interventions are (1) Do participants in programs benefit from them?, and (2) How would program impacts and costs differ if the features of the program were changed. We will consider different approaches to answering these questions in a way that recognizes heterogeneity in how individuals respond to treatment. We distinguish two types of evaluations, *ex post* evaluations, which analyze effects of existing programs, and *ex ante* evaluations, which analyze effects of counterfactual programs. Most of this chapter considers empirical methods for ex post evaluation, which is the most common. Section four takes up the problem of evaluating programs prior their implementation, which is useful for designing new programs or comparing existing programs to alternatives.

The goals of this chapter are (i) to describe the identifying assumptions needed to justify the application of different kinds of program or treatment effect estimators, (ii) to discuss the behavioral implications of these assumptions, (iii) to illustrate how different kinds of estimators are related to one another, (iv) to summarize the data requirements of the estimators and (v) to provide examples of how these evaluation methods have been applied in the development literature to assess the effects of program interventions.

2 The Evaluation Problem

We begin by introducing some notation for describing the evaluation problem and the parameters of interest. For simplicity, suppose there are only two states of the world, corresponding to the state of being with and without a treatment. For example, the outcome of interest could be a health measure and the treatment could be a health or nutrition intervention. Let $D = 1$ for persons who receive the intervention and $D = 0$ for persons who do not receive it. Associated with each state is a potential outcome. Y_0 denotes the potential outcome in the untreated state and Y_1 the potential outcome in the treated state. Each person has a (Y_0, Y_1) pair that represents the outcomes that would be realized in the two states. Because the person can only be in one state at a time, at most one of the two states is observed at any given point in time. The observed outcome is

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

The gain from moving an individual from the state "without treatment" to the state "with treatment" is

$$\Delta = Y_1 - Y_0.$$

Because of the missing data problem, the gain from treatment is not directly observed for anyone. Inferring gains from treatment requires solving a missing data problem. The evaluation literature has developed a variety of approaches.

2.1 Parameters of interest

In evaluating the effects of a social program, there may be many questions of interest, such as the benefits accruing to participants, any effects on nonparticipants, and program costs, which may include tax receipts used to finance the program. For example, consider the effects of a school subsidy program that provides incentive payments to parents for sending their children to school. If the subsidies are sufficiently large, we would expect such a program to have direct effects on the families participating in it. The program may also have indirect effects on nonparticipating families, perhaps through program-induced changes in the schools attended by nonparticipating children. If the program is financed from general taxes, the indirect effects might include any disincentives to work due to higher taxes. Thus, we distinguish between

direct effects: effects of the program on outcomes of program participants

indirect effects: effects of the program that are not directly related to program participation.

The program evaluation literature has focused mainly on estimating direct effects of the program and also on understanding what the program effects would be if the program offer were extended to other individuals not currently participating. Nonparticipants are often used as a source of control group data, so the maintained assumption typically is that the indirect effects on nonparticipants are negligible or that the goal of the evaluation is to uncover program effects on participants, conditional on any indirect effects.

Because program impacts are not directly observed for any individual, researchers usually hope to uncover only some features of the treatment impact distribution, such as its mean or median. Typical parameters of interest are the following:

(a) the proportion of program participants that benefit from the program

$$\Pr(Y_1 > Y_0|D = 1) = \Pr(\Delta > 0|D = 1)$$

(b) the proportion of the total population benefitting from the program:

$$\Pr(\Delta > 0|D = 1)\Pr(D = 1)$$

(c) quantiles of the impact distribution (such as the median), where q is the selected quantile

$$\inf_{\Delta}\{\Delta : F(\Delta|D = 1) > q\}$$

(d) the distribution of gains for individuals with some characteristics X_0

$$F(\Delta|D = 1, X = X_0),$$

where X represents some individual characteristics that are not affected by the program, such as age, education, race, or poverty level prior to the program intervention.

Much of the program evaluation literature develops methods for estimating two key parameters of interest:[1]

(e) the average gain from the program for persons with characteristics X

$$E(Y_1 - Y_0|X) = E(\Delta|X).$$

(f) the average gain from the program for program participants with characteristics X :

$$E(Y_1 - Y_0|X) = E(\Delta|D = 1, X).$$

The parameter (e) is commonly referred to as the *average impact of treatment (ATE)* and parameter (f) is known as the *average impact of treatment on the treated (TT).* If the individuals who take the program are the ones who tend to receive the greatest benefit from it, then we expect TT(X)>ATE(X).

2.2 What is the distinction between average program gain and average program gain for participants?

We will next consider the distinction between the ATE and the TT parameters and the conditions under which the two are the same. Suppose the outcomes in the treated and untreated states can be written as an additively separable function of observables (X) and unobservables (U_0 and U_1):

$$\begin{aligned} Y_1 &= \varphi_1(X) + U_1 \\ Y_0 &= \varphi_0(X) + U_0. \end{aligned}$$

The observed outcome $Y = DY_1 + (1 - D)Y_0$ can thus be written as:

$$Y = X\beta_0 + D(\varphi_1(X) - \varphi_0(X)) + \{U_0 + D(U_1 - U_0)\}.$$

[1] See, *e.g.*, Rosenbaum and Rubin (1985), Heckman and Robb (1985), or Heckman, Lalonde and Smith (1999).

Assume that $E(U_0|X) = E(U_1|X) = 0$. The gain to an individual from participating in the program is $\Delta = D(\varphi_1(X) - \varphi_0(X)) + D(U_1 - U_0)$. Individuals may or may not know their values of U_1 and U_0 at the time of deciding whether to participate in a program. If people self-select into the program based on their anticipated gains from the program, then we might expect that $E(U_0|X, D) \neq 0$ and $E(U_1|X, D) \neq 0$. That is, if the gain from the program depends on U_1 and U_0 and if people know their values of U_1 and U_0, or can to some extent forecast the values, then we would expect that people make use of this information when they decide whether to select into the program.

In the notation of the above statistical model for outcomes, the *average impact of treatment (ATE)* for a person with characteristics X is

$$\begin{aligned} E(\Delta|X) &= \varphi_1(X) - \varphi_0(X) + E(U_1|X) - E(U_0|X) \\ &= \varphi_1(X) - \varphi_0(X). \end{aligned}$$

The *average impact of treatment on the treated (TT)* is

$$E(\Delta|X) = \varphi_1(X) - \varphi_0(X) + E(U_1 - U_0|X, D = 1).$$

Note that the average effect of treatment on the treated combines the "structural parameters" (the parameters of the functions $\varphi_0(X)$ and $\varphi_1(X)$) with means of the unobservables.(See Heckman, 2000)

For completeness, we can also define the *average impact of treatment on the untreated (UT)* as

$$E(\Delta|X) = \varphi_1(X) - \varphi_0(X) + E(U_1 - U_0|X, D = 0),$$

which gives the impact of a program or intervention on the group that currently does not participate in it. This parameter may be of interest if there are plans to expand the scope of the program to include those currently nonparticipating.

Observe that if $U_1 = U_0$, then the TT, ATE and UT parameters are all the same. Allowing the random draw to differ in treated and untreated states is critical to allowing for unobserved heterogeneity in how people respond to treatment. There is, however, a special case where the parameters may be equal even if $U_1 \neq U_0$, that is, when

$$E(U_1 - U_0|X, D) = 0.$$

Under this restriction, D is uninformative on $U_1 - U_0$, but it is not necessarily the case that $U_1 = U_0$. The conditional mean restriction might be satisfied if agents making the participation decisions (e.g. individuals, program administrators or others) do not act on $U_1 - U_0$ in making the decision, perhaps because they do not know anything about their own indiosyncractic gain from participating in the program at the time of deciding whether to participate. In this special case, there is *ex post* heterogeneity in how people respond to treatment, but it is not acted upon *ex ante.*

As discussed in Heckman, Lalonde and Smith (1999), there are three different types of assumptions that can be made in the evaluation model that vary in their level of generality:

(A.1) conditional on X, the program effect is the same for everyone ($U_1 = U_0$)

(A.2) the program effect (given X) varies across individuals but $U_1 - U_0$ does not help predict participation in the program

(A.3) the program effect (given X) varies across individuals and $U_1 - U_0$ does predict who participates in the program.

(A.1) is the most restrictive and (A.3) the most general. We will consider ways of estimating the TT and ATE parameters of interest under these three different sets of assumptions.

2.3 Sources of Bias in Estimating $\mathbf{E}(\Delta|X, D = 1)$ and $\mathbf{E}(\Delta|X)$

The model of the last section can be rewritten as

$$Y = \varphi_0(X) + D(\varphi_1(X) - \varphi_0(X)) + \{U_0 + D(U_1 - U_0)\}.$$

In terms of the two parameters of interest, (ATE=$E(\Delta|X)$ and TT=$E(\Delta|X, D = 1)$), the outcomes model can be written as:

$$Y = \varphi_0(X) + DE(\Delta|X) + \{U_0 + D(U_1 - U_0)\} \tag{*}$$

or

$$Y = \varphi_0(X) + DE(\Delta|X, D = 1) + \{U_0 + D[U_1 - U_0 - E(U_1 - U_0|X, D = 1)]\}.$$

For simplicity, suppose the X variables are discrete and that we estimated the effects of the intervention (D) by the coefficients $\hat{b}_x$ from an ordinary least squares regression:

$$y = aX + b_x XD + v.^2$$

This model, which is popular in applied work, is known as the "common effect" model. In light of the true outcome model, bias for the ATE parameter ($E(\Delta|X)$) arises if the mean of the error term does not have conditional mean zero, i.e.

$$E(U_0 + D(U_1 - U_0)|X, D)) \neq 0.$$

Under assumption A.1 and A.2, bias arises only from $E(U_0|X, D) \neq 0$, but under the more general assumption A.3, there is also potential bias arising from $E(U_1 - U_0|D, X) \neq 0$. For estimating the TT parameter $E(\Delta|X, D = 1)$, under assumptions A.1-A.3, bias arises if $E(U_0|X, D) \neq 0$.

3 Solutions to the Evaluation Problem

3.1 Traditional Regression Estimators

Nonexperimental estimators of program impacts typically use two types of data to impute the missing counterfactual outcomes for program participants: data on participants at a point in time prior to entering the program and data on nonparticipants. We next consider three widely used methods for estimating the (TT) parameter, $E(\Delta|X, D = 1)$, using nonexperimental data: (a) the *before-after* estimator, (b) the *cross-section* estimator, and (c) the *difference-in-difference* estimator. The extensions to estimating the ATE parameter are straightforward.

To describe the estimators and their assumptions, we introduce a panel data regression framework. Using the same notation as previously, denote the outcome measures by Y_{1it} and Y_{0it}, where i denotes the individual and t the time period of observations, where

$$\begin{aligned} Y_{1it} &= \varphi_1(X_{it}) + U_{1it} \qquad (1) \\ Y_{0it} &= \varphi_0(X_{it}) + U_{0it}. \end{aligned}$$

U_{1it} and U_{0it} are assumed to be distributed independently across persons and to satisfy $E(U_{1it}|X_{it}) = 0$ and $E(U_{0it}|X_{it}) = 0$. X_{it} represents conditioning variables that may either be fixed or time-varying (such as gender or age), but whose distributions are assumed to be unaffected by whether an individual participates in the program. We can write the observed outcome at time t as

$$Y_{it} = \varphi_0(X_{it}) + D_{it}\alpha^*(X_{it}) + U_{0it}, \qquad (2)$$

where D_{it} denotes being a program participant in the program and $\alpha^*(X_{it}) = \varphi_1(X_{it}) - \varphi_0(X_{it}) + U_{1it} - U_{0it}$ is the treatment impact for an individual. Prior to the program intervention, we observe $Y_{0it} = \varphi_0(X_{it}) + U_{0it}$ for everyone. After the intervention we observe $Y_{1it} = \varphi_1(X_{it}) + U_{1it}$ for those who received the intervention (for whom $D_i = 1$) and $Y_{0it} = \varphi_0(X_{it}) + U_{0it}$ for those who did not receive it.

This model is a random coefficient model, because the impact of treatment is assumed to vary across persons even after conditioning on X_{it}. Assuming that $U_{0it} = U_{1it} = U_{it}$, so that the unobservable is the same in both the treated and untreated states, and assuming that $\varphi_1(X_{it}) - \varphi_0(X_{it})$ is constant with respect to X_{it} yields the fixed coefficient or "common effect" version of the model.

3.1.1 Before-After Estimators

As described above, the evaluation problem can be viewed as a missing data problem, because each person is only observed in one of two potential states at any point in time. The before-after estimator addresses the missing data problem by using pre-program data to impute the missing counterfactual outcomes for program participants.

To simplify notation, assume that the treatment impact is constant across individuals (i.e. the common effect assumption where $\varphi_1(X_{it}) = \varphi_0(X_{it}) + \alpha^*$). Let t' and t denote two time periods, one before and one after the program intervention. In a regression model, the before-after estimator is the least squares solution to α^* in

$$Y_{it} - Y_{it'} = \varphi_0(X_{it}) - \varphi_0(X_{it'}) + \alpha^* + U_{it} - U_{it'}.$$

Consistency of the estimator for α^* requires that $E(U_{it} - U_{it'}|D_{it} = 1, D_{it'} = 0, X_{it}) = 0$. A special case where this assumption is satisfied is if U_{it} can be decomposed into a fixed effect error structure, $U_{it} = f_i + v_{it}$, where f_i depends on i but does not vary over time and v_{it} is a iid random error term. For this error structure, it is necessary to assume $E(v_{it} - v_{it'}|D_i = 1, X_{it}) = 0$. Note that this assumption allows selection into the program to be based on f_i , so the estimation strategy admits to person-specific permanent unobservables.

