Chapter 60 Evaluating Social Programs with Endogenous Program Placement and Selection of the Treated

ARTICLE in HANDBOOK OF DEVELOPMENT ECONOMICS · JANUARY 2007

Source: RePEc
Chapter 60

EVALUATING SOCIAL PROGRAMS WITH ENDOGENOUS PROGRAM PLACEMENT AND SELECTION OF THE TREATED

PETRA E. TODD

Department of Economics, University of Pennsylvania, McNeil 520, 3718 Locust Walk, Philadelphia, PA 19104, USA

Contents

Abstract 3848
Keywords 3848
1. Introduction 3849
2. The evaluation problem 3849
   2.1. Parameters of interest 3850
   2.2. What is the distinction between average program gain and average program gain for participants? 3852
   2.3. Sources of bias in estimating $E(\Delta \mid X, D = 1)$ and $E(\Delta \mid X)$ 3853
3. Solutions to the evaluation problem 3854
   3.1. Traditional regression estimators 3854
      3.1.1. Before–after estimators 3855
      3.1.2. Cross-sectional estimators 3856
      3.1.3. Difference-in-differences estimators 3857
      3.1.4. Within estimators 3858
      3.1.5. Applications 3859
   3.2. Matching methods 3861
      3.2.1. Reducing the dimensionality of the conditioning problem 3863
      3.2.2. Alternative matching estimators 3864
      3.2.3. Matching with choice-based sampled data 3868
      3.2.4. When does bias arise in matching? 3868
      3.2.5. Some additional considerations in applying matching methods 3868
      3.2.6. Assessing the variability of matching estimators 3870
      3.2.7. Applications 3871
   3.3. Control function methods 3873

* 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, John Strauss, Andrew Foster and other conference participants for very helpful comments. I also benefitted greatly from collaborations with James J. Heckman, Hidehiko Ichimura and Ken Wolpin. This chapter summarizes some results from that research.

Handbook of Development Economics, Volume 4
© 2008 Elsevier B.V. All rights reserved
DOI: 10.1016/S1573-4471(07)04060-0
3.3.1. Methods for estimating control functions
3.3.2. A comparison of control function and matching methods
3.3.3. Applications
3.4. Instrumental variables, local average treatment effects (LATE), and LIV estimation
3.4.1. The Wald estimator
3.4.2. Marginal treatment effects (MTE) and local instrumental variables (LIV) estimation and its relationship to TT, ATE, LATE
3.4.3. Applications
3.5. Regression-discontinuity methods
3.5.1. Identification of treatment effects under sharp and fuzzy data designs
3.5.2. Estimation
3.5.3. Applications of RD methods
4. Ex ante program evaluation
4.1. An illustrative model of identification of subsidy effects
5. Conclusions
References

Abstract

This chapter considers methods for evaluating the impact of social programs in the presence of nonrandom program placement or program selection. It first presents the evaluation problem as a missing data problem and then considers various solutions proposed in the statistics and econometrics literature. For ex post evaluation, the following estimation methods are discussed: traditional regression methods, matching, control function methods, instrumental variable and local instrumental variable (LIV) methods, and regression-discontinuity. Alternative estimators are described along with their identifying assumptions, the behavioral implications of those assumptions, and the data requirements for implementation. The chapter also considers methods for ex ante evaluation, which can be used to assess the effects of programs prior to their implementation, for example, in trying to design a program that achieves some desired outcomes for a given cost. Throughout the chapter, numerous examples from the development literature illustrate applications of the different estimation methods and highlight factors affecting estimator performance.

Keywords
program evaluation, self-selection, treatment effects, matching, control function, regression-discontinuity, instrumental variables

JEL classification: C21, C53, I38
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 non-randomly 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 is usually restricted to individuals or families who meet some criteria. The focus of this chapter is on estimating the effects of program interventions using 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 programs 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? This chapter considers alternative approaches to answering these questions, recognizing that how individuals respond to treatment is potentially heterogeneous.

We distinguish two types of evaluations, *ex post* evaluations, which analyze effects of existing programs, and *ex ante* evaluations, which analyze effects of programs that have not yet been implemented, often termed *counterfactual* programs. Most of this chapter considers methods for ex post evaluation, which is the most common and preoccupies most of the evaluation literature. Section four takes up the problem of evaluating programs prior their implementation, which is useful for designing new programs or for comparing existing programs to alternative ones. For example, if the program includes either conditional or unconditional subsidies, it could be of interest to assess program impacts for a range of subsidy levels.

The goals of this chapter are (i) to describe the identifying assumptions needed to justify the application of different kinds of 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.

2. The evaluation problem

We begin by defining some notation for describing the evaluation problem and key parameters of interest. For simplicity, suppose there are only two states of the world, corresponding to the state of being with and without some treatment intervention. For example, the outcome of interest could be a health indicator and the treatment could be participating in a health or nutrition program.

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 associated 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 potential outcomes 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 only one state is observed at any given time, the gain from treatment is not directly observed for anyone. Inferring gains from treatment therefore requires solving a missing data problem, and the evaluation literature has developed a variety of approaches to solve this problem.

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, spillover 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 to send 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; and
- **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 investigating program effects if the program offer were extended to individuals not currently participating. Nonparticipants are often used as a source of control group data, under the assumption that the indirect effects on nonparticipants are negligible, an assumption that is also maintained throughout this chapter.

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

---

1. For example, the program might induce changes in school quality measures, such as pupil–teacher ratios.
(a) the proportion of program participants that benefit from the program
\[ \Pr(Y_1 > Y_0 \mid D = 1) = \Pr(\Delta > 0 \mid D = 1); \]
(b) the proportion of the total population benefitting from the program:
\[ \Pr(\Delta > 0 \mid D = 1) \Pr(D = 1); \]
(c) quantiles of the impact distribution (such as the median), where \( q \) is the selected quantile
\[ \inf_{\Delta} \{ \Delta : F(\Delta \mid D = 1) > q \}; \]
(d) the distribution of gains for individuals with some characteristics \( X_0 \)
\[ F(\Delta \mid D = 1, X = X_0), \]
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:
(e) the average gain from the program for persons with characteristics \( X \)
\[ E(Y_1 - Y_0 \mid X) = E(\Delta \mid X); \]
(f) the average gain from the program for program participants with characteristics \( X \):
\[ E(Y_1 - Y_0 \mid D = 1, X) = E(\Delta \mid D = 1, X). \]

The parameter (e) is commonly referred to as the \textit{average impact of treatment (ATE)} and parameter (f) is known as the \textit{average impact of treatment on the treated (TT)}. The ATE parameter is the gain from the program that would be experienced on average if a randomly chosen person with characteristics \( X \) were assigned to participate in the program. The TT parameter is the average gain experienced for the subset of individuals who actually participated in the program (for whom \( D = 1 \)). If the individuals who take the program tend to be the ones that receive the greatest benefit from it, then one would 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 further 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 \((U_0 \text{ and } U_1)\):

\[
Y_1 = \varphi_1(X) + U_1, \\
Y_0 = \varphi_0(X) + U_0.
\]

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

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

Assume that \(E(U_0 \mid X) = E(U_1 \mid X) = 0\). 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 would expect that \(E(U_0 \mid X, D) \neq 0\) and \(E(U_1 \mid X, D) \neq 0\). That is, if the gain from the program depends on \(U_1\) and \(U_0\) and if people know their future values of \(U_1\) and \(U_0\), or can to some extent forecast the values, then we would expect people to 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

\[
E(\Delta \mid X) = \varphi_1(X) - \varphi_0(X) + E(U_1 \mid X) - E(U_0 \mid X)
\]

\[
= \varphi_1(X) - \varphi_0(X).
\]

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

\[
E(\Delta \mid X) = \varphi_1(X) - \varphi_0(X) + E(U_1 - U_0 \mid X, D = 1).
\]

As discussed in Heckman (2000), the average effect of treatment on the treated in unconventional in the sense that it combines the “structural parameters” (the parameters of the functions \(\varphi_0(X)\) and \(\varphi_1(X)\)) with means of the unobservables.