A major drawback of a before-after estimation strategy is that identification of α^* breaks down in the presence of time-specific intercepts, making it impossible to separate effects of the program from other general time effects on outcomes.[3] Before-after estimates can also be sensitive to the choice of time periods used to construct the estimator. For example, many studies of employment and training programs in the U.S. and in other countries have noted that earnings and employment of training program participants dip down in the time period just prior to entering the program, a pattern now known as *Ashenfelter's Dip.*(See Ashenfelter, 1978, Heckman and Smith, 1999, and Heckman LaLonde and Smith, 1999). The dip pattern can arise from serially correlated transitory shocks to earnings that may have been the impetus for the person applying to the training program.[4] Another potential explanation for the dip pattern are the program eligibility criteria that are often imposed to select out the most disadvantaged persons for participation in programs. These criteria will

[3] Suppose $\varphi(X_{it}) = X_{it}\beta + \gamma_t$, where γ_t is a time specific intercept common across individuals. Such a common time effect may arise, e.g., from life-cycle wage growth over time or from shocks to the economy. In this case, α^* is confounded with $\gamma_t - \gamma_{t'}$.

[4] A fixed effect error structure would not generate a dip pattern.

select into the program persons with low transitory earnings shocks. A simple before-after estimation strategy that includes the preprogram "dip" period typically gives an upward biased estimate of the effect of the program.

An advantage of the before-after estimator relative to other classes of estimators is that it can be implemented even when data are available only on program participants. At a minimum, two cross-sections of data, one pre-program and post-program, are required to implement the estimator.

3.1.2 Cross-section Estimators

A *cross-section* estimator uses data on a comparison group of nonparticipants to impute counterfactual outcomes for program participants. The data requirements of this estimator are minimal, only one post-program cross section of data on $D_{it} = 1$ and $D_{it} = 0$ persons. Define $\hat{\alpha}_{CS}$ as the ordinary least squares solution to α^* in

$$Y_{it} = \varphi_0(X_{it}) + D_{it}\alpha^* + U_{it},$$

where the regression is estimated using data on persons for which $D_{it} = 1$ and $D_{it} = 0$. Consistency of the cross-section estimator requires that $E(U_{it}|D_{it}, X_{it}) = 0$. In a more general model where $U_{0it} \neq U_{1it}$, this restriction rules out the possibility that people select into the program based on expectations about their idiosyncratic gain from the program.

3.1.3 Difference-in-Differences Estimators

The *difference-in-differences* (DID) estimator is commonly used in evaluation work, as seen in the applications described below (in 3.1.5). It measures the impact of the program intervention by the difference in the before-after change in outcomes between participants and nonparticipants, which requires pre- and post-program data (t and t' data) on program participants and nonparticipants.

Define an indicator that equals 1 for participants (for whom $D_{it'} = 0$ and $D_{it} = 1$), denoted by I_i^D and zero otherwise. The difference-in-differences treatment effect estimator is the least squares solution for α^* in

$$Y_{it} - Y_{it'} = \varphi_0(X_{it}) - \varphi_0(X_{it'}) + D_i\alpha^* + \{U_{it} - U_{it'}\},$$

which allows for individual fixed effects that are differenced-out. Alternatively, the DID estimator is often implemented using a regression

$$Y_{it} = \varphi_0(X_{it}) + I_i^D\gamma + D_{it}\alpha^* + U_{it} \text{ for } t = t', .., t,$$

where I_i^D is an intercept that denotes whether a member of the treatment group.[5] This regression is estimated using participant $(D_{it'} = 0, D_{it} = 1)$ and nonparticipant $(D_{it'} = 0, D_{it} = 0)$ observations.The DID estimator addresses the main shortcoming of the before-after estimator in that it allows for time-specific intercepts that are common across groups (included in $\varphi_0(X_{it})$). The estimator is consistent if $E(U_{it} - U_{it'}|D_{it} - D_{it'}, X_{it}) = 0$. The data requirements are either longitudinal or repeated cross-section data on both participants and nonparticipants.

3.1.4 Within Estimators

Within estimators identify program impacts from changes in outcomes within some unit, such as an individual, a family, a school or a village. The previously described before-after and DID estimators, when implemented using longitudinal data, are examples of within estimators. We next describe other types of within estimators where the size of the unt is broader than a single individual and may, for example, represent a family or village.

Let Y_{0ijt} and Y_{1ijt} denote the outcomes for individual i, from unit j, observed at time t. Write the model for outcomes as:

$$Y_{ijt} = \varphi_0(X_{ijt}) + I_{ij}^D\gamma + D_{ijt}\alpha^* + U_{ijt}$$

and assume that the error term U_{ijt} can be decomposed as: $U_{ijt} = \theta_j + v_{ijt}$, where θ_j represents the effects of unobservables that vary across units but are constant for individuals within the same unit and v_{ijt} are *iid*.

[5] The specification could include individual-specific fixed effects, but estimating them consistently would require an assumption that the panel length T go to infinity.

Consider taking differences between two individuals from the same unit observed in the same time period:

$$Y_{ijt} - Y_{i'jt} = \varphi_0(X_{ijt}) - \varphi_0(X_{i'jt}) + (I_{ij}^D - I_{i'j}^D)\gamma + (D_{ijt} - D_{i'jt})\alpha^* + (v_{ijt} - v_{i'jt}).$$

Consistency of the ols estimator of α^* requires that $E(v_{ijt} - v_{i'jt}|X_{ijt}, X_{i'jt}, D_{ijt}, D_{i'jt}) = 0$. This assumption implies that within a particular unit, which individual receives the treatment is random with respect to the error term v_{ijt}.

As with the before-after and difference-in-differences estimation approaches, this estimator allows treatment to be selective across units; namely, it allows $E(U_{ijt}|D_{ijt}, X_{ijt}) \neq 0$, because treatment selection can be based on the unobserved heterogeneity term θ_j. The data requirements of this within estimator are a single cross-section of data. Because the estimator relies on comparisons between the outcomes of treated and untreated persons within the same unit, the approach implicitly requires that there be no spillover effects from treating one individual on other individuals within the same unit.

Sometimes it happens that all individuals within a unit receive treatment at the same time, in which case $D_{ijt} = D_{i'jt}$ for all i in j and the above approach is not feasible. In that situation, the within estimator can still be implemented if preprogram data (t') are available by taking differences across individuals in the same unit observed at different time periods:

$$Y_{ijt} - Y_{i'jt'} = \varphi_0(X_{ijt}) - \varphi_0(X_{i'jt'}) + (I_{ij}^D - I_{i'j}^D)\gamma + (D_{ijt} - D_{i'jt'})\alpha^* + (v_{ijt} - v_{i'jt'}),$$

where $D_{i'jt'} = 0$. Consistency of this estimator requires that $E(v_{ijt} - v_{i'jt'}|D_{ijt}, D_{i'jt'}, X_{ijt}, X_{i'jt'}) = 0$. When $I_{ij}^D = 1$ for all i, j, the estimation method is analogous to a before-after estimator, except that comparisons are between different individuals within the same unit across time. [6]

3.1.5 Applications

The above estimators are widely used in empirical evaluation research on developing countries. One of the earliest applications of the within estimator is Rosenzweig and Wolpin

[6] In that case, the estimator suffers from the same drawback as the before-after estimator of not being able to separately identify time effects.

(1986), which evaluates the impact of a family planning and health counseling program on child outcomes in the Phillipines. Their study also provides an early discussion in economics of the statistical problems created by nonrandom program placement, in particular, when the placement of a program may depend on the outcome variable of interest. Their empirical analysis adopts the following estimating framework:

$$H_{ijt}^{a} = \rho_{ij}^{a}\beta + \mu_i + \mu_j + \varepsilon_{ijt},$$

where H_{ijt}^{a} is a child health measure (height, weight) for child i of age a, living in locality j at time t. ρ_{ij}^{a} represents the length of time that child was exposed to the program intervention. μ_i is a time invariant, child-specific unobserved health endowment and μ_j a locality level effect. The estimation approach they adopt compares changes in health outcomes for children who were exposed to the program to changes for children who were not exposed to it.[7] This evaluation method allows the allocation of the program to be selective based on locality level and individual level unobserved characteristics. Another study by Rosenzweig and Wolpin (1988) uses a similar within child estimation strategy to evaluate the effects of a Colombian child health intervention.

A recent evaluation study that adopts a similar evaluation approach is that of Duflo (2001), in which a within estimator is used to evaluate the effects of a school construction program in Indonesia on education and wages. She notes that the new schools were in part locally financed, which led to nonrandom placement of schools in the more affluent communities. Because individuals from those communities likely experience better outcomes even in the absence of the intervention, it is difficult to draw reliable inferences from cross-sectional comparisons of localities with and without the new schools. Duflo observed that exposure to the school construction program varied by region of birth and date of birth, so that the education of individuals who were young when the program began should be more affected by the program than that of older individuals. Also, individuals in regions where a larger numbers of schools were built should also be more likely to have been affected by

[7] In the specification, locality level effects can be separately identified from individual effects using observations on families that migrated across localities.

the building programs. Essentially, her identification strategy draws comparisons between outcomes of older and younger individuals in regions where the school construction program was very active with those of similar individuals in regions where the school construction program was less active.

In a recent paper, Glewwe, Kremer, Moulin and Zitzewitz (2004) question the reliability of a difference-in-difference estimator in an application to evaluating the effectiveness of an educational intervention in Kenya, which provided schools with flip-charts to use as teaching aids in certain subjects. One of the goals of their study is to assess the efficacy of a nonexperimental DID estimation approach, by comparing the results obtained by DID to those obtained from a randomized social experiment. Their DID estimation strategy compares changes over time in test scores in flip-chart and non-flip-chart subjects within the schools that received the intervention. The experiment randomly allocated the schooling intervention (flip-charts) to a subset of schools. When Glewwe et. al. (2004) compare the nonexperimental DID estimates to the experimental estimates, they find substantial differences between the two sets of estimates. The experimental estimates suggest that flip-charts had little effect on test scores, while the DID estimates are statistically significantly different from zero at conventiona levels. The authors conclude that the difference-in-difference estimator is unrealiable.[8] Glewwe, Kremer, and Moulin (2000, 2003) carry out a similar comparison between a nonexperimental DID estimator and an experimental estimator, in which they evaluate other schooling inteventions.

3.2 Matching Methods

Matching is a widely-used method of evaluation that compares the outcomes of program participants with the outcomes of similar, matched nonparticipants. Their use in evaluating the effects of program interventions in developing country settings is relatively new. Some

[8] In their application, an implicit assumption of the DID estimator is that having flip-charts in certain subjects does not affect students' achievements in other subjects. For example, the DID estimator could be invalid if teachers spent more time teaching flip-chart subjects as a result of the intervention and less time on other subjects. This may account for the deviation between the experimental and the nonexperimental DID estimates.

of the earliest applications of matching to evaluate programs in economic development were World Bank evaluations of anti-poverty programs. [9]

One of the main advantages of matching estimators over other kinds of evaluation estimators is that they do not require specifying the functional form of the outcome equation and are therefore not susceptible to bias due to misspecification along that dimension. For example, they do not require specifying that outcomes are linear in observables. Traditional matching estimators pair each program participant with an observably similar nonparticipant and interpret the difference in their outcomes as the effect of the program intervention (see, e.g., Rosenbaum and Rubin, 1983). More recently developed methods pair program participants with more than one nonparticipant observation, using statistical methods to estimate the matched outcome. In this discussion, we focus on a class of matching estimators called *propensity score matching* estimators, because these methods are the most commonly used and have been shown in some studies to be reliable, under the conditions described below.[10]

Matching estimators typically assume that there exist a set of observed characteristics Z such that outcomes are independent of program participation conditional on Z. That is, it is assumed that the outcomes (Y_0, Y_1) are independent of participation status D conditional on Z,[11]

$$(Y_0, Y_1) \perp\!\!\!\perp D \; |Z \; . \tag{3}$$

It is also assumed that for all Z there is a positive probability of either participating ($D = 1$) or not participating ($D = 0$) in the program, i.e.,

$$0 < \Pr(D = 1|Z) < 1. \tag{4}$$

This second assumption is required so that a matches for $D = 0$ and $D = 1$ observations can be found. If assumptions (3) and (4) are satisfied, then the problem of determining mean

[9] See the applications described below.

[10] For discussions of other kinds of matching estimators, see e.g. Cochran (1973), Rubin (1980, 1984).

[11] In the terminology of Rosenbaum and Rubin (1983) treatment assignment is "strictly ignorable" given Z.

program impacts can be solved simply by substituting the Y_0 distribution observed for the matched non-participant group for the missing Y_0 distribution for program participants.

Heckman, Ichimura and Todd (1998) show that the above assumptions are overly strong if the parameter of interest is the mean impact of treatment on the treated (TT), in which case a weaker conditional mean independence assumption on Y_0 suffices:

$$E(Y_0|Z, D = 1) = E(Y_0|Z, D = 0) = E(Y_0|Z). \tag{5}$$

Furthermore, when TT is the parameter of interest, the condition $0 < \Pr(D = 1|Z)$ is also not required, because that condition only guarantees the possibility of a participant analogue for each non-participant. The TT parameter requires only

$$\Pr(D = 1|Z) < 1. \tag{6}$$

Under these assumptions, the mean impact of the program on program participants can be written as

$$\begin{aligned} \Delta &= E(Y_1 - Y_0|D = 1) \\ &= E(Y_1|D = 1) - E_{Z|D=1}\{E_Y(Y|D = 1, Z)\} \\ &= E(Y_1|D = 1) - E_{Z|D=1}\{E_Y(Y|D = 0, Z)\}, \end{aligned}$$

where the second term can be estimated from the mean outcomes of the matched (on Z) comparison group.[12] Assumption (5) implies that D does not help predict values of Y_0 conditional on Z. Thus, selection into the program cannot be based directly on values of Y_0.However, no restriction is imposed on Y_1, so the method does allow individuals who expect high levels of Y_1 to be selecting into the program. Thus, it permits assumption (A-3) discussed in section 1.3, with some restrictions on the nature of the selection process governing program participation decisions.