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

\[
E(\Delta \mid X) = \varphi_1(X) - \varphi_0(X) + E(U_1 - U_0 \mid X, D = 0),
\]

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. The relationship between
TT, ATE and UT is:

\[ ATE = \Pr(D = 1 \mid X)TT + \Pr(D = 0 \mid X)UT. \]

Observe that if \( U_1 = U_0 \), then the TT, ATE and UT parameters are the same. The advantage of allowing the residual term to differ in treated and untreated states is that it allows the potential for unobserved heterogeneity in how people respond to treatment. Under a special case, the parameters may be equal even if \( U_1 \neq U_0 \). That case arises when

\[ E(U_1 - U_0 \mid X, D) = 0. \]

This restriction implies the participation decision \( (D) \) is uninformative on \( U_1 - U_0 \), so that \( U_1 - U_0 \) could not have been a factor in the decision to participation. The restriction might be satisfied if the agents making the participation decisions (e.g. individuals, program administrators or others) do not act on \( U_1 - U_0 \), perhaps because agents do not know the idiosyncratic gain from participating in the program (and cannot forecast it) at the time of deciding whether to participate. In this special case, there is said to be ex post heterogeneity in how people respond to treatment, which is not acted upon ex ante.

As discussed in Heckman, La Londe 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. In order of increasing generality, they are:

(A.1) conditional on \( X \), the program effect is the same for everyone \((U_1 = U_0)\);
(A.2) conditional on \( X \), the program effect varies across individuals but \( U_1 - U_0 \) does not help predict program participation;
(A.3) conditional on \( X \), the program effect varies across individuals and \( U_1 - U_0 \) does predict who participates in the program.

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 \( E(\Delta \mid X, D = 1) \) and \( E(\Delta \mid X) \)

Consider again the model of the previous section

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

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

\[ Y = \varphi_0(X) + DE(\Delta \mid X) + \{U_0 + D(U_1 - U_0)\} \quad (*) \]

or

\[ Y = \varphi_0(X) + DE(\Delta \mid X, D = 1) \]

\[ + \{U_0 + D[U_1 - U_0 - E(U_1 - U_0 \mid X, D = 1)]\}. \]
For simplicity, suppose the $X$ variables are discrete and that we estimate the effects of the intervention ($D$) by the coefficients $b_X$ from an ordinary least squares regression$^3$:

$$Y = aX + b_X XD + v.$$  

This model is known as the common effect model and is popular in applied work. A special case of the model assumes that $b_X$ is constant across $X$:

$$Y = aX + bD + v.$$  

In light of the true 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 assumptions (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 the potential of bias from $E(U_1 - U_0 | D, X) \neq 0$. For estimating the $IT$ 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 ($Y_0$) 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 ($ITT$) parameter, $E(\Delta | X, D = 1)$, using nonexperimental data: (a) the before–after estimator, (b) the cross-sectional estimator, and (c) the difference-in-difference estimator. In each case, we illustrate the assumptions required to justify application of the estimator. 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 observation,

$$Y_{1it} = \varphi_1(X_{it}) + U_{1it},$$
$$Y_{0it} = \varphi_0(X_{it}) + U_{0it}.$$ \(1\)

$U_{1it}$ and $U_{0it}$ are assumed to be distributed independently across persons and to satisfy $E(U_{1it} | X_{it}) = 0$ and $E(U_{0it} | X_{it}) = 0$. Here, $X_{it}$ represents conditioning variables that may either be fixed or time-varying (such as gender or age), but whose distributions

$^3$ Here, we allow the effects of treatment to differ by the observed $X$, as reflected in the $X$ subscript on $b$. 
are assumed to be unaffected by whether an individual participates in the program. The observed outcome at time \( t \) can be written as

\[
Y_{it} = \phi_0(X_{it}) + D_{it}\alpha^*(X_{it}) + U_{0it},
\]

where \( D_{it} \) denotes being a program participant in the program and \( \alpha^*(X_{it}) = \phi_1(X_{it}) - \phi_0(X_{it}) + U_{1it} - U_{0it} \) is the treatment impact for an individual. Prior to the program intervention, we observe \( Y_{0it} = \phi_0(X_{it}) + U_{0it} \) for everyone. After the intervention we observe \( Y_{1it} = \phi_1(X_{it}) + U_{1it} \) for those who received the intervention (for whom \( D_{it} = 1 \), for \( t > t_0 \), the time of the intervention) and \( Y_{0it} = \phi_0(X_{it}) + U_{0it} \) for those who did not receive it (for whom \( D_{it} = 0 \) in all time periods).

This model is a random coefficient model, because the treatment impact can vary across persons even after conditioning on \( X_{it} \). Assuming that \( U_{1it} = U_{0it} = U_{it} \), so that the unobservable is the same in both the treated and untreated states, yields the fixed coefficient or common effect version of the model. In this model, the \( TT \) parameter is given by:

\[
\alpha^*_{TT}(X_{it}) = E(\alpha^*(X_{it}) | D_{it} = 1, D_{it'} = 0, X_{it}),
\]

where the conditioning on \( D_{it} = 1, D_{it'} = 0 \) denotes that the person was not in the program at time \( t' \) but did participate by time \( t \).

3.1.1. Before–after estimators

As noted 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.

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 for the \( TT \) parameter \( (\alpha^*_{TT}(X_{it}) = E(\alpha^*(X_{it}) | D_{it} = 1, D_{it'} = 0, X_{it})) \) is obtained by

\[
Y_{it} - Y_{it'} = \phi_0(X_{it}) - \phi_0(X_{it'}) + \alpha^*_{TT}(X_{it}) + \epsilon_{it}
\]

where

\[
\epsilon_{it} = [U_{1it} - U_{0it} - E(U_{1it} - U_{0it} | D_{it} = 1, D_{it'} = 0, X_{it})] + U_{0it} - U_{0it'}.
\]

Consistency of the estimator for \( \alpha^*_{TT}(X_{it}) \) requires that \( E(\epsilon_{it} | D_{it} = 1, D_{it'} = 0, X_{it}) = 0 \). The term in brackets has conditional mean zero by construction, so the assumption required to justify application of this estimator is \( E(U_{0it} - U_{0it'} | D_{it} = 1, D_{it'} = 0, X_{it}) = 0 \).

4 For example, if the set of conditioning variables \( X_{it} \) includes marital status and the program intervention is a job training program, we need to assume that the job training program does not affect marital status.
\(D_{it'} = 0, X_{it'}) = 0\). A special case where this assumption would be satisfied is if \(U_{0it}\) can be decomposed into a fixed effect error structure, \(U_{0it} = f_i + v_{it}\), where \(f_i\) does not vary over time and \(v_{it}\) is a i.i.d. random error that satisfies \(E(v_{it} - v_{it'} \mid D_t = 1, D_{it'} = 0, X_{it'}) = 0\). 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 that may affecting the program participation decision.

One drawback of a before–after estimation strategy is that identification 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.\(^5\) Before–after estimates can also be sensitive to the choice of time periods used to construct the estimator.

Many studies of employment and training programs in the US 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, La Londe and Smith, 1999.) The dip pattern can arise from serially correlated transitory downward shocks to earnings that may have been the impetus for the person applying to the training program.\(^6\) Another potential explanation for the observed dip pattern are the program eligibility criteria that are often imposed that tend to select out the most disadvantaged persons for participation in programs. These criteria will select into the program persons with low transitory earnings shocks. A simple before–after estimation strategy that includes the pre-program “dip” period typically gives an upward biased estimate of the effect of the program if the estimator is based on a pre-program period that occurs during the “dip”.

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

### 3.1.2. Cross-sectional estimators

A cross-sectional 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 post-program data on \(D_{it} = 1\) and \(D_{it} = 0\) persons. Define \(\hat{\alpha}_{CS}(X_{it})\) as the ordinary least squares solution to

\[
Y_{it} = \varphi_0(X_{it}) + D_{it} \alpha^{CS}_T(X_{it}) + \epsilon_{it},
\]

where

\[
\epsilon_{it} = U_{0it} + D_{it} [(U_{1it} - U_{0it}) - E(U_{0it} - U_{1it} \mid D_t = 1, X_{it})]
\]

\(^5\) Such a common time effect may arise, e.g., from life-cycle wage growth over time or from shocks to the economy.

\(^6\) A fixed effect error structure would not generate a dip pattern.
where the regression is estimated on \( D_{it} = 1 \) and \( D_{it} = 0 \) persons observed at time \( t \). Consistency of the cross-sectional estimator requires that \( E(\varepsilon_{it} \mid D_{it}, X_{it}) = 0 \). This restriction rules out the possibility that people select into the program based on expectations about their idiosyncratic gain from the program, which could violate the assumption that \( E(U_{i0} \mid X_{it}, D_{it}) = 0 \). Importantly, the cross-sectional estimator does not admit to unobservable variables affecting outcomes and the decision to participate in the program, which is a strong assumption.