With nonexperimental data, there may or may not exist a set of observed conditioning variables for which (3) and (4) hold. A finding of Heckman, Ichimura and Todd (1997)

[12] The notation $E_{Z|D=1}$ denotes that the expectation is taken with respect to the $f(Z|D = 1)$ density.

and HIST (1996,1998) in their application of matching methods to JTPA data is that (4) was not satisfied, meaning that for a fraction of program participants no match could be found. If there are regions where the support of Z does not overlap for the $D = 1$ and $D = 0$ groups, then matching is only justified when performed over the *region of common support*.[13] The estimated treatment effect must then be defined conditionally on the region of overlap. Empirical methods for determining the region of overlap are described below.

3.2.1 Reducing the Dimensionality of the Conditioning Problem

Matching can be difficult to implement when the set of conditioning variables Z is large.[14] Rosenbaum and Rubin (1983) provide a theorem that is useful in reducing the dimension of the conditioning problem in implementing the matching method. They show that for random variables Y and Z and a discrete random variable D

$$E(D|Y, P(D = 1|Z)) = E(E(D|Y, Z)|Y, \Pr(D = 1|Z)),$$

so that

$$E(D|Y, Z) = E(D|Z) \Longrightarrow E(D|Y, \Pr(D = 1|Z) = E(D|\Pr(D = 1|Z)).$$

This result implies that when Y_0 outcomes are independent of program participation conditional on Z, they are also independent of participation conditional on the probability of participation, $P(Z) = \Pr(D = 1|Z)$. Thus, when matching on Z is valid, matching on the summary statistic $\Pr(D = 1|Z)$ (the *propensity score*) is also valid. Provided that $P(Z)$ can be estimated parametrically (or semiparametrically at a rate faster than the nonparametric rate), matching on the propensity score reduces the dimensionality of the matching prob-

[13] An advantage of randomized experiments noted by Heckman (1997), as well as Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998), is that they guarantee that the supports are equal across treatments and controls, so that the mean impact of the program can always be estimated over the entire support.

[14] If Z is discrete, small cell problems may arise. If Z is continuous and the conditional mean $E(Y_0|D = 0, Z)$ is estimated nonparametrically, then convergence rates will be slow due to the "curse of dimensionality" problem.

lem to that of a univariate matching problem. For this reason, much of the literature on matching focuses on propensity score matching methods.[15]

Using the Rosenbaum and Rubin (1983) theorem, the matching procedure can be broken down into two stages. In the first stage, the propensity score $\Pr(D=1|Z)$ is estimated, using a binary discrete choice model such as a logit or probit. In the second stage, individuals are matched on the basis of their predicted probabilities of participation, obtained from the first stage.

The literature has developed a variety of matching estimators. We next describe some of the leading examples.

3.2.2 Alternative Matching Estimators

For notational simplicity, let $P = P(Z)$. A typical matching estimator takes the form

$$\hat{\alpha}_M = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} [Y_{1i} - \hat{E}(Y_{0i}|D=1, P_i)] \tag{8}$$

$$\hat{E}(Y_{0i}|D=1, P_i) = \sum_{j \in I_0} W(i,j) Y_{0j},$$

where I_1 denotes the set of program participants, I_0 the set of non-participants, S_P the region of common support (see below for ways of constructing this set). n_1 denotes the number of persons in the set $I_1 \cap S_P$. The match for each participant $i \in I_1 \cap S_P$ is constructed as a weighted average over the outcomes of non-participants, where the weights $W(i,j)$ depend on the distance between P_i and P_j.

Define a neighborhood $C(P_i)$ for each i in the participant sample. Neighbors for i are non-participants $j \in I_0$ for whom $P_j \in C(P_i)$. The persons matched to i are those people in set A_i where $A_i = \{j \in I_0 \mid P_j \in C(P_i)\}$. Alternative matching estimators (discussed below) differ in how the neighborhood is defined and in how the weights $W(i,j)$ are constructed.

[15] Heckman, Ichumura and Todd (1998) and Hahn (1998) consider whether it is better in terms of efficiency to match on $P(X)$ or on X directly. For the TT parameter, they show that neither is necessarily more efficient than the other. If the treatment effect is constant, then it is more efficient to condition on the propensity score.

Nearest Neighbor matching Traditional, pairwise matching, also called *nearest-neighbor matching*, sets

$$C(P_i) = \min_j \|P_i - P_j\|, \; j \in I_0.$$

That is, the non-participant with the value of P_j that is closest to P_i is selected as the match and A_i is a singleton set. The estimator can be implemented either matching with or without replacement. When matching is performed with replacement, the same comparison group observation can be used repeatedly as a match. A drawback of matching without replacement is that the final estimate will likely depend on the initial ordering of the treated observations for which the matches were selected. The nearest neighbor matching estimator is often used in practice, in part due to ease of implementation.

Caliper matching *Caliper matching* (Cochran and Rubin, 1973) is a variation of nearest neighbor matching that attempts to avoid "bad" matches (those for which P_j is far from P_i) by imposing a tolerance on the maximum distance $\|P_i - P_j\|$ allowed. That is, a match for person i is selected only if $\|P_i - P_j\| < \varepsilon$, $j \in I_0$, where ε is a pre-specified tolerance. For caliper matching, the neighborhood is $C(P_i) = \{P_j \mid \|P_i - P_j\| < \varepsilon\}$. Treated persons for whom no matches can be found (within the caliper) are excluded from the analysis. Thus, caliper matching is one way of imposing a common support condition. A drawback of caliper matching is that it is difficult to know a priori what choice for the tolerance level is reasonable.

Stratification or Interval Matching In this variant of matching, the common support of P is partitioned into a set of intervals. Within each interval, a separate impact is calculated by taking the mean difference in outcomes between the $D = 1$ and $D = 0$ observations within the interval. A weighted average of the interval impact estimates, using the fraction of the $D = 1$ population in each interval for the weights, provides an overall impact estimate. Implementing this method requires a decision on how wide the intervals should be. Dehejia and Wahba (1999) implement interval matching using intervals that are selected such that the mean values of the estimated P_i's and P_j's are not statistically different from each other within intervals.

Kernel and Local Linear matching More recently developed matching estimators construct a match for each program participant using a weighted average over multiple persons in the comparison group. Consider, for example, the nonparametric *kernel matching estimator,* given by

$$\hat{\alpha}_{KM} = \frac{1}{n_1} \sum_{i \in I_1} \left\{ Y_{1i} - \frac{\sum_{j \in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)} \right\}.$$

where $G(\cdot)$ is a kernel function and a_n is a bandwidth parameter.[16] In terms of equation (8), the weighting function, $W(i,j)$, is equal to $\frac{G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}$. For a kernel function bounded between -1 and 1, the neighborhood is $C(P_i) = \{|\frac{P_i - P_j}{a_n}| \leq 1\}$, $j \in I_0$. Under standard conditions on the bandwidth and kernel , $\frac{\sum_{j \in I_0} Y_{0j} G\left(\frac{P_j - P_i}{a_n}\right)}{\sum_{k \in I_0} G\left(\frac{P_k - P_i}{a_n}\right)}$ is a consistent estimator of $E(Y_0|D=1,P_i)$.[17]

Heckman, Ichimura and Todd (1997) also propose a generalized version of kernel matching, called local linear matching.[18] The local linear weighting function is given by

$$W(i,j) = \frac{G_{ij} \sum\limits_{k \in I_0} G_{ik}(P_k - P_i)^2 - [G_{ij}(P_j - P_i)][\sum\limits_{k \in I_0} G_{ik}(P_k - P_i)]}{\sum\limits_{j \in I_0} G_{ij} \sum\limits_{k \in I_0} G_{ij}(P_k - P_i)^2 - \left(\sum\limits_{k \in I_0} G_{ik}(P_k - P_i)\right)^2}. \qquad (9)$$

As demonstrated in research by Fan (1992a,b), local linear estimation has some advantages over standard kernel estimation. These advantages include a faster rate of convergence near boundary points and greater robustness to different data design densities. (See Fan (1992a,b).) Thus, local linear regression would be expected to perform better than kernel estimation in cases where the nonparticipant observations on P fall on one side of the participant observations.

[16] See Heckman, Ichimura and Todd (1997, 1998) and Heckman, Ichimura, Smith and Todd (1998),

[17] Specifically, we require that $G(\cdot)$ integrates to one, has mean zero and that $a_n \to 0$ as $n \to \infty$ and $na_n \to \infty$.

[18] Recent research by Fan (1992a,b) demonstrated advantages of local linear estimation over more standard kernel estimation methods. These advantages include a faster rate of convergence near boundary points and greater robustness to different data design densities. See Fan (1992a,b).

To implement the matching estimator given by equation (8), the region of common support S_P needs to be determined. To determine the support region, Heckman, Ichimura and Todd (1997) use kernel density estimation methods. The common support region can be estimated by

$$\hat{S}_P = \{P : \hat{f}(P|D=1) > 0 \text{ and } \hat{f}(P|D=0) > c_q\},$$

where $\hat{f}(P|D=d)$, $d \in \{0,1\}$ are nonparametric density estimators given by

$$\hat{f}(P|D=d) = \sum_{k \in I_d} G\left(\frac{P_k - P}{a_n}\right),$$

where a_n is a bandwidth parameter. To ensure that the densities are strictly greater than zero, it is required that the densities be strictly positive density (i.e. exceed zero by a certain amount), determined using a "trimming level" q. That is, after excluding any P points for which the estimated density is zero, we exclude an additional small percentage of the remaining P points for which the estimated density is positive but very low. The set of eligible matches is therefore given by

$$\hat{S}_q = \{P \in \hat{S}_P : \hat{f}(P|D=1) > c_q \text{ and } \hat{f}(P|D=0) > c_q\},$$

where c_q is the density cut-off level that satisfies:

$$\sup_{c_q} \frac{1}{2J} \sum_{\{i \in I_1 \cap \hat{S}_P\}} \{1(\hat{f}(P|D=1) < c_q + 1(1(\hat{f}(P|D=0) < c_q\} \leq q.$$

Here, J is the cardinality of the set of observed values of P that lie in $I_1 \cap \hat{S}_P$. That is, matches are constructed only for the program participants for which the propensity scores lie in $\hat{S}_q$.

The above estimators are straightforward representations of matching estimators and are commonly used. The recent literature has developed some alternative, more efficient estimators. See, for example, Hahn (1998) and Hirano, Imbens and Ridder (2003). In addition, Heckman, Ichimura and Todd (1998) propose a regression-adjusted matching estimator that replaces Y_{0j} as the dependent variable with the residual from a regression of Y_{0j} on a vector of exogenous covariates. The estimator explicitly incorporates exclusion restrictions, i.e. that some of the conditioning variables in the outcome equation do not enter into the participation equation or vice versa. In principal, imposing exclusions restrictions can increase efficiency. In practice, though, researchers have not observed much gain from using the regression-adjusted matching estimator.

Difference-in-difference matching The estimators described above assume that after conditioning on a set of observable characteristics, mean outcomes are conditionally mean independent of program participation. However, for a variety of reasons there may be systematic differences between participant and nonparticipant outcomes, even after conditioning on observables, that could lead to a violation of the identification conditions required for matching. Such differences may arise, for example, because of program selectivity on unmeasured characteristics, or because of levels differences in outcomes across different labor markets in which the participants and nonparticipants reside.

A difference-in-differences (DID) matching strategy, as defined in Heckman, Ichimura and Todd (1997) and Heckman, Ichimura, Smith and Todd (1998), allows for temporally invariant differences in outcomes between participants and nonparticipants. This type of estimator is analogous to the standard differences-in-differences regression estimator defined in Section 3.1, but it reweights the observations according to the weighting functions used by the propensity score matching estimators defined above. The DID matching estimator requires that

$$E(Y_{0t} - Y_{0t'}|P, D = 1) = E(Y_{0t} - Y_{0t'}|P, D = 0),$$

where t and t' are time periods after and before the program enrollment date. This estimator also requires the support condition given in (7), which must now hold in both periods t and t'. The local linear difference-in-difference estimator is given by

$$\hat{\alpha}_{KDM} = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} \left\{ (Y_{1ti} - Y_{0t'i}) - \sum_{j \in I_0 \cap S_P} W(i,j)(Y_{0tj} - Y_{0t'j}) \right\},$$

where the weights correspond to the local linear weights defined above. If repeated cross-section data are available, instead of longitudinal data, the estimator can be implemented as

$$\hat{\alpha}_{KDM} = \frac{1}{n_{1t}} \sum_{i \in I_{1t} \cap S_P} \left\{ (Y_{1ti} - \sum_{j \in I_{0t} \cap S_P} W(i,j) Y_{0tj} \right\} - \frac{1}{n_{1t'}} \sum_{i \in I_{1t'} \cap S_P} \left\{ (Y_{1t'i} - \sum_{j \in I_{0t'}} W(i,j) Y_{0t'j} \right\},$$

where $I_{1t}, I_{1t'}, I_{0t}, I_{0t'}$ denote the treatment and comparison group datasets in each time period.

Finally, the DID matching estimator allows selection into the program to be based on anticipated gains from the program, in the sense of assumption (A-3) described in section 1.3. That it, D can help predict the value of Y_1 given P. However, the method assumes that D does not help predict changes in the value of Y_0 $(Y_{0t} - Y_{0t'})$ conditional on P. .Thus, individuals who participate in the program may be the ones who expect the highest values of Y_1 , but they may not be systematically difference in terms of their changes in Y_0.

3.2.3 Matching when the Data are Choice-based Sampled

The samples used in evaluating the impacts of programs are often choice-based, with program participants oversampled relative to their frequency in the population of persons eligible for the program. Under choice-based sampling, weights are generally required to consistently estimate the probabilities of program participation.[19] When the weights are unknown, Heckman and Todd (1995) show that with a slight modification, matching methods can still be applied, because the odds ratio $(P/(1-P))$ estimated using a logistic model with incorrect weights (i.e., ignoring the fact that samples are choice-based) is a scalar multiple of the true odds ratio, which is itself a monotonic transformation of the propensity scores. Therefore, matching can proceed on the (misweighted) estimate of the odds ratio (or of the log odds ratio).[20]

[19] See, e.g., Manski and Lerman (1977) for discussion of weighting for logistic regressions.