### 3.1.3. Difference-in-differences estimators

The *difference-in-differences* (DID) estimator is commonly used in evaluation work, as in the numerous applications of it described below. This estimator measures the impact of the program intervention by the difference in the before–after change in outcomes between participants and nonparticipants.

Define an indicator that equals 1 for participants (for whom \( D_{it'} = 0 \) and \( D_{it} = 1 \)), denoted by \( IP \) and zero otherwise. The difference-in-differences treatment effect estimator is the least squares solution for \( \alpha_{TT}(X_{it}) \) in

\[
Y_{it} - Y_{it'} = \varphi_0(X_{it}) - \varphi_0(X_{it'}) + IP \alpha_{TT}(X_{it}) + \varepsilon_{it},
\]

\[
\varepsilon_{it} = D_{it}[U_{it} - U_{i0t} - E(U_{i1t} - U_{i0t} \mid D_{it} = 1, D_{it'} = 0, X_{it})]
\]

\[
+ U_{i0t} - U_{i0t'}.
\]

This regression is similar to a before–after regression, except that now the model is estimated using both participant and nonparticipant 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 (which can be included in \( \varphi_0(X_{it}) \)). The time effects are separately identified from the nonparticipant observations. The estimator is consistent if \( E(\varepsilon_{it} \mid D_{it}, X_{it}) = 0 \), which would be satisfied under a fixed effect error structure (see the discussion above for the before–after estimator). The data requirements to implement the DID estimator are either longitudinal or repeated cross section data on program participants and nonparticipants.7

Alternatively, the DID estimator can be implemented from a regression

\[
Y_{it} = \varphi_0(X_{it}) + IP \gamma + D_{it}\alpha_{TT}(X_{it}) + \tilde{\varepsilon}_{it} \quad \text{for } t = t', \ldots, t,
\]

\[
\tilde{\varepsilon}_{it} = U_{i0t} + D_{it}[U_{i1t} - U_{i0t} - E(U_{i1t} - U_{i0t} \mid D_{it} = 1, X_{it})].
\]

The main advantage of longitudinal estimators (before–after or difference-in-difference) over cross-sectional methods are that they allow there to be unobservable

---

7 For discussion of how repeated cross-sectional data can be used to implement difference-in-difference estimators, see Heckman and Robb (1985).
determinants of program participation decisions and outcomes. However, the fixed effect error structure that is usually assumed to justify application of these estimators only incorporates the potential influence of time-invariant unobservables.8

3.1.4. Within estimators

Within estimators identify program impacts from changes in outcomes within some unit, such as within a family, a school or a village. The previously described before–after and DID estimators fall within the class of within estimators, where the variation exploited is for a given individual over time. We next describe other kinds of within estimators where the unit of observation is broader than a single individual, representing, for example, a family or village.

Let \( Y_{0ijt} \) and \( Y_{1ijt} \) denote the outcomes for individual \( i \), from unit \( j \), observed at time \( t \), and for simplicity at first assume that \( U_{1it} = U_{0it} \). Write the model for outcomes as:

\[
Y_{ijt} = \phi_0(X_{ijt}) + I_{ijt}^D \gamma + D_{ijt} \alpha_{TT}(X_{ijt}) + \epsilon_{ijt}
\]

and assume that the error term \( \epsilon_{ijt} = U_{0it} \) can be decomposed as:

\[
\epsilon_{ijt} = \theta_j + v_{ijt},
\]

where \( \theta_j \) represents the effects of unobservables that vary across units but are constant for individuals within the same unit and \( v_{ijt} \) are i.i.d.

Taking differences between two individuals from the same unit observed in the same time period gives

\[
Y_{ijt} - Y_{i'jt} = \phi_0(X_{ijt}) - \phi_0(X_{i'jt}) + (I_{ijt}^D - I_{i'jt}^D) \gamma + (D_{ijt} - D_{i'jt}) \alpha_{TT}(X_{ijt}) + (v_{ijt} - v_{i'jt}).
\]

Consistency of the OLS estimator of \( \alpha_{TT}(X_{ijt}) \) requires that

\[
E(v_{ijt} - v_{i'jt} \mid X_{ijt}, X_{i'jt}, D_{ijt}, D_{i'jt}) = 0.
\]

This assumption implies that within a particular unit, which individual receives the treatment is random with respect to the error term \( v_{ijt} \). In the more general random coefficients version of the model, it has to be assumed that within a particular unit, which individual receives the treatment is random with respect to that individual’s idiosyncratic gain from the program. That is, the program may be targeted at specific units (e.g. families or villages), but within those units, for a within estimation strategy to be valid, it is necessary to assume that which individuals participated in the program is unrelated to the idiosyncratic gain from the program. Also, because the estimator relies on comparisons between the outcomes of treated and untreated persons, the approach requires assuming that there be no spillover effects from treating one individual on other individuals within the same unit. As with the before–after and difference-in-differences estimation approaches, the within estimator allows treatment to be selective across units.

---

8 For example, an unobservable attribute such as an individual’s “motivation” could affect the outcome measure and also influence the program participation decision.
Namely, it allows \( E(\varepsilon_{ijt} \mid D_{ijt}, X_{ijt}) \neq 0 \), because treatment selection can be based on the unobserved heterogeneity term \( \theta_j \).

When the variation being exploited for identification of the treatment effect is variation within a family, village, or school at a single point in time, then the within estimator can be implemented with as little as a single cross section of data.

Sometimes it happens that all individuals within a unit receive treatment at the same time, in which case \( D_{ijt} = D_{ij't} \) for all \( i \) in \( j \) and the above approach is not feasible. In that situation, a within-estimation strategy may still be feasible if pre-program data (\( t' \)) are available by taking differences across individuals in the same unit observed at different time periods:

\[
Y_{i'jt'} - Y_{ijt} = \varphi_0(X_{i'jt'}) - \varphi_0(X_{ijt}) + (I^D_{ij} - I^D_{ij't}) \gamma + (D_{ijt} - D_{ij't}) \alpha^* + (v_{ijt} - v_{ij't}),
\]

where \( D_{ij't} = 0 \). Consistency requires that \( E(v_{ijt} - v_{ij't} \mid D_{ijt}, D_{ij't}, X_{ijt}, X_{ij't}) = 0 \). When \( I^D_{ij} = 1 \) 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.  

3.1.5. Applications

The above described estimators are widely used in empirical evaluation research in development. One of the earliest applications of the within estimator is by Rosenzweig and Wolpin (1986), which assesses the impact of a family planning and health counseling program on child outcomes in the Philippines. Their study provides an early discussion of the statistical problems created by nonrandom program placement, in particular, when the placement of a program potentially depends on the outcome variable of interest. For example, family planning programs are often placed in areas where the need is considered to be the greatest. Not accounting for nonrandom placement would lead to the erroneous conclusion that family planning programs cause fertility.

Rosenzweig and Wolpin’s (1986) empirical analysis adopts the following statistical model:

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

where \( H_{ijt}^a \) is a child health measure (height, weight) for child \( i \) observed at age \( a \), living in locality \( j \) at time \( t \). \( \rho_{ij}^a \) represents the length of time that child was exposed to the program intervention. \( \mu_i \) is a time invariant, child-specific unobserved health endowment and \( \mu_j \) is an unobserved locality level effect. The estimation approach compares changes in health outcomes for children who were exposed to the program to changes for children who were not exposed to it.  

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

Locality level effects are separately identified from individual effects using observations on families that migrated across localities. Without migration, they would not be separately identified.
the allocation of the program to be selective on unobservables, namely locality level or individual level unobserved characteristics. A subsequent study by Rosenzweig and Wolpin (1988a, 1988b) adopts a similar within child estimation strategy to evaluate the effects of a Colombian child health intervention.

A more recent evaluation that adopts a similar identification and estimation strategy 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. The paper notes that the new schools were in part locally financed, which led to nonrandom placement of schools into more affluent communities. Because individuals from those communities usually 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 and year. For this reason, the education of individuals who were young when the program began would be more affected by the school building program than that of older individuals. Also, individuals in regions where a larger numbers of schools were built are more likely to have been affected by the 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.

A recent paper by Glewwe et al. (2004) questions the reliability of a difference-in-difference estimation approach in an application that evaluates the effectiveness of an educational intervention in Kenya. The program intervention provided schools with flip-charts to use as teaching aids in certain subjects. One of the goals of their study is to compare the estimates obtained by a nonexperimental DID estimation approach to those obtained from a randomized social experiment. Their DID estimator 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 and compares the schools that did and did not receive the intervention. When Glewwe et al. (2004) compare the nonexperimental to the experimental estimates, they find substantial differences. The experimental results indicate that flip-charts had little effect on test scores, while the DID estimates are statistically significantly different from zero at conventional levels. The authors conclude that the difference-in-difference estimator is unreliable. 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 interventions.

11 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.
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 of the earliest applications of matching to evaluate economic development programs were World Bank evaluations of anti-poverty programs.\(^{12}\)

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.\(^ {13}\)

There are two different variants of matching estimators, cross-sectional matching and difference-in-difference matching. Cross-sectional matching estimators allow for selection on unobservables only in a very limited sense, as described below. For the most part, these estimators are only applicable in contexts where the researcher is relatively certain that the major determinants of program participation are accounted for and that any remaining variation in who participates is due to random factors. Difference-in-difference matching estimators identify treatment effects by comparing the change in outcomes for treated persons to the change in outcomes for matched, untreated persons. Difference-in-difference matching estimators allow for selection into the program to be based on unobserved time-invariant characteristics of individuals. Below, we first describe cross-sectional matching estimators, which are the type considered in most of the statistical literature on matching and are the most widely used. Then we discuss difference-in-difference matching, which is a more recent variant introduced in the econometrics literature.

Cross-sectional matching estimators 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 \(Z\),\(^ {14}\)

\[
(Y_0, Y_1) \indep | Z. \tag{3}
\]

---

\(^{12}\) See the applications discussed below.

\(^{13}\) For discussions of other kinds of matching estimators, see e.g. Cochran and Rubin (1973), Rubin (1980, 1984).

\(^{14}\) In the terminology of Rosenbaum and Rubin (1983) treatment assignment is “strictly ignorable” given \(Z\).
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.,

$$0 < \Pr(D = 1 \mid Z) < 1.$$  \hspace{1cm} (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 program impacts can be solved by simply substituting the $Y_0$ distribution observed for the matched nonparticipant 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:

$$E(Y_0 \mid Z, D = 1) = E(Y_0 \mid Z, D = 0) = E(Y_0 \mid Z).$$  \hspace{1cm} (5)

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

$$\Pr(D = 1 \mid Z) < 1.$$  \hspace{1cm} (6)

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

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

$$= E(Y_1 \mid D = 1) - E_{Z \mid D = 1}\{E_Y(Y \mid D = 1, Z)\}$$

$$= E(Y_1 \mid D = 1) - E_{Z \mid D = 1}\{E_Y(Y \mid D = 0, Z)\},$$

where the second term can be estimated from the mean outcomes of the matched (on $Z$) comparison group. 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 anticipated 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, the estimator accommodates selection on unobservables (assumption (A-3) discussed in Section 1.3), but only in a very limited sense, because there is a strong restriction on the nature of the selection process.

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) and Heckman, Ichimura, Smith and Todd (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

---

15 The notation $E_{Z \mid D = 1}$ denotes that the expectation is taken with respect to the $f(Z \mid D = 1)$ density. $E_{Z \mid D = 1}(E_Y(Y \mid D = 0, Z)) = \int_z \int_y y f(y \mid D = 0, z) f(z \mid D = 1) dy dz$. 

---
not overlap for the $D = 1$ and $D = 0$ groups, then matching is only justified when performed over the region of common support. 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. Rosenbaum and Rubin (1983) provide a theorem that is useful in reducing the dimension of the conditioning problem. They show that for random variables $Y$ and $Z$ and a discrete random variable $D$

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

so that

$$E(D \mid Y, Z) = E(D \mid Z)$$

$$\implies E(D \mid Y, Pr(D = 1 \mid Z)) = E(D \mid Pr(D = 1 \mid Z)).$$

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, $P(Z) = Pr(D = 1 \mid Z)$. Thus, when matching on $Z$ is valid, matching on the summary statistic $Pr(D = 1 \mid Z)$ (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 problem to that of a univariate problem. Because they are much easier to implement, much of the literature on matching focuses on propensity score matching methods.

Using the Rosenbaum and Rubin (1983) theorem, the matching procedure can be broken down into two stages. In the first stage, the propensity score $Pr(D = 1 \mid Z)$ 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 first stage estimated probabilities of participation.

The literature has developed a variety of matching estimators. The next section describes some of the leading examples.

---

16 An advantage of randomized experiments noted by Heckman (1997), as well as Heckman, Ichimura and Todd (1997) and Heckman et al. (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.

17 If $Z$ is discrete, small cell problems may arise. If $Z$ is continuous and the conditional mean $E(Y_0 \mid D = 0, Z)$ is estimated nonparametrically, then convergence rates will be slow due to the “curse of dimensionality” problem.

18 Heckman, Ichimura 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 $JT$ 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.

19 Semiparametric estimation methods, such as Ichimura’s (1993) semiparametric least squares (SLS) method can also be used.
3.2.2. Alternative matching estimators

For notational simplicity, let \( P = P(Z) \). A typical cross-sectional matching estimator takes the form

\[
\hat{\alpha}_M = \frac{1}{n_1} \sum_{i \in I_1 \cap S_P} \left[ Y_{1i} - \hat{E}(Y_{0i} \mid D = 1, P_i) \right].
\]

\[
\hat{E}(Y_{0i} \mid D = 1, P_i) = \sum_{j \in I_0} W(i, j)Y_{0j}, \tag{8}
\]

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.

3.2.2.1. Nearest neighbor matching  
Traditional, pairwise matching, also called nearest-neighbor matching, sets

\[
C(P_i) = \min_j \| P_i - P_j \|, \quad j \in I_0.
\]

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.

3.2.2.2. 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 \( \varepsilon \) is a prespecified 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 is a reasonable choice for the tolerance level.
3.2.2.3. 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.

3.2.2.4. 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

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

where $G(\cdot)$ is a kernel function and $a_n$ is a bandwidth parameter.\(^{20}\) In terms of Eq. (8), the weighting function, $W(i, j)$, is equal to

$$\frac{G(P_j - P_i)}{\sum_{k \in I_0} G(P_k - P_i)}.$$  

For a kernel function bounded between $-1$ and $1$, the neighborhood is $C(P_i) = \{ |P_i - P_j| \leq 1 \}, j \in I_0$.

Under standard conditions on the bandwidth and kernel, \(|\sum_{j \in I_0} Y_{0j} G(P_j - P_i)\) is a consistent estimator of $E(Y_0 \mid D = 1, P_i)$.\(^{21}\)

Heckman, Ichimura and Todd (1997) also propose a generalized version of kernel matching, called local linear matching.\(^{22}\) The local linear weighting function is given by

$$W(i, j) = \frac{G_{ij} \sum_{k \in I_0} G_{ik}(P_k - P_i)^2 - [G_{ij}(P_j - P_i)][\sum_{k \in I_0} G_{ik}(P_k - P_i)]}{\sum_{j \in I_0} G_{ij} \sum_{k \in I_0} G_{ij}(P_k - P_i)^2 - (\sum_{k \in I_0} G_{ik}(P_k - P_i))^2}.$$  

As demonstrated in research by Fan (1992a, 1992b), 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, 1992b).


\(^{21}\) Specifically, we require that $G(\cdot)$ integrates to one, has mean zero and that $a_n \to 0$ as $n \to \infty$ and $na_n \to \infty$.

\(^{22}\) Recent research by Fan (1992a, 1992b) 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, 1992b).
convergence near boundary points and greater robustness to different data design densities. (See Fan, 1992a, 1992b.) 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.