[20] With nearest neighbor matching, it does not matter whether matching is performed on the odds ratio or on the propensity scores (estimated using the wrong weights), because the ranking of the observations is the same and the same neighbors will be selected either way. Thus, failure to account for choice-based sampling will not affect nearest-neighbor point estimates. However, it will matter for kernel or local linear matching methods, because these methods take into account the absolute distance between the P observations.

3.2.4 When Does Bias Arise in Matching?

The success of a matching estimator depends on the availability of observable data to construct the conditioning set Z, such that (5) and (6) are satisfied. Suppose only a subset $Z_0 \subset Z$ of the variables required for matching is observed. The propensity score matching estimator based on Z_0 then converges to

$$\alpha_M' = E_{P(Z_0)|D=1}\left(E(Y_1|P(Z_0), D=1) - E(Y_0|P(Z_0), D=0)\right). \tag{7}$$

The bias for the parameter of interest, $E(Y_1 - Y_0|D=1)$, is

$$bias_M = E(Y_0|D=1) - E_{P(Z_0)|D=1}\{E(Y_0|P(Z_0), D=0)\}.$$

3.2.5 Using Balancing Tests for Check the Specification of the Propensity Score Model

As described earlier, the propensity score matching estimator requires that the outcome variable is mean independent of the treatment indicator conditional on the propensity score, $P(Z)$. An important consideration in implemention is how to choose which variables to include in estimating the propensity score. Unfortunately, there is no theoretical basis for how to choose a particular set Z to satisfy the identifying assumptions. Moreover, the set Z that satisfies the matching conditions is not necessarily the one the most inclusive one, as augmenting a set that satisfies the identification conditions for matching could lead to a violation of the conditions. Using more conditioning variables could also exacerbate a common support problem.

To guide in the selection of Z, there is some accumulated empirical evidence on how bias estimates of matching estimators depended on the choice of Z in particular applications. For example, Heckman Ichimura Smith and Todd (1998), Heckman Ichimura and Todd (1999) and Lechner (2001) show that which variables are included in the estimation of the propensity score can make a substantial difference to the performance of the estimator. These papers found that biases tended to be higher when cruder sets of conditioning variables where used.

These papers selected the set Z to maximize the percent of people correctly classified by treatment status under the model.

Also, they found that the matching estimators performed best when the treatment and control groups were were located in the same geographic area, so that regional effects on outcomes were the same across groups. Lastly, they studied the performance of matching estimators when a different survey instrument is used to collect the comparison group data from that used to collect the treatment group data.[21] They found that matching estimators generally performed poorly when the survey instrument is not the same. They conclude that matching estimators do not compensate for biases caused by differences in how variables are measured across difference surveys, a purpose for which they were not designed. The performance of matching method relies crucially on the data being of relatively high quality.

Rosenbaum and Rubin (1983) present a theorem that does not aid in choosing which variables to include in Z, but which can help in determining which interactions and higher order terms to include in the propensity score model for a given set of Z variables. The theorem states that

$$Z \perp\!\!\!\perp D | \Pr(D = 1|Z),$$

or equivalently

$$E(D|Z, \Pr(D = 1|Z)) = E(D| \Pr(D = 1|Z)).$$

The basic intuition is that after conditioning on $\Pr(D = 1|Z)$, additional conditioning on Z should not provide new information about D. Thus, if after conditioning on the estimated values of $P(D = 1|Z)$ there is still dependence on Z, this suggests misspecification in the model used to estimate $\Pr(D = 1|Z)$. Note that the theorem holds for any Z, including sets Z that do not satisfy the conditional independence condition required to justify matching. As such, the theorem is not informative about what set of variables to include in Z.

This result motivates a specification test for $\Pr(D = 1|Z)$. The general idea is to

[21] It is often the case in evaluation work that the comparison group data are collected using a different survey instrument. (See Lalonde (1986), and Smith and Todd (200?)).

test whether or not there are differences in Z between the $D = 1$ and $D = 0$ groups after conditioning on $P(Z)$. The test has been implemented in the literature a number of ways. Eichler and Lechner (2001) use a variant of a measure suggested in Rosenbaum and Rubin (1985) that is based on standardized differences between the treatment and matched comparison group samples in terms of means of each variable in Z, squares of each variable in Z and first-order interaction terms between each pair of variables in Z. An alternative approach used in Dehijia and Wahba (1999,2001) divides the observations into strata based on the estimated propensity scores. These strata are chosen so that there is not a statistically significant difference in the mean of the estimated propensity scores between the experimental and comparison group observations within each strata, though how the initial strata are chosen and how they are refined if statistically significant differences are found is not made precise. The problem of choosing the strata in implementing the balancing test is analogous to the problem of choosing the strata in implementing the interval matching estimator, described earlier. A common practice is to use five strata (e.g. quintiles of the propensity score). Within each stratum, t-tests are used to test for mean differences in each Z variable between the experimental and comparison group observations.

An alternative way of implementing the balancing test estimates a regression of each element of the set Z, Z_k on D interacted with a power series expansion in $P(Z)$:

$$\begin{aligned} Z_k &= \alpha + \beta_1 P(Z) + \beta_2 P(Z)^2 + \beta_3 P(Z)^3 + ... + \beta_j P(Z)^j + \\ &\quad \gamma_1 P(Z)D + \gamma_2 P(Z)^2 D + \gamma_3 P(Z)^3 D + ... + \gamma_j P(Z)^j D + \nu, \end{aligned}$$

and then tests whether the estimated γ coefficients are jointly insignificantly different from zero.

When significant differences are found for particular variables, higher order and interaction terms in those variables are added to the logistic model and the testing procedure is repeated, until such differences no longer emerge.

3.2.6 Assessing the Variability of Matching Estimators

Distribution theory for cross-sectional and difference-in-difference kernel and local linear matching estimators is derived in Heckman, Ichimura and Todd (1998). However, implementing the asymptotic standard error formulae can be cumbersome, so standard errors for matching estimators are often instead generating using bootstrap resampling methods.[22] A recent paper by Imbens and Abadie (2004a) shows that standard bootstrap resampling methods are not valid for asessing the varibility of nearest neighbor estimators, although their criticism does not apply for kernel or local linear matching estimators. Imbens and Abadie (2004b) present alternative standard error formulae for assessing the variability of nearest neighbor matching estimators.

3.2.7 Applications

Matching estimators have only recently been applied in evaluating the impacts of program interventions in developing countries. In one of the early applications, Jyotsna and Ravaillon (1999) use propensity score matching techniques to assess the impact of a workfare program in Argentina (the *Trabajar* program) on the wages of individuals who took part in the program. Their study finds sizable average wage gains due to the program. In another application, Jyotsna and Ravaillon (2003) use propensity score matching methods to study the effects of public investments in piped water in rural India on child health outcomes, where the matching estimators are used to control for nonrandomness in which households have access to piped water. Their study finds statistically significant impacts of having piped water on reducing the prevalence and duration of diarrhea among children under five.[23]

Matching methods were also used in the 2005 large-scale evaluation of the urban *Oportunidades* program in Mexico. The program is described in detail in chapter ? of this handbook. Briefly, the *Oportunidades* program provides monetary subsidies to families for

[22] See Efron and Tibshirani (1993) for an introduction to boostrap methods, and Horowitz (2001) for a recent survey of bootstrapping in econometrics.

[23] Upon more detailed examination of the distribution of treatment effects, however, Joytsna and Ravaillon (2003) also observe that the observed health gains largely bypass children from the poorest families, particularly those where the mother is poorly educated.

sending their children to school and for attending health clinics. The rural version of the program was evaluated using a place-based randomized experiment, which randomized a set of 506 villages in or out of the program. Because of high cost and out of ethical concerns, this type of randomization was deemed infeasible in high density urban areas. The alternative evaluation design adopted was a matched comparison group study. Matches for treatment group households were drawn from two data sources: families living in intervention areas who did not sign up for the program but who otherwise met the eligibility criteria, and families who met the eligibility criteria for the program but who were living in areas where the program was not yet available.[24] The propensity score model was estimated using data on program participants and nonparticipants living in intervention areas, and then used to impute propensity scores for the families living in nonintervention areas. The scores represent the probability that these families would participate in the program if it were offered to them. Program impact estimates were obtained using kernel and local linear regression matching estimators with bootstrapped standard errors. The analysis of children and youth age 6-20 indicated statistically significant program impacts on school enrollment, educational attainment, dropout rates, employment and earnings of youth, and on the numbers of hours spent doing homework.[25]

In another recent application of matching methods, Galiani, Gertler, and Schargrodsky (2005) analyze effects of privatization of water services on child mortality in Argentina. Variation of ownership in water provision over time provides a source of information that can be used to identify the effect of privatization, but which municipalities privatized first was nonrandom. To take into account unobserved municipality characteristics that may affect the decision to privatize, Galiani, Gertler, and Schargrodsky (2005) use a difference-in-difference kernel matching estimator. Their study finds that privatization of water services significantly reduced child mortality, especially in the poorest areas.

Behrman, Cheng and Todd (2004) develop a modified version of a propensity score match-

[24] To participate in the program, families had to attend sign-up modules during a time period when the modules were open.

[25] See Behrman, Garcia-Gallardo, Parker, and Todd, and Velez-Grajales (2005).

ing estimator that they use to evaluate the effects of a preschool program in Bolivia on child health and cognitive outcomes. Their approach identifies program effects by comparing children with different lengths of duration in the program. Instead of controlling for selectivity in program participation, as is usually done, their method controls for selectivity into alternative program participation durations, conditional on having chosen to participate. The estimator matches on the hazard rate and nonparametrically recovers the relationship between program duration and magnitude of treatment impact.

Other applications of matching methods in the economic development literature are Gertler, Levine and Ames (2004), in a study of the effects of parental death on child outcomes, Lavy (2004), in a study of the effects of a teacher incentive program in Israel on student performance, Angrist and Lavy (2001), in a study of the effects of teacher training on children's test scores in Israel, and Chen and Ravaillon (2003), a study of a poverty reduction project in China. There are numerous applications of matching estimators in the job training literature, many of which are discussed in Heckman, Lalonde and Smith (1999).

3.3 Control Function Methods

Another class of evaluation estimators are control function methods, which are also known as generalized residual methods. These methods were proposed as a solution to the evaluation problem in Heckman and Robb (1986).[26] Like the regression estimators discussed in section 3.1, they are usually defined within the context of an econometric model for the outcome process. Control function esitmators explicitly recognize that nonrandom selection into the program gives rise to an endogeneity problem in the model and try to obtain unbiased parameter estimates by modeling the source of the endogeneity. In contract, the matching estimators discussed in the previous section assume that selection on unobservables is not a problem after conditioning on a set of observed covariates.[27]

To see how the generalized residual method applies to the evaluation problem, write the

[26] The methods build on earlier selection bias correction methods developed in Heckman (1976, 1979, 1980).

[27] Or after taking differences in outcomes and conditioning on observed covariates, in the case of difference-in-difference matching.

model for outcomes as

$$Y = \varphi_0(X) + D\alpha^*(X) + \tilde{\varepsilon},$$

where

$$\alpha^*(X) = E(Y_1 - Y_0|X, D = 1) = \varphi_1(X) - \varphi_0(X) + E(U_1 - U_0|X, D = 1)$$

is the parameter of interest (TT(X)) and

$$\tilde{\varepsilon} = U_0 + D(U_1 - U_0 - E(U_1 - U_0|X, D = 1)).$$

Because the decision to participate may be endogenous with respect to the outcomes, we might expect that $E(U_0|X, D) \neq 0, i = 0, 1$. Heckman (1976,1979) showed that the endogeneity problem can be viewed as an error in model specification analogous to the problem of omitted variables. By adding and subtracting $E(U_0|X, D) = DE(U_0|D = 1, X) + (1-D)E(U|D = 0, X)$, we can rewrite the outcome model as

$$\begin{aligned} Y &= \varphi_0(X) + D\alpha^*(X) + E(U|D = 0, X) + \\ & D[E(U_0|D = 1, X) - E(U_0|D = 0, X)] + \varepsilon \\ &= \varphi_0(X) + D\alpha^*(X) + K_0(X) + D[K_1(X) - K_0(X)] + \varepsilon \end{aligned} \tag{8}$$

where

$$\begin{aligned} \varepsilon &= D\{U_0 - E(U_0|D = 1, X)\} + (1 - D)\{U_0 - E(U_0|D = 0, X)\} \\ + & D\{U_1 - E(U_1|D = 1, X)\} \end{aligned}$$

By construction, the residual ε has conditional mean equal to 0. The functions $K_1(X)$ and $K_0(X)$ are termed “control functions." When these functions are known up to some finite number of parameters, they can be included in the model to control for the endogeneity and regression methods (either linear or nonlinar) applied to consistently estimate program.

3.3.1 Econometric Methods for Estimating Control Functions

If no restrictions where placed on either $\alpha^*(X)$, $K_1(X)$, or $K_0(X)$, then the treatment impact parameter could not be separately identified from the control functions. Therefore, some identifying restrictions are necessary. Different implementations of control function estimators in the literature impose different types of restrictions. Usually, they consist of functional form restrictions and/or exclusion restrictions. In this context, exclusion restrictions are requirements that some variables that determine the participation process (i.e. the equation for D) be excluded from the outcome equation. These excluded variables generate variation in $K_1(X)$ and $K_0(X)$ that is independent from $\alpha^*(X)$. The following types of restrictions could be imposed: (1) functional form restrictions on $\alpha^*(X)$ and on $K_1(X)$ and $K_0(X)$ without exclusion restrictions; (2) Exclusion restrictions without functional form assumptions; (for example, if all the regressors in the outcome and participation equations were mutually exclusive and linearly independent) and (3) a combination of functional form and exclusion restrictions.[28]

Heckman and Robb (1986) motivate particular functional form restrictions on $K_d(X)$, $d \in \{0,1\}$, through an economic model of the participation process. Participation is assumed to depend on characteristics Z through an index $h(Z\gamma)$ and on unobservable characteristics V as follows:

$$D = \begin{cases} 1 \text{ if } h(Z\gamma) + V > 0 \\ 0 \text{ if } h(Z\gamma) + V \leq 0 \end{cases}.$$

In a random utility framework, $h(Z\gamma) + V$ represents the net utility from participating in a program. (McFadden, 1981, and Manski and McFadden, 1981).

Under this model, the function $K_0(X) = E(U_0|D=1,X)$ can be written as

$$\begin{aligned} E(U_0|D=1,X) &= E(U_0|h(Z\gamma)+V>0,X) \\ &= \frac{\int_{-h(Z\gamma)}^{\infty}\int_{-\infty}^{\infty} uf(u,v|X)dudv}{\int_{-h(Z\gamma)}^{\infty}\int_{-\infty}^{\infty} f(u,v|X)dudv}. \end{aligned}$$

If $F(U_0,V|X)$ is assumed to be continuous with full support in R^2 and $F_V(\cdot)$ is invertible,

[28] There is also another source of identification considered in the econometric evaluation literature called "identification at infinity." (See Heckman, 1979, Andrews and Schafgans, 1998). This type of identification is possible when there is a subgroup in the data for which $\Pr(D=1|Z)=1$ for some set Z, meaning that individuals with that set of characteristics always select into the program and there is no selection problem for them. This subgroup can be used to identify some of the model parameters that are not otherwise identified.

then the index $Z\gamma$ can be written as a function of the conditional probability of participation.

$$\begin{aligned}
\Pr(D=1|Z) &= \Pr(V > -h(Z;\gamma)) \\
&= 1 - F_V(-h(Z;\gamma)). \\
&\implies h(Z;\gamma) = -F_v^{-1}(-\Pr(D=1|Z))
\end{aligned}$$

Heckman and Robb (1986) note that if we make the additional assumption that the joint distribution of the unobservables, U_0 and V, does not depend on X, except possibly though the index, $h(Z;\gamma), i.e.$

$$f(U_0, V|X) = f(U_0, V|h(Z;\gamma)),$$

then $E(U_0|D=1,X)$ can be written as a function of the probability of participating in the program, $\Pr(D=1|Z)$, and D and no other variables:

$$\begin{aligned}
E(U_0|D &= 1, X) = E(U_0|D=1, P(Z)) = K_1(P(Z)) \\
E(U_0|D &= 0, X) = E(U_0|D=0, P(Z)) = K_0(P(Z)).^{29}
\end{aligned}$$

Assuming that a linear index is sufficient to represent the bias control function (so-called *index sufficiency*) greatly simplifies the problem of estimating the $K_d(X)$, $d \in \{0,1\}$ functions. It also helps the identification problem. For example, suppose $\varphi_0(X)$ and $h(Z\gamma)$ were both linear in the regressors. Then, under the index assumption, we can allow for overlap between X and Z, as long as there is at least one continuous variable included in Z excluded from X and no combination of X is a function of the Z.(Cosslett, 1984)

In the original formulation of the control function method in Heckman (1976, 1979), it was assumed that U_0 and V were jointly normal which implies a parametric form for $K_1(P(Z))$ and $K_0(P(Z))$. In Heckman, Ichimura, Smith and Todd (1996), the index sufficiency assumption is invoked and the $K(\cdot)$ functions are estimated nonparametrically as a function of the probability of participating in the program. [30]

[30] In that study, estimates of the probabilities of participating in the program (the propensity scores) are

3.3.2 A Comparison of Control Function and Matching Methods

Control function and matching methods were developed largely in separate literatures in econometrics and statistics, but the two methods are related and both make use of propensity scores in implementation. Conventional matching estimators can in some cases be viewed as a restricted form of a control function estimator. Recall that traditional matching methods assume that selection is on observables, whereas control function methods explicitly allow selection into programs to be on the basis of observables Z or on unobservables V. Assume the model for outcomes given in (1). The assumption that justifies matching outcomes on the basis of Z characteristics is

$$E(Y_0|D=1,Z) = E(Y_0|D=0,Z).$$

If $X \subset Z$, then, in the context of the outcomes model above, this assumption is implies that[31]

$$E(U_0|D=1,Z) = E(U_0|D=0,Z).$$

Under the control function approach, this assumption is equivalent to assuming that the control functions are equal for both the $D=0$ and $D=1$ groups

$$K_1(P(Z)) - K_0(P(Z)) = 0, \tag{9}$$

in which case the model for outcomes can be written as

$$Y_0 = \varphi_0(X) + D\alpha^*(X) + K_0(P(Z)) + D\{U_1 - U_0 - E(U_1 - U_0|D=1,X)\}.$$

In the literature, assumption (9) is referred as the special case of "selection on observables." (see Heckman and Robb, 1986; Heckman, Ichimura, Smith and Todd, 1995; and Barnow, Cain and Goldberger, 1980).

first obtained by a discrete choice model and then control functions are estimated nonparametrically using the predicted probabilities. That paper also develops a test for index sufficiency and finds that it cannot be rejected for a data sample of adult male applicants to the U.S. JTPA (Job Training and Partnership Act) program.

[31] See Heckman,Ichimura, Todd (1996a,b) for the more general case where Z contains variables not in X.

When selection is of this form, many of the identification problems that arise in trying to separate the treatment impact $\alpha^*(X)$ from the bias function $K_1(X)$ go away. That is, $\alpha^*(X)$ could be estimated without imposing functional form restrictions or exclusion restrictions. The functions $\varphi_0(X)$ and $K_0(P(Z))$ cannot be separately identified without additional restrictions, but if the goal of the estimation is to recover treatment impacts then there may be no need to separately identify these functions. As seen in the previous section, matching estimators recover $E(Y_0|D,X)$ directly with any attempt to separate the different components and without restrictions on the functional form of the conditional mean of the outcome equation.

In traditional implementations of the control function method, it is common to assume that (U_0, V) are joint normally distributed. Under the normal model, the restriction that $K_0(P(Z)) = K_1(P(Z))$ will, in general, not be satisifed unless the errors have zero covariance, $\sigma_{U_0V} = 0$. To see why that is the case, note that under joint normality

$$\begin{aligned} E(U_0|D &= 1, Z) = K_1(P(Z)) = \frac{\sigma_{U_0V}}{\sigma_{V^2}} \frac{\phi(-h(Z\gamma))}{1-\Phi(-h(Z\gamma))} \\ E(U_0|D &= 0, Z) = K_0(P(Z)) = \frac{\sigma_{U_0V}}{\sigma_{V^2}} \frac{-\phi(-h(Z\gamma))}{\Phi(-h(Z\gamma))}. \end{aligned}$$

Thus, $K_1(P(Z))$ equals $K_0(P(Z))$ only if $\sigma_{U_0V} = 0$.

3.3.3 Applications

Control function methods are not widely used in evaluating development programs. For discussion of their application in the context of evaluating job training programs, see Heckman, Lalonde and Smith (1999).

3.4 Instrumental Variables, Local Average Treatment Effects (LATE), and LIV Estimation

In this section, we consider the application of instrumental variables estimators for estimating program effects.

3.4.1 The Wald Estimator

We consider again the treatment effect model of the previous section:

$$Y = \varphi_0(X) + D\alpha^*(X) + \tilde{\varepsilon},$$

where

$$\alpha^*(X) = E(Y_1 - Y_0|X, D = 1) = \alpha(X) + E(U_1 - U_0|X, D = 1)$$

is the parameter of interest (TT) and

$$\tilde{\varepsilon} = U_0 + D(U_1 - U_0 - E(U_1 - U_0|X, D = 1)).$$

Suppose that there is an exclusion restriction, a variable Z that affects the program participation decision but does not enter into the outcome equation. Also, for simplicity of exposition, assume that the conditioning variables X and that the instrument Z is binary and takes on the values Z_1 and Z_2. We first partition the dataset by X and then use the instrument to estimate the program effect using the method of instrumental variables. The identifying assumption is that

$$E(U_0|X, Z) = E(U_0|X).$$

The so-called Wald estimator (applied within X strata) is given by

$$\begin{aligned}\hat{\alpha}^*_{IV}(X) &= \frac{\hat{E}(Y|Z = Z_1, X) - \hat{E}(Y|Z = Z_2, X)}{\hat{E}(D|Z = Z_1, X) - \hat{E}(D|Z = Z_2, X)} \\ &= \frac{\hat{E}(Y|Z = Z_1, X) - \hat{E}(Y|Z = Z_2, X)}{\widehat{\Pr}(D = 1|Z = Z_1, X) - \widehat{\Pr}(D = 1|Z = Z_2, X)},\end{aligned}$$

where the denominator is the difference in the probability of participating in the program under the two different values of the instrument. As noted in Heckman (1992), the estimator $\hat{\alpha}^*(X)$ recovers the effect of treatment on the treated, $\alpha^*_{IV}(X)$, only under one of two alternative assumptions on the error term:

Case I: $U_1 = U_0$

or

Case II: $U_1 \neq U_0$ and $E(U_1 - U_0|X, Z, D = 1) = E(U_1 - U_0|X, D = 1)$.

In the first case, the average impact of treatment on the treated (TT) is assumed to be the same as the average treated effect (ATE). Under the second case, the ATE and TT parameters differ, but the instrument does not forecast the idiosyncratic gain from the program. Heckman (1992) provides several examples where the assumption that the instrument does not help forecast the program gain may be problematic. Whether such an assumption is tenable or not will depend on the particular application at hand.

If assumptions I or II are not satisfied, then the Wald estimator can no longer be interpreted as estimating the average effect of treatment on the treated. Nonetheless, it has an alternative interpretation as a *Local Average Treatment Effect* (See Imbens and Angrist, 1994), which is the average effect of treatment for the subset of persons induced by the change in the instrument to receive the treatment. In the above example, the LATE estimator gives the average treatment impact for the subset of individuals who would not get treatment if $Z = Z_2$ but do get treatment if $Z = Z_1$. These are people who are induced to change their treatment status by the value of the instrument. The LATE parameter is further discussed below.

3.4.2 Marginal Treatment Effects (MTE) and Local Instrumental Variables (LIV) Estimation and its Relationship to TT, ATE, LATE

Recent advances in the program evaluation literature have led to a better understanding of the relationship between the TT, ATE and LATE parameters. Heckman and Vytlacil (2005) develop a unifying theory of how the different parameters relate to one another using a new concept, called a marginal treatment effect (MTE). Here, we summarize some key points of their argument. Consider the treatment effect model of the previous sections, written in slightly more general form, where there is again an outcome equation and a participation equation:

$$\begin{aligned}
Y_i &= D_i Y_{1i} + (1 - D_i) Y_{0i} \\
Y_{1i} &= \mu_1(X_i, U_{1i}) \\
Y_{0i} &= \mu_0(X_i, U_{0i}) \\
D_i &= 1 \text{ if } \mu_0(Z_i) - U_{D_i} \geq 0
\end{aligned}$$

It is assumed that $\mu_0(Z_i)$ is nondegenerate conditional on X_i, so that there is variation in who participates in the program holding X_i constant. The error terms are assumed to be independent of Z_i conditional on X_i.[32] As before, denote the propensity score as $P(Z) = \Pr(D = 1|Z = z) = F_{U_D}(\mu_0(Z_i))$ and assume that there is full support $(0 < \Pr(D = 1|Z) < 1)$. Heckman and Vytlacil (2005) show that without loss of generality, one can assume U_{D_i} distributed uniformly. Suppose that

$$D_i = 1 \text{ if } \ \varphi(Z_i) - v \geq 0$$

so that

$$\Pr(v < c) = F_V(c).$$

Since $F_V(\cdot)$ is a monotone transformation of the random variable v, we have

$$\Pr(F_V(v) < F_V(c)) = F_V(c).$$

Define $U_{D_i} = F_V(v)$. Because $\Pr(U_{D_i} < t) = t$, U_{D_i} is uniformly distributed between 0 and 1.

Next, note that when U_{D_i} is uniformly distributed,

$$E(D|Z) = \Pr(D = 1|Z) = F_{U_D}(\mu_0(Z_i)) = \mu_0(Z_i).$$

Let Z and Z' be two values of the instrument such that $\Pr(D = 1|Z) < \Pr(D = 1|Z')$. The threshold crossing model of program participation implies that some individuals who would

[32] See Heckman and Vytlacil (2005) for other technical conditions that are not central to the argument here.

have chosen $D = 0$ with $Z = Z$ will chose instead $D = 1$ when $Z = Z'$, but no individual with $D = 1$ when $Z = Z$ would choose $D = 0$ when $Z = Z'$.[33]

Using this framework, we can define different parameters of interest:

(i) The average treatment effect (ATE) is given by $\Delta^{ATE}(X) = E(\Delta|X = x)$

(ii) The average effect of treatment on the treated, conditional on a value of $P(Z)$, is given by $\Delta^{TT}(X, P(Z), D = 1) = E(\Delta|X = x, P(z) = P(Z), D = 1)$

(iii) The marginal treatment effect (MTE) conditions on a value of the unobservable: MTE $= \Delta_{MTE}(X) = E(\Delta|X = x, U_D = u)$

(iv) The local average treatment effect (LATE) parameter is given by

$$\text{LATE}{=}\Delta_{LATE}(X, P(Z), P(Z')) = \frac{E(Y|P(Z){=}P(Z),X){-}E(Y|P(Z){=}P(Z'),X)}{P(Z){-}P(Z')}.$$

The MTE is a new concept that Heckman and Vytlavil (2005) introduced. If $U_D = P(Z)$, then the index $\mu_0(Z_i) - U_{D_i} = 0$ (by the above reasoning, $\mu_0(Z_i) = P(Z)$ when U_{D_i} is uniformly distributed). People with the index equal to zero have unobservables that make them just indifferent between participating or not participating in the program. People with $U_{D_i} = 0$ have unobservables that make then most inclined to participate, while people with $U_{D_i} = 1$ have unobservables that make them the least inclined to participate.