To implement the matching estimator given by Eq. (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

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

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

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

and $a_n$ is the bandwidth parameter. To ensure that the densities are strictly greater than zero, it is required that the densities be strictly positive (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:

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

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 (2000). 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.
3.2.2.5. Difference-in-difference matching  The cross-sectional 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 non-participant outcomes, even after conditioning on observables, which could lead to a violation of the above maintained assumptions. 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-difference (DID) matching strategy, as defined in Heckman, Ichimura and Todd (1997) and Heckman et al. (1998), better accommodates the potential for selection on unobservables by allowing 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 implied by 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

\[
\hat{\alpha}_{DID} = \frac{1}{n_1} \sum_{i \in I_{1t} \cap \tilde{S}_q} \left\{ (Y_{1ti} - Y_{0t'i}) - \sum_{j \in I_{0t'} \cap \tilde{S}_q} W(i, j)(Y_{0tj} - Y_{0t'j}) \right\},
\]

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

\[
\hat{\alpha}_{DID} = \frac{1}{n_{1t}} \sum_{i \in I_{1t} \cap \tilde{S}_q} \left\{ Y_{1ti} - \sum_{j \in I_{0t} \cap \tilde{S}_q} W(i, j)Y_{0tj} \right\}
\]

\[
- \frac{1}{n_{1t'}} \sum_{i \in I_{1t'} \cap \tilde{S}_q} \left\{ Y_{1t'i} - \sum_{j \in I_{0t'} \cap \tilde{S}_q} W(i, j)Y_{0t'j} \right\},
\]

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 also allows selectivity into the program to be based on anticipated gains from the program, in the sense of assumption (A.2) described in Section 2.2. That is, \( D \) can help predict the value of \( Y_1 \) given \( P \). However, a maintained assumption is that \( D \) does not help predict changes in the value of \( Y_0 \) (i.e. \( Y_{0t} - Y_{0t'} \)) 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 different in terms of their changes in \( Y_0 \).
3.2.3. Matching with choice-based sampled data

The samples used in evaluating the impacts of programs are often choice-based, with program participants being oversampled relative to their frequency in the population. Under choice-based sampling, weights are required to consistently estimate the probabilities of program participation, where the weights correspond to the ratio of the proportion of program participants in the population relative to the proportion in the sample. The true population proportions usually are not obtainable from the sample and have to be derived from some other sources. When the weights are known, the Manski and Lerman (1977) procedure can be implement to consistently estimate propensity scores. However, oftentimes the population weights are unknown. Heckman and Todd (1995) show that in the case where the weights are unknown, with a slight modification, matching methods can still be applied, because the odds ratio \( \frac{P}{1-P} \) 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).

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

\[
\alpha'_M = E_{P(Z_0) \mid D=1}(E(Y_1 \mid P(Z_0), D = 1) - E(Y_0 \mid P(Z_0), D = 0)).
\]  

(7)

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

\[
\text{bias}_M = E(Y_0 \mid D = 1) - E_{P(Z_0) \mid D=1}(E(Y_0 \mid P(Z_0), D = 0)).
\]

3.2.5. Some additional considerations in applying matching methods

3.2.5.1. Choosing the set of matching variables  As described earlier, the propensity score matching estimator requires that the outcome variable be mean independent of the treatment indicator conditional on the propensity score, \( P(Z) \). An important consideration in implementation is how to choose the set of conditioning variables used in

---

23 See, e.g., Manski and Lerman (1977) for discussion of weighting for logistic regressions.
24 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.
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 conditions for matching could lead to a violation of the conditions. Using too many 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 and Todd (1997) and Lechner (2002) show that which variables are included in the estimation of the propensity score can make a substantial difference to the estimator’s performance. These papers found that biases tended to be more substantial 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.

3.2.5.2. Other determinants of the performance of matching estimators

Empirical explorations have shown that matching estimators perform best when the treatment and control groups are located in the same geographic area, so that regional effects on outcomes are held constant across groups. Lastly, a few papers have 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. Smith and Todd (2005) and Heckman, Ichimura and Todd (1997) found that matching estimators performed poorly when the survey instrument is not the same and concluded from that evidence that matching estimators do not compensate for biases caused by differences in how variables are measured across surveys, a purpose for which they were not designed. The results also indicated that difference-in-difference matching methods are more reliable than cross-sectional matching methods, particularly when treatments and controls are mismatching geographically or in terms of the survey instrument. In general, the success of matching approaches to evaluation depends strongly on the data being of relatively high quality.

3.2.5.3. Using balancing tests to check the propensity score specification

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

$$Z \perp D \mid \Pr(D = 1 \mid Z),$$

or equivalently

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

25 It is often the case in evaluation work that the comparison group data are collected using a different survey instrument. (See La Londe, 1986; Dehejia and Wahba, 1998, 1999; and Smith and Todd, 2005.)
The basic intuition is that after conditioning on $\Pr(D = 1 \mid Z)$, additional conditioning on $Z$ should not provide new information about $D$. Thus, if after conditioning on the estimated values of $P(D = 1 \mid Z)$ there is still dependence on $Z$, this suggests misspecification in the model used to estimate $\Pr(D = 1 \mid Z)$. 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 $\Pr(D = 1 \mid Z)$, which tests whether or not there are differences in $Z$ between the $D = 1$ and $D = 0$ groups after conditioning on $P(Z)$. Various testing approaches have been proposed in the literature. Eichler and Lechner (2001) use a variant of a test 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 Dehejia and Wahba (1998, 1999) 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.

Another 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 $P(Z)$:

$$Z_k = \alpha + \beta_1 P(Z) + \beta_2 P(Z)^2 + \beta_3 P(Z)^3 + \cdots + \beta_j P(Z)^j + \gamma_1 P(Z)D + \gamma_2 P(Z)^2D + \gamma_3 P(Z)^3D + \cdots + \gamma_j P(Z)^jD + \nu,$$

and then tests whether the estimated $\gamma$ 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. In this way, the specification for the propensity score is iteratively refined.

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. A recent paper by Imbens and Abadie (2004a) shows that standard bootstrap resampling methods are not valid for assessing the variability of nearest neighbor estimators. Their criticism does not, however, apply for kernel or local linear matching estimators, for which bootstrap methods are valid. 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, Jalan and Ravallion (2003) 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, Jalan and Ravallion (2001) 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.

Handa and Maluccio (2006) study the performance of matching estimators by comparing nonexperimental estimates obtained by matching to estimates obtained from a randomized social experiment. They find that the matching estimators perform well in replicating the experiment only for outcomes that are relatively easily measured, such as schooling attainment, but perform less well for more complex outcomes such as expenditures. They find that stringently imposing common support and choosing highly comparable comparison groups from which to draw the matched outcomes improves the performance of their propensity score matching estimators.

Matching methods were also used in the large-scale evaluation of the urban Oportunidades program in Mexico that was carried out in 2005. The program is described in detail in Chapter 62 of this handbook. Briefly, the Oportunidades program provides monetary subsidies to families for 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

---


27 Upon more detailed examination of the distribution of treatment effects, however, Jalan and Ravallion (2003) also observe that the observed health gains largely bypass children from the poorest families, particularly those where the mother is poorly educated.
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. 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. When longitudinal variation in the outcome of interest was available, difference-in-difference matching estimators were applied. 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.

In another recent application of difference-in-difference matching methods, Galiani, Gertler and Schargrodsky (2005) analyze effects of privatization of water services on child mortality in Argentina. Temporal variation of ownership in water provision provides a source of variation for identifying the effects of privatization, although which municipalities privatized first was nonrandom. To take into account unobserved municipality characteristics that may affect the decision to privatize and may affect child health outcomes, Galiani, Gertler and Schargrodsky (2005) use a difference-in-difference kernel matching. Their study finds that privatization of water services significantly reduced child mortality and that the effects were most pronounced in the poorest areas. Godtland et al. (2004) apply cross-sectional propensity score matching estimators to evaluate the effectiveness of a farmer field school that provided agricultural extension services to farmers in Peru. They find a statistically significantly positive effect of the program on measures of farmer knowledge.

Behrman, Cheng and Todd (2000) use matching methods 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 is used to nonparametrically recover the relationship between program duration and magnitude of treatment impact.

Other applications of matching methods in the recent economic development literature are Gertler, Levine and Ames (2004), in a study of the effects of parental death on

---

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

29 See, for example, Behrman et al. (2005), which analyzes the impact of Oportunidades on education outcomes.

3.3. Control function methods

As noted in the previous section, matching estimators do not require specifying functional form assumptions for the outcome equation, but they make strong assumptions about how unobservables are allowed to affect program participation decisions. The most general variant of matching, difference-in-difference matching, allows individuals to select into programs based on time-invariant unobservable characteristics. It does not allow time-varying unobservables to affect participation decisions.

Another class of evaluation estimators are control function methods, also known as generalized residual methods. These methods were proposed as a solution to the evaluation problem in Heckman and Robb (1985), but they are closely related to the earlier selection bias correction methods developed in Heckman (1979). 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 estimators explicitly recognize that nonrandom selection into the program gives rise to an endogeneity problem and aim to obtain unbiased parameter estimates by explicitly modeling the source of the endogeneity. They allow selection into the program to be based on time varying unobservable variables at the expense of stronger functional form assumptions needed to secure identification.