Heckman and Vytlacil (2005) show that all the parameters of interest can be written in terms of the marginal treatment effect $\Delta_{MTE}(X)$ as follows:

$$\begin{aligned}
\Delta_{TT}(X) &= \frac{\int_0^{P(Z)} E(\Delta|X = x, U_D = u) dU_D}{P(Z)} \\
\Delta_{TT}(X) &= \int_0^1 E(\Delta|X = x, U_D = u) dU_D \\
\Delta_{LATE}(X, P(Z), P(Z')) &= \frac{\int_{P(Z')}^{P(Z)} E(\Delta|X = x, U_D = u) dU_D}{P(Z) - P(Z')}
\end{aligned}$$

That is, each of the parameters of interest can be written as an average of $\Delta_{MTE}(X)$ for values of U_D lying in different intervals. This result implies that knowledge of the MTE

[33] As shown in Vytlacil (2002), the assumptions required to justify a threshold crossing model are the same as the monotonocity conditions typically assumed to justify application of LATE estimators, proposed in Imbens and Angrist (1994).

function enables computation of all the other parameters of interest. However, the MTE function depends on a value of an unobservables, raising the question of how to estimate the MTE function. Heckman and Vyltacil (2005) propose an following estimation strategy that is implementable when the researcher has access to a continuous instrumental variable, Z, that enters into the participation equation but not the outcome equation. First, define a *local instrumental variables* estimator as

$$\begin{aligned}
\Delta_{LIV}(X, P(Z)) &= \frac{\partial E(Y|P(Z) = P(Z), X)}{\partial P(Z)} \\
&= \lim_{P(Z')->P(Z)} \frac{E(Y|P(z) = P(Z), x = X) - E(Y|P(z) = P(Z'), x = X)}{P(Z) - P(Z')} \\
&= MTE(X, U_D = P(Z)).
\end{aligned}$$

The $\Delta_{LIV}(X, P(Z))$ parameter can be obtained by (i) estimating the program participation (propensity score) model to get $\hat{P}(Z)$, and then (ii) estimating $\frac{\partial E(Y|P(Z)=P(Z),X)}{\partial P(Z)}$ nonparametrically (which can be done by local linear regression). Step (i) can be carried out using a parametric, semiparametric or nonparametric estimator for the binary choice model. Step (ii) can be performed by local linear regression of the outcome ($Y = DY_1 + (1-D)Y_0$) on the estimated $\hat{P}(Z)$. Evaluating this function for different values of $P(Z)$ traces out the MTE function. The different estimands TT, ATE, LATE can be obtained by integrating under the MTE function.

3.4.3 Applications

LIV estimators have only been recently developed, and there are thus far no applications to evaluating effects of program interventions in developing country settings. For a recent application to estimating returns to education using U.S. data, see Carniero, Heckman and Vytlacil (2001).

3.5 Regression-Discontinuity Methods

Sometimes, in evaluating effects of a program intervention, there is information available on the rule generating assignment of individuals into treatment. For example, suppose that individuals who apply to the program are assigned a program eligibility score (based on their characteristics) and that only individuals with a score below a threshold are allowed to enter the program. This type of data design was first considered in Thistlethwaite and Campbell (1960) in an application in which they estimated the effect of receiving a National Merit Scholarship Award has on students' success in obtaining additional college scholarships and on their career aspirations. They observed that the awards are given on the basis of whether a test score exceeds a threshold, so one can take advantage of knowing the cut-off point to learn about treatment effects for persons near the cut-off. [34] The defining characteristic of regression discontinuity (RD) data designs is that the treatment variable changes discontinuously as a function of one or more underlying variables.

In the evaluation literature, there are several papers considering identificiation of treatment effects under a RD data design along with different kinds of assumptions on the processing governing the outcome variables and on the distribution of treatment effects. Trochim (1984) discusses alternative parametric and semiparametric RD estimators that have been proposed in the statistics literature. Van der Klaauw (1996) considers identification and estimation in a semiparametric model under a constant treatment effect assumption. Hahn, Todd and Van der Klaauw (2000) consider a more general case that allows for variable treatment effects and that imposes weak assumptions on the distribution (or conditional mean function) of the outcome variables. The discussion below follows along the lines of the Hahn et. al. (2000).

Suppose that the goal of the evaluation is to determine the effect that some binary treatment variable D_i has on an outcome Y_i. The model for the observed outcome can be written as

$$Y_i = Y_{0i} + D_i \cdot \Delta_i, \tag{10}$$

If the data are purely observational (or nonexperimental), then little may be known a priori about the process by which individuals are selected into treatment. With data from a RD

[34] Other applications of the regression-discontinuity methods include Berk and Rauma (1983), Van der Klaauw (1996), Angrist and Lavy (1996), and Black (1996).

design, the analyst has some information about the treatment assignment mechanism.

There are two main types of discontinuity designs considered in the literature - the *sharp design* and the so-called *fuzzy design* (see *e.g.* Trochim, 1984). With a sharp design, treatment D_i is known to depend in a deterministic way on some observable variable Z_i, $D_i = f(Z_i)$, where Z takes on a continuum of values and the point z_0 where the function $f(Z)$ is discontinuous is assumed to be known.

With a fuzzy design, D_i is a random variable given Z_i, but the conditional probability $f(Z) \equiv E[D_i| Z_i = z] = \Pr[D_i = 1| Z_i = z]$ is known to be discontinuous at z_0.[35] Next we consider formally why knowing that the probability of receiving treatment changes discontinuously as a function of an underlying variable is a valuable source of identifying information.

3.5.1 Identification of Treatment Effects under Sharp and Fuzzy Data Designs

Sharp Design To simplify the exposition, consider the special case of a sharp discontinuity design. Treatment is assigned based on whether Z_i crosses a threshold value z_0:

$$\begin{aligned} D_i &= 1 \text{ if } Z_i > z_0 \\ &= 0 \text{ if } Z_i \leq z_0. \end{aligned}$$

As z may be correlated with the outcome variable, the assignment mechanism is clearly not random and a comparison of outcomes between persons who received and did not receive treatment will generally be a biased estimator of treatment impacts. However, we may have reason to believe that persons close to the threshold z_0 are similar. If so, we may view the design as almost experimental near z_0.

To make ideas concrete, let $e > 0$ denote an arbitrary small number. Comparing conditional means for persons who received and did not receive treatment gives

$$\begin{aligned} E[Y_i| Z_i = z_0 + e] - E[Y_i| Z_i = z_0 - e] &= E[\Delta_i| Z_i = z_0 + e] \\ &\quad + E[Y_{0i}| Z_i = z_0 + e] - E[Y_{0i}| Z_i = z_0 - e]. \end{aligned}$$

[35] For example, in the application of Van der Klaauw (1996), the probability that a student receives financial aid changes discontinuously as a function of a known index of the student's GPA and SAT scores. However, there are other factors, some of which are unobserved, which affect the financial aid decision, so the data fits a fuzzy rather than a sharp design.

When persons near the threshold are similar, we would expect $E\left[Y_{0i}\right|Z_i = z_0 + e] \cong E\left[Y_{0i}\right|Z_i = z_0 - e]$. This intuition motivates the following assumptions:

RD-1: $E\left[Y_{0i}\right|z_i = z]$ is continuous in Z at z_0.[36]

RD-2 The limit $\lim_{e\to 0^+}\left[\Delta_i\right|Z_i = z_0 + e]$ is well defined.

Under Conditions (RD-1) and (RD-2), it is easy to see that

$$\lim_{e\to 0^+}\left\{E\left[Y_i\right|Z_i = z_0 + e] - E\left[Y_i\right|Z_i = z_0 - e]\right\} = E\left[\Delta_i\right|z_0]\,. \tag{11}$$

By comparing persons arbitrarily close to the point z_0 who did and did not receive treatment, we can in the limit identify $E\left[\Delta_i\right|z_i = z_0]$, which is the average treatment effect for people with values of Z_i at the point of discontinuity z_0. Conditions (RD-1) and (RD-2) are all that is required for identification.

It is a limitation of a RD design that we can only learn about treatment effects for persons with z values near the point of discontinuity. Sometimes, however, the treatment effects near the boundary are of particular interest, for example, if the policy change being considered were that of expanding the cut-off.

Fuzzy Design The fuzzy design differs from the sharp design in that the treatment assignment is not a deterministic function of z_i, because there are additional unobserved variables that determine assignment to treatment. The common feature it shares with the sharp design is that the probability of receiving treatment (the propensity score), $\Pr\left[D_i = 1\right|Z_i]$, viewed as a function of z_i, is discontinuous at z_0. As shown in Hahn, Todd and Van der Klaauw (2000), mean treatment effects can be identified even under a fuzzy design under different some assumptions on the heterogeneity of impacts.

Common Treatment Effects Suppose that the treatment effect is constant across different individuals and is equal to Δ.The mean difference in outcomes for persons above and below the discontinuity point z_0 is

$$\Delta\cdot\left\{E\left[D_i\right|Z_i = z_0 + e] - E\left[D_i\right|Z_i = z_0 - e]\right\} + E\left[Y_{0i}\right|Z_i = z_0 + e] - E\left[Y_{0i}\right|Z_i = z_0 - e]\,.$$

[36] It is assumed that the density of Z_i is positive in the neighborhood containing z_0.

Under (RD-1), we have

$$\lim_{e\to 0^+} E\left[Y_i \middle| Z_i = z_0 + e\right] - E\left[Y_i \middle| Z_i = z_0 - e\right] = \Delta \cdot \lim_{e\to 0^+} \left\{E\left[D_i \middle| Z_i = z_0 + e\right] - E\left[D_i \middle| Z_i = z_0 - e\right]\right\}.$$

Thus, we can identify Δ by the ratio

$$\frac{\lim_{e\to 0^+} E\left[y_i \middle| z_i = z_0 + e\right] - \lim_{e\to 0^+} E\left[y_i \middle| z_i = z_0 - e\right]}{\lim_{e\to 0^+} E\left[x_i \middle| z_i = z_0 + e\right] - \lim_{e\to 0^+} E\left[x_i \middle| z_i = z_0 - e\right]}. \tag{12}$$

The denominator is nonzero because the fuzzy RD design guarantees that $\Pr\left[D_i = 1 \middle| z_i = z\right]$ (the propensity score) is discontinuous at z_0.

Variable Treatment Effects Now suppose treatment effects are heterogeneous, and in addition to assumptions (RD-1) and (RD-2), we assume

RD-3: D_i is independent of Δ_i conditional on Z_i near z_0: $D_i \perp \Delta_i | \, Z_i = z_0$

Then the same ratio identifies $E(\Delta_i | Z_i = z_0)$. In addition to the cases considered above, Hahn et. al. (2000) also consider an alternative local average treatment effect (LATE) interpretation of the same ratio.[37]

3.5.2 Estimation

We next describe an estimation approach proposed in Hahn et. al. (2000).[38] For both the sharp design and fuzzy design, (12) identifies the treatment effect at $z = z_0$. Thus, given consistent estimators of the four one-sided limits in (12), the treatment effect can be consistently estimated. One simple nonparametric estimator would estimate the limits by

[37] Extending the idea of Imbens and Angrist (1994) or Angrist, Imbens, and Rubin (1994) to the RD design, the ratio gives the average impact for people induced to receive treatment by whether the instrument is above or below the cut-off z_0.

[38] One estimation approach proposed by van der Klaauw (1996) for the sharp design is to assume (in addition to continuity) a flexible parametric specification for $g(Z) = E\left[Y_{0i} \middle| z_i\right]$ and add this as a 'control function' to the regression of Y_i on D_i. For the fuzzy design he proposes a similar approach but where D_i in the control function-augmented regression equation is now replaced by a first stage estimate of $E\left[D_i \middle| Z_i\right]$. This estimation approach is consistent under correct specification but can be sensitive to misspecification.

averages over the Y_i values and the D_i values within a specified distance of the boundary points (the bandwidth). Let $\hat{\Delta}$ denote an estimator for the treatment impact

$$\hat{\Delta} = \frac{\hat{y}^{+} - \hat{y}^{-}}{\hat{x}^{+} - \hat{x}^{-}},$$

where $\hat{y}^{+}, \hat{y}^{-}, \hat{x}^{+}$, and $\hat{x}^{-}$ are estimators for each of the limit expressions. Given appropriate bandwidths h_{+} and h_{-}, we would estimate the limits by

$$\hat{y}^{+} = \frac{\sum_i Y_i \cdot 1\left(z_0 < Z_i < z_0 + h_{+}\right)}{\sum_i 1\left(z_0 < Z_i < z_0 + h_{+}\right)}, \quad \hat{y}^{-} = \frac{\sum_i Y_i \cdot 1\left(z_0 - h_{-} < Z_i < z_0\right)}{\sum_i 1\left(z_0 - h_{-} < Z_i < z_0\right)},$$

and

$$\hat{x}^{+} = \frac{\sum_i D_i \cdot 1\left(z_0 < Z_i < z_0 + h_{+}\right)}{\sum_i 1\left(z_0 < Z_i < z_0 + h_{+}\right)}, \quad \hat{x}^{-} = \frac{\sum_i D_i \cdot 1\left(z_0 - h_{-} < Z_i < z_0\right)}{\sum_i 1\left(z_0 - h_{-} < Z_i < z_0\right)}.$$

The RD estimator can also be implemented using local linear regression methods, as proposed in Hahn et. al. (2000), which have have better performance than simple averaging methods or kernel methods at boundary points.(See Fan, 1992)[39] For this problem, all the estimation points are boundary points.

3.5.3 Applications of RD Methods

Regression-discontinuity methods have only rarely been used in the evaluation of social programs in developing country settings. Buddelmeier and Skoufias (2004) study the performance of RD methods using data from the Mexican PROGRESA experiment.[40] As discussed in section 3.2, the PROGRESA program was a school and health subsidy program introduced by the Mexican government in rural areas. The experiment randomized villages in and out of the program. Within each village, only families who were eligible for the program according to an eligibility index were allowed to participate in it, where the index was derived from poverty criteria, such as whether the family had a dirt floor or a bathroom in their home. Most families deemed eligible for the program decided to participate in it to some extent.

[39] Boundary points are points within one bandwidth of the boundary. See Härdle (1990) or Härdle and Linton (1994) for discussion of the boundary bias problem.

[40] This experiment is described in detail in Chapter ? of this handbook.

Buddelmeier and Skoufias (2004) observe that families with eligibility index values just above the cut-off who received the program are highly similar to families with eligible values just below the cut-off. The criteria for eligibility were not made public, which alleviates concerns that households could have manipulated their poverty status to become eligible for the program.[41] Using a RD estimation approach, Buddlemeier and Skoufias (2004) calculate program impacts for the households near the eligibility cut-off by comparing households living in treated communities with scores just above and below the cut-off. Their results show that the estimates based on the RD method are close to those observed from the experiment, lending credibility to the RD approach. Moreover, most of the households in their sample have scores near the cut-off values, making the sample near the cut-off an interesting subsample to study.

In another application, Lavy (2004) uses an RD estimator to evaluate the effects of a teacher incentive program on student performance. The program introduced a rank-order tournament (among teachers of English, Hebrew, and mathematics in Israel) that rewarded teachers with cash bonuses for improving their students' performance on high-school matriculation exams. The regression discontinuity method of Lavy (2004) exploits both a natural experiment stemming from measurement error in the assignment variable and a sharp discontinuity in the assignment-to-treatment variable. The results show that performance incentives significantly affect students in the treatment group, with some minor spillover effects on untreated subjects. A recent study by Chay, McEwan, and Urquiola (2005) also evaluates the effects of an incentive program using a RD design. The program is a school resource program in Chile that awards resources to schools based on cut-offs in the school's test scores. Their results indicate that the program had statistically significant effects on test score gains.

[41] If households were selecting nonrandomly into the program around the cut-off, then this could invalidate the assumption RD-1.

4 Ex Ante Program Evaluation

Thus far, we have considered the problem of evaluating effects of existing programs. All of the evaluation methods described in the previous sections require access to data on program participants, which would typically be available in an ex post evaluation. However, policy makers are sometimes interested in evaluating effects of possibly a range of programs before deciding which type of program to implement. For example, the goal may be to (i) to optimally design a social program to achieve some desired effects, (ii) to forecast the take-up rates of alternative programs, or (iii) to study the effectiveness of alternatives to an existing program. Evaluating effects of programs that do not yet exist requires an evaluation method that makes use only of data on people who have not participated in the program. Answering question (iii) requires a way of extrapolating from experience with an existing program to a range of alternative programs.

The problem of forecasting the effects of social programs is part of the more general problem of studying the effects of policy changes prior to their implementation that was studied by Marshak (1953). He described it as one of the most challenging problems facing empirical economists. In the early discrete choice literature, the problem appeared as the "forecast problem," whereby researchers were trying to predict the demand for a new good prior to its being introduced into the choice set. For example, McFadden (1977) used a discrete choice random utility model to forecast the demand for the San Francisco BART subway system prior to its being built.

There are a few empirical studies that study the performance of economic models in forecasting program effects by comparing models' forecasts of treatment effects to those estimated experimentally. For example, Wise (1985) develops and estimates a model of housing demand that he uses to forecast the effects of a housing subsidy. The housing subsidy program was implemented as a randomized experiment, so he is able to compare forecasts he obtains from alternative models models to the experimental subsidy effects.

In a more recent application, Todd and Wolpin (2004) develop and estimate a dynamic behavioral model of family decision making about child schooling and fertility that they use to forecast the effects the PROGRESA program (see section 3.4) on choices about children's schooling and work and on family fertility.[42] Todd and Wolpin (2004) compare the model's predictions about program impacts to those observed under the randomized experiment. They find that the model provides relatively accurate forecasts of program effects on school enrollment and child work patterns. They then use the model to evaluate the take-up rates, costs and program effects of a variety of counterfactual programs, such as changes to the subsidy schedule (how the subsidy varies by gender and grade). Lastly, they use the model to evaluate effects of some radically different programs, such as an income subsidy program that removes the school attendance requirement.

To illustrate how a behavioral model can be used to predict the impacts of a program that has not been implemented, we next describe a simple model of schooling choice that shows how the effects of a school subsidy program can potentially be identified even when none of the families in the data receives a subsidy. This example generalizes an example in Todd and Wolpin (2004).

4.1 An Illustrative Model of Identification of Subsidy Effects

Consider a household making a single period decision about whether to send a single child to school or to work. Let the utility of the household be separable in consumption (C) and school attendance (s), namely $u = C + (\alpha + \varepsilon)s$, where $s = 1$ if the child attends school, $= 0$ otherwise and ε is a preference shock. Assume that the cost of attending schoo depends on distance to the school, denoted k. Children who work contribute to family income, so the family's income is $y + w(1 - s) - \delta ks$, where y is parent's income, w is the child's earnings, and δks is the distance cost that is only incurred if the child attends school. Under utility maximization, the family chooses to have the child attend school if $\varepsilon > w - \alpha + \sigma k$.

Suppose that wages are only onserved for children who work and specify a child wage

[42] The PROGRESA program is described in detail in chapter ? of this handbook.

offer equation:

$$w = z\gamma + v$$

where z are characteristics (such as age or sex) that are determinants of wage offers and are observed for all children. The equation governing whether a family sends a child to school or work is

$$s = 1 \text{ if } \alpha - \delta k + \varepsilon > z\gamma + v, \text{ else } s = 0.$$

The probability that a child attends school can be written as

$$\begin{aligned} \Pr(s &= 1|z) = \Pr(z\gamma - \alpha + \delta k < \varepsilon - \nu) \\ &= F_{\varepsilon - v}(z\gamma - \alpha + \delta k), \end{aligned}$$

where $F_{\varepsilon-v}(\cdot)$ is the cdf of $\varepsilon-\nu$. Under an assumption that the median of $\varepsilon-\nu$ is 0 conditional on z, the parameters γ, α and δ can be estimated up to scale by either a parametric or semiparametric discrete choice estimation method.[43]

Next, consider estimation of the child wage offer equation only using data on children who work ($s = 0$) for whom wages are observed. We can write the wage equation as

$$y = z\gamma + E(\nu|z, s = 0) + \{\nu - E(\nu|z, s = 0)\},$$

where the error term in brackets ($\eta = \nu - E(\nu|z, s = 0)$) has conditional mean zero by construction.

As described in section 3.3, we can consistently estimate the parameter γ by including a control function to capture $E(\nu|z, s = 0)$.[44] Under the assumption that (i) v and ε are jointly distributed with density $f(\nu, \varepsilon)$ and (ii) the conditional density equals the unconditional density, $f(v, \varepsilon|z, k) = f(\nu, \varepsilon)$, as described along in section 3.3, we obtain

$$w = z\gamma + K(P) + \eta,$$

where P is the probability of working. If there is a continuous exclusion restriction that affects the work decision but not the wage offer equation (in this case, the distance variable

[43] See Manski, 1988.

[44] Also, see Heckman (1980).

k), then the parameter γ can be nonparametrically identified under very weak assumptions on the K function.[45] To see why, note that an exclusion restriction allows us to hold constant z at some value and vary P, thereby tracing out the K function. Then, fixing K at some value, we can estimate γ.[46] Once γ is identied, we can use the results of the discrete choice estimation to obtain α and δ.[47]

Next, we examine how the estimated model can be used for ex-ante evaluation. Suppose that the government is contemplating a program to increase school attendance of children though the introduction of a subsidy to parents in the amount b if they send their child to school. Under such a program, the probability that a child attends school will increase by $F_{\varepsilon-v}(z\gamma-\alpha-b+\delta k)-F_{\varepsilon-v}(z\gamma-\alpha+\delta k)$. The function $F_{\varepsilon-v}(s)$ can be estimated nonparametrically by a nonparametric regression of the school attendance indicator, s, on $z\hat{\gamma}-\hat{\alpha}+\hat{\delta}k$.[48] To assess the effect of the subsidy on the probability of attending school, we simply evaluate the $F_{\varepsilon-v}(s)$ function at the point $z\hat{\gamma}-\hat{\alpha}+\hat{\delta}k$.

5 Conclusions

A common problem in evaluating the effects of program interventions using nonexperimental data is that the program recipients may differ in systematic ways from nonrecipients. Such differences may arise either because the programs are nonrandomly placed, are targeted at certain groups of people, or because individuals self-selected into the program. Of course, all three of these factors may occur simultaneously, posing challenges to the evaluator.

This chapter has reviewed a range of estimation methods developed in the evaluation literature for evaluating the impact of program interventions in these kinds of situations. The goal of the chapter was to identify different parameters of interest in an evaluation,

[45] Only weak assumptions on the continuity of the K function are required.

[46] The intercept of the wage offer equation will, in general, not be separately identified from the K function unless there is a subset of the data for which $\Pr(s=0|z,k)=1$. On this point, known in the literature as identification at infinity, see Heckman (1980) and Andrews and Schafgans (1998).

[47] Given an estimate of γ, the scaling factor in the discrete choice problem can also be obtained.

[48] Here, we use the fact that the conditional expectation of s, $E(s|z\hat{\gamma}-\hat{\alpha}+\hat{\delta}k=\tau)=\Pr(s=1|z\hat{\gamma}-\hat{\alpha}+\hat{\delta}k=\tau)$.

illustrate different methods for estimating them, review their identifying assumptions and describe how different methods relate to one another.

Each of the econometric evaluation estimators discussed in this chapter invokes a different set of assumptions to justify its application. The question of which method to adopt in any particular circumstance will be context-specific and will also depend on the quality of data available. For example, matching methods impose weak assumptions on the conditional mean of the outcome equation, but they make the strong assumptions that which unit receives treatment is ignorable after conditioning on a set of observed covariates. Such a method should only be adopted only in situations where the available conditioning variables are rich enough to make the required assumption plausible. Control function estimators are the most general class of estimators, in that they explicitly allow possibly time-varying unobservables to affect program participation decisions. The implementation of parametric control function estimators are straightforward, but a drawback to them is that they typically assume that error terms are normally distributed. Semiparametric control function estimators provide a more flexible alternative, but they often require additional assumptions to achieve identification of parameters of interest.

Regression-discontinuity estimators can be applied in situations where there is a known discontinuity in the treatment assignment rule as a function of some underlying variable, such as a score determining who is eligible for the treatment. These estimators can be justified under weak assumptions, but they usually only provide information on treatment effects at the points of discontinuity.

Lastly, the evaluator has access to a class of instrumental variables estimators, that can be applied In situations where there is a variable affecting the program participation decision but not affecting the outcome. When there is a valid exclusion restriction, one option is simply to apply the Wald IV estimator, which recovers the local average treatment effect (LATE) parameter. The newest of the evaluation methods considered in this chapter are Local Instrumental Variable (LIV) estimators that can be applied when there is a continuous instrumental variable. LIV estimators provide a means of learning about the distribution of treatment effects and can be used to generate other parameters of interest, including LATE, treatment on the treated (TT), average treatment effect (ATE).

References

[1] Abadie, Alberto and Guido Imbens (2004a): "On the Failure of the Bootstrap for Matching Estimators," manuscript, Harvard University.

[2] Abadie, Alberto and Guido Imbens (2004b): "Large Sample Properties of Matching Estimators for Average Treatment Effects," manuscript, Harvard University.

[3] Andrews, Donald and Schafgans, (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65, 497-518.

[4] Angrist, J., G. Imbens and D. Rubin (1994): "Identification of Causal Effects using Instrumental Variables," *Journal of the American Statistical Association,*

[5] Angrist, J. and Lavy, V. (1999): "Using Maimonides Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quarterly Journal of Economics*, May.

[6] Angrist, J. and Lavy, V. (2001): "Does Teacher Training Affect Pupil Learning? Evidence from Matched Comparisons in Jerusalem Public Schools," *Journal of Labor Economics*, 19(2), 343-369.

[7] Ashenfelter, Orley (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60, 47-57.

[8] Ashenfelter, Orley and David Card (1985): "Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs," *Review of Economics and Statistics*, 67, 648-660.

[9] Barnow, B., G. Cain, and A. Goldberger (1980): "Issues in the Analysis of Selectivity Bias," in Ernst Stromsdorfer and George Farkas, eds., *Evaluation Studies Review Annual Volume 5* (San Fransisco: Sage), 290-317.

[10] Bassi, Lauri (1984): “Estimating the Effects of Training Programs with Nonrandom Selection, ” *Review of Economics and Statistics*, 66, 36-43.

[11] Behrman, Jere, Cheng, Yingmei and Petra Todd (2000): "Evaluating Preschool Programs when Length of Exposure to the Program Varies: A Nonparametric Approach," *Review of Economics and Statistics,* 86(1), 108-132.

[12] Behrman, Jere, Jorge Garcia-Gallardo, Susan Parker, Petra Todd, and Viviana Velez-Grajales (2005): "How Conditional Cash Transfers Impact School and Working Behavior of Children and Youth in Urban Mexico," manuscript.

[13] Buddelmeyer, Hielke and Emmanuel Skoufias (2004): "An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA," Policy Research Working Paper Series 3386, The World Bank.

[14] Carniero, P., Heckman, J. J. and Vyltacil, E (2001): "Understanding What Instrumental Variables Estimate: Estimating Marginal and Average Returns to Education," manuscript, University of Chicago.

[15] Chay, Kenneth Y., Patrick J. McEqan and Miquel Urquiola (2005): "The Central Role of Noise in Evaluating Interventions that use Test Scores to Rank Schools," American Economic Review

[16] Chen, Shaohua and Martin Ravallion, 2003. "Hidden Impact? Ex-Post Evaluation of an Anti-Poverty Program," Policy Research Working Paper Series 3049, The World Bank.