To see how the generalized residual method applies to the evaluation problem, write the model for outcomes as

$$Y = \varphi_0(X) + D\alpha_{TT}(X) + \tilde{\epsilon},$$

where

$$\alpha_{TT}(X) = E(Y_1 - Y_0 \mid X, D = 1) = \varphi_1(X) - \varphi_0(X) + E(U_1 - U_0 \mid X, D = 1)$$

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

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

Because the decision to participate may be endogenous with respect to the outcomes, we expect that $E(U_0 \mid X, D) \neq 0, i = 0, 1$.

Heckman (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 $E(U_0 \mid X, D) = DE(U_0 \mid D = 1, X) + (1 - D)E(U \mid D = 0, X)$, we can rewrite the outcome model as
\[ Y = \varphi_0(X) + D\alpha^{*\ast}_T(X) + E(U \mid D = 0, X) \]
\[ + D\left[ E(U_0 \mid D = 1, X) - E(U_0 \mid D = 0, X) \right] + \varepsilon \]
\[ = \varphi_0(X) + D\alpha^{*\ast}_T(X) + K_0(X) + D\left[ K_1(X) - K_0(X) \right] + \varepsilon \] (8)

where
\[ K_0(X) = E(U_0 \mid D = 0, X), \]
\[ K_1(X) = E(U_0 \mid D = 1, X), \]
\[ \varepsilon = D\left[ U_0 - E(U_0 \mid D = 1, X) \right] + (1 - D)\left[ U_0 - E(U_0 \mid D = 0, X) \right] \]
\[ + D\left[ U_1 - E(U_1 \mid D = 1, X) \right]. \]

By construction, \( \varepsilon \) 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 nonlinear) applied to consistently estimate program.

3.3.1. Methods for estimating control functions

If no restrictions where placed on either \( \alpha^{*\ast}_T(X) \), \( K_1(X) \), or \( K_0(X) \), then the treatment impact parameter \( (\alpha^{*\ast}_T(X)) \) could not be separately identified from the control functions. For example, suppose \( K_1(X) - K_0(X) = \rho_0 + \rho_1 X + \rho_2 X^2 \). Clearly, \( \alpha^{*\ast}_T(X) \) could not be separately identified from the difference in the control functions. Some identifying restrictions are necessary, and different implementations of control function estimators proposed in the literature impose different kinds of restrictions. Usually, the restrictions 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 choice of \( D \)) be excluded from the outcome equation. These excluded variables generate variation in \( K_1(X) \) and \( K_0(X) \) that is independent from \( \alpha^{*\ast}_T(X) \).

The following types of restrictions could be imposed: (1) functional form restrictions on \( \alpha^{*\ast}_T(X) \) 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.

One interesting approach to identification considered in the econometric evaluation literature is called identification at infinity. (See Heckman, 1990; Andrews and Schafgans, 1998.) This approach is feasible only when there is a subgroup in the data for which \( \Pr(D = 1 \mid Z) = 1 \) 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, under the requirement that there is at least one continuous variable included in \( Z \) but not contained in \( X \) (termed an exclusion restriction).

Heckman and Robb (1985) 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 a set of 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, 1984, and Manski and McFadden, 1981.)

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

$$E(U_0 | D = 1, X) = E(U_0 | h(Z\gamma) + V > 0, X)$$

$$= \frac{\int_{h(Z\gamma)}^{\infty} \int_{-\infty}^{\infty} uf(u, v | X) du dv}{\int_{h(Z\gamma)}^{\infty} \int_{-\infty}^{\infty} f(u, v | X) du dv}.$$  

If $F(u, v | X)$ is assumed to be continuous with full support in $R^2$ and $F_V(\cdot)$ is invertible, then the index $Z\gamma$ can be written as a function of the conditional probability of participation:

$$h(Z; \gamma) = -F^{-1}_V(-Pr(D = 1 | Z)).$$

Heckman and Robb (1985) note that with the additional assumption that the joint distribution of the unobservables, does not depend on $X$, except possibly through the index, $h(Z; \gamma)$:

$$f(u, v | X) = f(u, v | h(Z; \gamma)).$$

then $E(U_0 | D = 1, X)$ can be written solely as a function of the probability of participating in the program, $Pr(D = 1 | Z)$:

$$E(U_0 | D = 1, X) = E(U_0 | D = 1, P(Z)) = K_1(P(Z)).$$

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 aids in the identification problem. For example, suppose $\varphi_0(X)$ and $h(Z\gamma)$ were both linear in the regressors. Under the index assumption and

30 From the above expression, it can be seen that as $h(Z\gamma)$ approaches infinity, $E(U_0 | D = 1, Z)$ approaches 0 ($= E(U_0 | Z)$). For this reason, subgroups with a high probability of participating in the program can be used to secure identification of model parameters. (See Heckman, 1990.)

31 A stronger assumption that would all imply index sufficiency is independence, $f(u, v | X) = f(u, v).$
with one continuous variable included in $Z$ but excluded from $X$, we can allow for overlap between $X$ and $Z$ and even for the case where $X$ are fully contained in $Z$.\footnote{If the control functions are estimated nonparametrically, separately distinguishing the treatment effect from the control function requires the application of identification at infinity methods. These approaches have been criticized in the literature, because they base identification of a subset of model parameters on a typically small sample of individuals with a probability of participating close to 1.}

In the original formulation of the control function method in\footnote{In that study, estimates of the probabilities of participating in the program (the propensity scores) are 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 US JTPA (Job Training and Partnership Act) program.} Heckman (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 et al. (1998), 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.\footnote{See Heckman, Ichimura and Todd (1997, 1998) for the more general case where $Z$ contains variables not in $X$.}

### 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 both make use of propensity scores in implementation and are related. Conventional matching estimators can in some cases be viewed as a restricted form of a control function estimator. Recall that traditional cross-sectional 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 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 notation of the previous model for outcomes, this assumption is implies that\footnote{See Heckman, Ichimura and Todd (1997, 1998) for the more general case where $Z$ contains variables not in $X$.}

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

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

$$K_1(P(Z)) - K_0(P(Z)) = 0.$$  \hspace{1cm} \text{(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 to as the special case of selection on observables (see Heckman and Robb, 1985; Heckman et al., 1996; and Barnow, Cain and Goldberger, 1980).

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 was 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 satisfied unless the errors have zero covariance, $\sigma_{U_0V} = 0$. To see why that is the case, note that under joint normality

$$E(U_0 | D = 1, Z) = K_1(P(Z)) = \frac{\sigma_{U_0V}}{\sigma_V} \frac{\phi(-h(Z \gamma))}{1 - \Phi(-h(Z \gamma))},$$

$$E(U_0 | D = 0, Z) = K_0(P(Z)) = \frac{\sigma_{U_0V}}{\sigma_V} \frac{-\phi(-h(Z \gamma))}{\Phi(-h(Z \gamma))}.$$ 

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

### 3.3.3. Applications

Control function methods have not yet been widely used in the context of evaluating development programs. For discussion of many applications to evaluating job training programs, see Heckman, La Londe and Smith (1999).

### 3.4. Instrumental variables, local average treatment effects (LATE), and LIV estimation

Instrumental variables methods provide another approach to estimating program effects in the presence of nonrandom self-selection. In this section, we consider both traditional instrumental variables estimators as well as more recently developed local instrumental variable (LIV) methods.

#### 3.4.1. The Wald estimator

Consider again the treatment effect model of the previous section:

$$Y = \varphi_0(X) + D\alpha^*_T(X) + \tilde{\varepsilon},$$
where
\[ \alpha^*_{TT}(X) = E(Y_1 - Y_0 \mid X, D = 1) = \alpha(X) + E(U_1 - U_0 \mid X, D = 1) \]
is the parameter of interest (TT) and
\[ \tilde{\epsilon} = U_0 + D(U_1 - U_0 - E(U_1 - U_0 \mid X, D = 1)) . \]

Suppose that there is an exclusion restriction, namely a variable \( Z \) that affects the program participation decision but does not enter into the outcome equation. Also, for ease of exposition, assume that the conditioning variables \( X \) and the instrument \( Z \) are binary variables and that the instrument 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 within \( X \) subsamples. The identifying assumption is that
\[ E(U_0 \mid X, Z) = E(U_0 \mid X) . \]
The so-called Wald estimator (applied within \( X \) strata) is given by
\[ \hat{\alpha}^*_{IV}(X) = \frac{\hat{E}(Y \mid Z = Z_1, X) - \hat{E}(Y \mid Z = Z_2, X)}{\hat{P}(D = 1 \mid Z = Z_1, X) - \hat{P}(D = 1 \mid Z = Z_2, X)} , \]
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 \( \hat{\alpha}^*_{IV}(X) \) recovers the average impact of treatment on the treated (the TT parameter) only under one of two alternative assumptions on the error term:

Case I: \( U_1 = U_0 \)

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

Either of these assumptions would give \( E(D(U_1 - U_0 - E(U_1 - U_0 \mid X, D = 1))) = 0 \). 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.\(^{35}\) Heckman (1992) provides several examples where the assumption that the instrument does not help forecast the program gain can be problematic. Whether such an assumption is tenable or not will depend on the particular application at hand.

\(^{35}\) Note that \( E(D(U_1 - U_0 - E(U_1 - U_0 \mid X, D = 1))) \mid X, Z) = \Pr(D = 1 \mid X)E(U_1 - U_0 - E(U_1 - U_0 \mid X, D = 1)) \mid X, Z, D = 1 \), so the required assumption is that \( E(U_1 - U_0 \mid X, Z, D = 1) = E(U_1 - U_0 \mid X, D = 1) \). For this reason, as shown in Heckman (1992), the required assumption to apply the IV estimator for the purpose of estimating the treatment on the treated parameter is that the instrument not help forecast the gain from the program.
If assumptions I or II are not satisfied, then the Wald estimator no longer recovers the average impact of treatment on the treated. Nonetheless, it has a meaningful 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 a change in the value of the instrument from $Z_1$ to $Z_2$ 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 $Z = Z_1$. The LATE parameter is discussed further 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 and of new ways to estimate them. Heckman and Vytlacil (2005) develop a unifying theory of how the parameters relate to one another using a new concept, called a marginal treatment effect (MTE). Here, we provide an overview of the major results of the paper. 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:

$$
Y = DY_1 + (1 - D)Y_0, \\
Y_1 = \mu_1(X, U_1), \\
Y_0 = \mu_0(X, U_0), \\
D = 1 \text{ if } \mu_0(Z) - U_D \geq 0.
$$

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 (i.e. that there is an exclusion restriction). The error terms are assumed to be independent of $Z_i$ conditional on $X_i$. 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 $U_D$ distributed uniformly. To see why, suppose that

$$
D = 1 \text{ if } \varphi(Z) - v \geq 0
$$

so that

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

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

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

36 See Heckman and Vytlacil (2005) for other technical conditions that are not central to the argument here.
Define $U_D = F_V(v)$ and note that $P(U_D < t) = t$. Thus, $U_D$ is uniformly distributed between 0 and 1.

Next, note that when $U_D$ is uniformly distributed,

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

Let $Z$ and $Z'$ be two values of the instrument such that $Pr(D = 1 \mid Z) < Pr(D = 1 \mid Z')$. The threshold crossing model of program participation implies that some individuals who would have chosen $D = 0$ with $Z = Z$ will instead choose $D = 1$ when $Z = Z'$, but no individual with $D = 1$ when $Z = Z$ would choose $D = 0$ when $Z = Z'$.\(^{37}\)

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

(i) The average treatment effect (ATE) is given by $\Delta_{ATE}(X) = E(\Delta \mid X = x)$.
(ii) The average effect of treatment on the treated, conditional on a value of $P(Z)$, is given by $\Delta_{TT}(X, P(Z), D = 1) = E(\Delta \mid X = x, P(z) = P(Z), D = 1)$.
(iii) The marginal treatment effect (MTE) conditions on a value of the unobservable: $MTE = \Delta_{MTE}(X) = E(\Delta \mid X = x, U_D = u)$.
(iv) The local average treatment effect (LATE) parameter is given by

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

The MTE is a new concept. 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 them most inclined to participate, while people with $U_{D_i} = 1$ have unobservables that make them the least inclined to participate.

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

$$\Delta_{TT}(X) = \int_0^{P(Z)} E(\Delta \mid X = x, U_D = u) \, dU_D,$$
$$\Delta_{TT}(X) = \int_0^{1} E(\Delta \mid X = x, U_D = u) \, dU_D,$$
$$\Delta_{LATE}(X, P(Z), P(Z')) = \frac{\int_{P(Z')}^{P(Z)} E(\Delta \mid X = x, U_D = u) \, dU_D}{P(Z) - P(Z')}.$$

\(^{37}\) As shown in \textit{Vytlacil (2002)}, the assumptions required to justify a threshold crossing model are the same as the monotonicity conditions typically assumed to justify application of LATE estimators, proposed in \textit{Imbens and Angrist (1994)}.
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. Knowledge of the MTE function therefore enables computation of all the other parameters of interest.

The MTE function depends on a value of an unobservables, raising the question of how to estimate the MTE function. Heckman and Vytlacil (2000) propose the 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

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

The $\Delta_{LIV}(X, P(Z))$ parameter can be obtained by first estimating the program participation (propensity score) model to get $\hat{P}(Z)$, and then estimating $\frac{\partial E(Y | P(z) = P(Z), X)}{\partial P(Z)}$ nonparametrically (which can be done by local linear regression). The first step can be carried out using a parametric, semiparametric or nonparametric estimator for the binary choice model. The second step 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 estimates $TT$, $ATE$, $LATE$ can then be obtained by integrating under different regions of the MTE function.

3.4.3. Applications

LIV estimators have only been recently developed, and there are so far no applications to evaluating effects of program interventions in developing country settings. For a recent application to estimating returns to education using US data, see Carneiro, 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

---

38 If the data are choice-based sampled, the choice-based sampling will have to be taken into account in both steps.
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. 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 identification 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 (2001) 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, Todd and Van der Klaauw (2001).

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

$$Y_i = Y_{0i} + D_i \cdot \Delta_i.$$  \hfill (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 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 $f(Z) \equiv E[D_i \mid Z_i = z] = \Pr[D_i = 1 \mid Z_i = z]$ is known to be discontinuous at $z_0$.\footnote{Other more recent applications of the regression-discontinuity methods include Van der Klaauw (1996), Angrist and Lavy (1999), and Black (1999).}

\footnote{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.}

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

3.5.1.1. 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 $z_0$:

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

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 $\epsilon > 0$ denote an arbitrary small number. Comparing conditional means for persons who received and did not receive treatment gives

$$E[Y_i \mid Z_i = z_0 + \epsilon] - E[Y_i \mid Z_i = z_0 - \epsilon] = E[\Delta_i \mid Z_i = z_0] + E[Y_{0i} \mid Z_i = z_0 + \epsilon] - E[Y_{0i} \mid Z_i = z_0 - \epsilon].$$

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

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

RD-2: The limit $\lim_{\epsilon \to 0^+} E[\Delta_i \mid Z_i = z_0 + \epsilon]$ is well-defined.

Under conditions (RD-1) and (RD-2),

$$\lim_{\epsilon \to 0^+} \{E[Y_i \mid Z_i = z_0 + \epsilon] - E[Y_i \mid Z_i = z_0 - \epsilon]\} = E[\Delta_i \mid z_0]. \quad (11)$$

By comparing persons arbitrarily close to the point $z_0$ who did and did not receive treatment, we can in the limit identify $E[\Delta_i \mid z_i = z_0]$, 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.

---

41 It is assumed that the density of $Z_i$ is positive in the neighborhood containing $z_0$. 
3.5.1.2. **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), $\Pr[D_i = 1 \mid Z_i]$, viewed as a function of $z_i$, is discontinuous at $z_0$. As shown in Hahn, Todd and Van der Klaauw (2001), mean treatment effects can be identified even under a fuzzy design under different some assumptions on the heterogeneity of impacts.

3.5.1.3. **Common treatment effects**  

Suppose that the treatment effect is constant across different individuals and is equal to $\Delta$. The mean difference in outcomes for persons above and below the discontinuity point $z_0$ is

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

Under (RD-1), we have

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

Thus, we can identify $\Delta$ by the ratio

$$
\lim_{e \to 0^+} \frac{E[y_i \mid z_i = z_0 + e] - E[y_i \mid z_i = z_0 - e]}{E[x_i \mid z_i = z_0 + e] - E[x_i \mid z_i = z_0 - e]}.
$$

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

3.5.1.4. **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 $\Delta_i$ conditional on $Z_i$ near $z_0$: $D_i \perp \Delta_i \mid Z_i = z_0$.