[17] Cochran, W. and Donald Rubin (1973): “Controlling Bias in Observational Studies,” *Sankyha*, 35, 417-446.

[18] Dehejia, Rajeev and Sadek Wahba (1998): “Propensity Score Matching Methods for Nonexperimental Causal Studies,” NBER Working Paper No. 6829.

[19] Dehejia, Rajeev and Sadek Wahba (1999): “Causal Effects in Noexperimental Studies: Reevaluating the Evaluation of Training Programs,” *Journal of the American Statistical Association*, 94(448), 1053-1062.

[20] Duflo, Esther (2000): "Child Health and Household Resources in South Africa: Evidence from Old Age Pension," *AEA Papers and Proceedings*, 90(2), 393-398.

[21] Duflo, Esther (2001): “Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment," *American Economic Review.*

[22] Eichler, Martin and Michael Lechner (2001): “An Evaluation of Public Employment Programmes in the East German State of Sachsen-Anhalt,” *Labour Economics*, forthcoming.

[23] Fan, J. (1992a): “Design Adaptive Nonparametric Regression, ” *Journal of the American Statistical Association,* 87, 998-1004.

[24] Efron, Bradley and Robert Tibshirani (1993): *An Introduction to the Bootstrap*, Chapman and Hall, New York: New York.

[25] Fan, J. (1992b): : “Local Linear Regression Smoothers and their Minimax Efficiencies,” *The Annals of Statistics,* 21, 196-216.

[26] Fan, J. and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*. New York: Chapman and Hall.

[27] Galiani, Sebastian, Gertler, Paul, and Ernesto Schargrodsky "Water for Life: The Impact of the Privatization of Water Services on Child Mortality in Argentina," *Journal of Political Economy.*

[28] Gertler, Paul, Levine, David and Minnie Ames (2004): "Schooling and Parental Death," *Review of Economics and Statistics, 86(1).*

[29] Gertler, Paul and Molyneaux (1994): “How Economic Develop and Family Planning Programs Combined to Reduce Indonesian Fertility," *Demography*, 31(1), 33-63.

[30] Glewwe, Paul, Kremer, Michael, Moulin, Sylvie, and Eric Zitzewitz (2004): "Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya," *Journal of Development Economics*, 74, 251-268.

[31] Glewwe, Paul, Kremer, Michael, Moulin, Sylvie (2000): "Textbooks and Test Scores," Evidence from a Prospective Evaluation in Kenya," manuscript.

[32] Glewwe, Paul, Kremer, Michael, Moulin, Sylvie (2003): "Teacher Incentives," NBER Working Paper #9671.

[33] Hahn, Jinyong (1998): “On the Role of the Propensity Score in Efficient Estimation of Average Treatment Effects,” *Econometrica*, 66(2), 315-331.

[34] Hahn, Jinyong, Todd, Petra and Wilbert Van der Klauww, “Identification of Treatment Effects by Regression-Discontinuity Design”,*Econometrica*, February, 2001, pp. 201-209.

[35] Härdle, W. (1990), Applied Nonparametric Regression. New York: Cambridge University Press.

[36] Härdle, W. and Linton, O. (1994), “Applied Nonparametric Methods,” in *Handbook of Econometrics, Volume 4,* ed. by D.F. McFadden and R.F. Engle. Amsterdam: North Holland, 2295 - 2339.

[37] Heckman, James (1979): “Sample Selection Bias as a Specification Error,” *Econometrica*, 47(1), 153-161.

[38] Heckman, J. (1990): “Varieties of Selection Bias,” *American Economic Review,* 80, 313-318.

[39] Heckman, James (1992): "Randomization and Social Policy Evaluation," in Charles Manski and Irwin Garfinkle, eds., *Evaluating Welfare and Training Programs* (Cambridge, Mass.: Harvard University Press), 201-230.

[40] Heckman, James (1997): "Randomization as an Instrumental Variables Estimator: A Study of Implicit Behavioral Assumptions in One Widely-used Estimator," *Journal of Human Resources*, 32, 442-462.

[41] Heckman, James, Neil Hohmann and Jeffrey Smith, with Michael Khoo (2000): "Substitution and Drop Out Bias in Social Experiments: A Study of an Influential Social Experiment," *Quarterly Journal of Economics*, 115(2), 651-694.

[42] Heckman, James and Joseph Hotz (1989): "Choosing Among Alternative Nonexperimental Methods for Estimating the Impact of Social Programs: The Case of Manpower Training", *Journal of the American Statistical Association*, 84 (408), 862-880.

[43] Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1996): "Sources of Selection Bias in Evaluating Social Programs: An Interpretation of Conventional Measures and Evdience on the Effectiveness of Matching as a Program Evaluation Method," *Proceedings of the National Academy of Sciences*, 93(23), 13416-13420.

[44] Heckman, James, Hidehiko Ichimura, Jeffrey Smith and Petra Todd (1998): "Characterizing Selection Bias Using Experimental Data," *Econometrica* , 66(5), 1017-1098.

[45] Heckman, James, Hidehiko Ichimura and Petra Todd (1997): "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program," *Review of Economic Studies,* 64(4), 605-654.

[46] Heckman, James, Hidehiko Ichimura and Petra Todd (1998), "Matching As An Econometric Evaluation Estimator," *Review of Economic Studies,* 65(2), 261-294.

[47] Heckman, James, Robert Lalonde and Jeffrey Smith (1999): "The Economics and Econometrics of Active Labor Market Programs" in Orley Ashenfelter and David Card,

eds., *Handbook of Labor Economics Volume 3A* (Amsterdam: North-Holland), 1865-2097.

[48] Heckman, James and Richard Robb (1985): “Alternative Methods for Evaluating the Impact of Interventions,” in James Heckman and Burton Singer, eds., *Longitudinal Analysis of Labor Market Data (*Cambridge, England: Cambridge University), 156-246.

[49] Heckman, James and Jeffrey Smith (1995): “Assessing the Case the Randomized Social Experiments,” *The Journal of Economic Perspectives,* 9, 85-110.

[50] Heckman, James and Jeffrey Smith (1999): “The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies,” *Economic Journal*, 109(457), 313-348.

[51] Heckman, James and Jeffrey Smith, with Nancy Clements (1997): “Making the Most Out of Social Experiments: Accounting for Heterogeneity in Programme Impacts,” *Review of Economic Studies*, 64(4), 487-536.

[52] Heckman, James and Petra Todd (1995): “Adapting Propensity Score Matching and Selection Models to Choice-based Samples,” manuscript, University of Chicago.

[53] Heckman, James and Edward Vytlacil (2000): “Causal Parameters, Structural Equations, Treatment Effects and Randomized Evaluations of Social Programs,” manuscript, University of Chicago.

[54] Heckman, James and Edward Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*.

[55] Heckman, J. J. (2000). ‘Causal parameters and policy analysis in economics: A Twentieth Century retrospective.’ *Quarterly Journal of Economics*, vol. 115, 45-97.

[56] Hirano, Keisuke, Imbens, Guido and Geert Ridder (2000): “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” manuscript, UCLA.

[57] Hollister, Robinson, Peter Kemper and Rebecca Maynard. 1984. *The National Supported Work Demonstration* (Madison: University of Wisconsin Press).

[58] Horowitz, Joel (2001):The Bootstrap," Handbook of Econometrics, Vol. 5, J.J. Heckman and E.E. Leamer, eds., Elsevier Science B.V.,Ch. 52, pp. 3159-3228.

[59] Ichimura, Hidehiko and Christopher Taber (1998): "Direct Estimation of Policy Impacts," manuscript, Northwestern University.

[60] Ichimura, Hidehiko and Christopher Taber (2002): "Semiparametric Reduced-Form Estimation of Tuition Subsidies" in *American Economic Review.* Vol. 92 (2). p 286-92.

[61] Imbens, G., and J. Angrist (1994): "Identification of Local Average Treatment Effects," in *Econometrica,* 62, 467-475.

[62] Jalan, Jyotsna and Martin Ravaillon (1999): "Income Gains to the Poor from Workfare: Evidence for Argentina's Trabajar Program," Policy Research Working Paper Series, the World Bank.

[63] Jalan, Jyotsna and Martin Ravaillon (2001): "Does Piped Water Reduce Diarrea for Children in Rural India," *Journal of Econometrics*, 112, 153-173..

[64] LaLonde, Robert (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review,* 76, 604-620.

[65] Lavy, Victor (2004): "Performance Pay and Teachers' Effort, Productivity and Grading Ethics," NBER Working Paper #10622, National Bureau of Economic Research.

[66] Lavy, Victor (2002): "Evaluating the Effects of Teachers' Group Performance Incentives on Pupil Achievement," Journal of Political Economics, 110(6), 1286-1387.

[67] Lechner, Michael (2002): "Some Practical Issues in the Evaluation of Heterogeneous Labour Market Programmes by Matching Methods" *Journal of the Royal Statistical Society*, Series A, 165(Part 1): 59-82.

[68] Lechner, Michael and Jefrey Smith (2000): “Some Exogenous Information Should Not Be Used in Evaluation Studies,” manuscript, University of Western Ontario.

[69] Manski, Charles and Steven Lerman (1977): “The Estimation of Choice Probabilities from Choice-Based Samples,” *Econometrica*, 45(8), 1977-1988.

[70] Manski, Charles and D. McFadden (1981): “Alternative Estimators and Sample Designs for Discrete Choice Analysis” in *Structural Analysis of Discrete Data with Economic Applications*, edited by C.F. Manski and D. McFadden (Cambridge, Mass.: MIT Press) 1-50.

[71] Manski, C. F. (1988): “Identification of Binary Reponse Models,” *Journal of the American Statistical Association*, 83, 403, 729-737.

[72] Manski, C. F. (1986): ”Semiparametric Analysis of Binary Response From Response-Based Samples,” *Journal of Econometrics*, 31, 31-40.

[73] Manski, C. F. and D. McFadden (1981): “Alternative Estimators and Sample Designs for Discrete Choice Analysis” in *Structural Analysis of Discrete Data with Economic Applications*, edited by C.F. Manski and D. McFadden (Cambridge, Mass.: MIT Press) 1-50.

[74] Manski, C. F. (1975): "The Maximum Score Estimation of the Stochastic Utility Model of Choice," in Journal of Econometrics, 27:313-333.

[75] Marschak, Jacob (1953): “Economic Measurements for Policy and Prediction,” in William Hood and Tjalling Koopmans, eds., Studies in Econometric Method (New York: John Wiley, 1953), pp. 1-26.

[76] McFadden, D. (1984): “Econometric Analysis of Qualitative Response Models,” *Handbook of Econometrics*, Vol. II, edited by Z. Griliches and M.D. Intriligator.

[77] McFadden, Daniel and A. P. Talvitie and Associates (1977): "Validation of Disaggregate Travel Demand Models: Some Tests" in Urban Demand Forecasting Project, Final Report, Volume V, Institute of Transportation Studies, University of California, Berkeley.

[78] Parker, Susan W., Todd, Petra E. and Kenneth I. Wolpin, "Within-Family Program Effect Estimators: The Impact of Oportunidades on Schooling in Mexico," manuscript.

[79] Pitt, Mark, Rosenzweig, Mark and Donna Gibbons (1993): "The Determinants and Consequences of Placement of Government Programs in Indonesia," *World Bank Economic Review*, 319-348.

[80] Porter, J. (1998): "Estimation of Regression Discontinuities", *Seminar Notes.*

[81] Raaum, Oddbjørn and Hege Torp (2001): "Labour Market Training in Norway – Effect on Earnings," *Labour Economics*, forthcoming.

[82] Ravaillon, Martin and Shaohua Chen (2005): "Hidden Impact? Household Savings in Response to a Poor Area Development Project," *Journal of Public Economics*, forthcoming.

[83] Rosenbaum, Paul and Donald Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects, "*Biometrika,* 70,41-55.

[84] Rosenbaum, Paul and Donald Rubin (1985): "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *American Statistician,* 39, 33-38.

[85] Rubin, D. B. (1980): "Bias Reduction Using Mahalanobis' Metric Matching," *Biometrics*, 36,2, pp. 295-298.

[86] Rubin, D. B. (1984): "Reducing BBias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association,* 79, pp. 516-524.

[87] Smith, Jeffrey and Petra Todd (2005): "Does Matching Address Lalonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics,* 125(1-2), March-April, 305-353.

[88] Todd, Petra and Kenneth I. Wolpin (2004): "Using Experimental Data to Validate a Dynamic Behavioral Model of Child Schooling: Assessing the Impact of a School Subsidy Program in Mexico," manuscript, University of Pennsylvania.

[89] Todd, Petra and Kenneth I. Wolpin (2005): "Ex-Ante Evaluation of Social Programs" manuscript, University of Pennsylvania.

[90] Thistlethwaite, D., and D. Campbell (1960) : "Regression-discontinuity Analysis: An alternative to the ex post facto experiment", *Journal of Educational Psychology*, 51, 309-317.

[91] Trochim, W. (1984): *Research Design for Program Evaluation: the Regression-Discontinuity Approach.* Beverly Hills: Sage Publications.

[92] Van der Klaauw, W. (1996): "A Regression-Discontinuity Evaluation of the Effect of Financial Aid Offers on College Enrollment," *International Economic Review.*

[93] Wise, David A. (1985): "A Behavioral Model Verses Experimentation: The Effects of Housing Subsidies on Rent" in *Methods of Operations Research*, 50, Verlag Anton Hain.

[94] Wolpin, Kenneth I. and Mark R. Rosenzweig (1988a): "Evaluating the Effects of Optimally Distributed Programs: Child Health and Family Planning Programs," *American Economic Review,* 76(3), 470-482.

[95] Wolpin, Kenneth I. and Mark R. Rosenzweig (1988b): "Migration Selectivity and Effects of Public Programs," in *Journal of Public Economics*, 37, 265-289.