Then the same ratio identifies $E(\Delta_i \mid Z_i = z_0)$. In addition to the cases considered above, Hahn, Todd and Van der Klaauw (2001) also consider an alternative local average treatment effect (LATE) interpretation of the same ratio.⁴²

---

⁴² 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 $z_0$. 
We next describe an estimation approach proposed in Hahn, Todd and Van der Klaauw (2001). 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 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

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

and

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

The RD estimator can also be implemented using local linear regression methods, as proposed in Hahn, Todd and Van der Klaauw (2001), which have better performance than simple averaging methods or kernel methods at boundary points. (See Fan, 1992a.) For this problem, all the estimation points are boundary points.

3.5.3. Applications of RD methods

Regression-discontinuity methods have only occasionally been used in evaluating social programs in developing country settings. Buddlemeier and Skoufias (2004) study the performance of RD methods using data from the Mexican PROGRESA experiment. As discussed in Section 3.2.7, the PROGRESA program was a school and health subsidy program introduced by the Mexican government. The PROGRESA experiment randomized villages in and out of the program. Within each village, only families who

---

43 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 \( g(Z) = E[Y_0 | z_i] \) 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 \( E[D_i | Z_i] \). This estimation approach is consistent under correct specification but can be sensitive to misspecification.

44 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.

45 This experiment is described in detail in Chapter 62 of this handbook.
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.

Buddlemeier 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, alleviating concerns that households could have manipulated their poverty status to become eligible for the program.46 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. They find that the estimates based on the RD method are close to those derived from the experiment, lending credibility to the RD approach. Moreover, most of the households in their sample have eligibility scores near the cut-off values, making the sample near the cut-off an interesting subsample to study.

In another application of RD methods, 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) similarly 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 child test scores.

4. Ex ante program evaluation

Thus far, we have considered the problem of how to evaluate effects of existing programs, and all of the methods described in the previous sections assumed access to data on program participants. However, policy makers are sometimes interested in evaluating effects of hypothetical programs before deciding which type of program to implement. Some of the desired goals may be (i) to optimally design a social program to achieve some desired effects, (ii) to forecast the take-up rates and costs of alternative programs, or (iii) to study the effectiveness of alternatives to an existing program. Evaluating the effects of programs that do not yet exist requires an evaluation method that makes use

46 If households were selecting nonrandomly into the program around the cut-off, then this could invalidate the assumption RD-1.
only of data on people who have never participated in the program. Answering question (iii) requires a way of extrapolating from experience with an existing program to alternative programs.

The problem of forecasting the effects of social programs is part of the more general problem of assessing the effects of policy changes prior to their implementation that was considered in the early work of Marshak (1953). He described it as one of the most challenging problems facing empirical economists. In the early discrete choice literature, the problem was cast as the “forecast problem,” whereby researchers were trying to predict the demand for a good prior to its being introduced into the choice set. For example, McFadden and Talvitie (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 for forecasting program effects by comparing models’ forecasts to program effects that are estimated using experimental data. 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 actually implemented as a randomized experiment, and Wise (1985) is able to compare forecasts he obtains from alternative models to the subsidy effects estimated experimentally.

More recently, Todd and Wolpin (2006) 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 discussion of this program in Section 3.2.7) on choices about children’s schooling and work and on family fertility.47 The PROGRESA program was evaluated by an experiment that randomly assigned 506 villages to treatment and control groups. To assess the efficacy of the economic model for ex ante evaluation purposes, Todd and Wolpin (2006) estimate the model using data only on untreated individuals and then compare the model’s predictions about program impacts to those observed under the experiment. They find that the model provides relatively accurate ex ante forecasts of program effects on school enrollment and child work patterns. They then use the model to evaluate 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, the next section presents a simple model of schooling choice and shows how the effects of a school subsidy program can potentially be identified, even when none of the families in the data actually receive a subsidy. This example is drawn from Todd and Wolpin (2005).

47 The PROGRESA program is described in detail in Chapter 62 of this handbook.
4.1. An illustrative model of identification of subsidy effects

Consider a household making a single period decision about whether to send a child to school or to work. Let the utility of the household be separable in consumption \((C)\) and school attendance \((s)\), namely \(u = C + (\alpha + \epsilon)s\), where \(s = 1\) if the child attends school, \(= 0\) otherwise and \(\epsilon\) is a preference shock. Assume that the cost of attending school 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 \(\delta ks\) is the cost that is only incurred if the child attends school. Under utility maximization, the family chooses to have the child attend school if \(\epsilon > w - \alpha + \delta k\).

Suppose that wages are only on-served for children who work and specify a child wage 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
\[
    s = 1 \quad \text{if} \quad \alpha - \delta k + \epsilon > z\gamma + v, \quad \text{else} \quad s = 0.
\]

The probability that a child attends school can be written as
\[
    \Pr(s = 1 \mid z) = \Pr(\alpha - \delta k + \epsilon < v - \nu) = F_{\nu - v}(z\gamma - \alpha + \delta k),
\]
where \(F_{\nu - v}(\cdot)\) is the cdf of \(\epsilon - \nu\). Under an assumption that the median of \(\epsilon - \nu\) is 0 conditional on \(z\), the parameters \(\gamma, \alpha, \text{and} \delta\) can be identified up to scale and estimated by either a parametric or semiparametric discrete choice estimation method.\(^{48}\)

Next, consider estimation of the child wage offer equation using data on children who work \((s = 0)\) for whom wages are observed. We can write the wage equation as
\[
    y = z\gamma + E(v \mid z, s = 0) + \{v - E(v \mid z, s = 0)\},
\]
where the error term in brackets \((\nu = v - E(v \mid z, s = 0))\) has conditional mean zero by construction.

As described in Section 3.3, the parameter \(\gamma\) can be consistently estimated by including a control function \(E(v \mid z, s = 0)\).\(^{49}\) Under the assumption that (i) \(\nu\) and \(\epsilon\) are jointly distributed with density \(f(v, \epsilon)\) and (ii) the conditional density equals the unconditional density, \(f(v, \epsilon \mid z, k) = f(v, \epsilon)\), as described along in Section 3.3, we

\(^{48}\) See Manski (1975, 1988).

\(^{49}\) See Heckman (1979).
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 \( k \)), then the parameter \( \gamma \) can be nonparametrically identified under very weak assumptions on the \( K \) function.\(^{50}\) To see why, note that an exclusion restriction allows us to hold constant \( z \) at some value and vary \( P \), thereby tracing out the shape of the \( K \) function. Then, fixing \( K \) at some value, we can estimate \( \gamma \)\.\(^{51}\) Once the scale of \( \gamma \) is identified, we can use the results of the discrete choice estimation to obtain \( \alpha \) and \( \delta \)\.\(^{52}\)

Next, consider 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 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 

\[ F_v(z\gamma - \alpha - b + \delta k) - F_v(z\gamma - \alpha + \delta k). \]

The function \( F_v(s) \) can be estimated nonparametrically by a nonparametric regression of the school attendance indicator, \( s \), on \( z\gamma - \alpha + \delta k \).\(^{53}\) To assess the effect of the subsidy on the probability of attending school, we simply evaluate the \( F_v(s) \) function at the point \( z\gamma - \alpha - b + \delta k \).

The above described approach is semiparametric and does not require specifying the distribution, for example, of \( \varepsilon - v \) or the specification of the \( K \) function. Of course, a fully parametric approach would also be feasible.

5. Conclusions

This chapter has considered traditional and new approaches to evaluating the effects of treatment interventions using nonexperimental data. The major challenge to the evaluator is how to take account both observable and potentially nonobservable preexisting differences between program participants and nonparticipants in drawing inferences about the program’s causal effect. Differences between participants and nonparticipants may arise either because programs are nonrandomly placed, are targeted at certain groups, or because of individual nonrandom self-selected into the program. All three of these factors may occur simultaneously.

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

---

\(^{50}\) Assumptions on the continuity of the \( K \) function are required.

\(^{51}\) 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 \( \Pr(s = 0 \mid z, k) = 1 \). See Heckman (1990) and Andrews and Schafgans (1998).

\(^{52}\) Given an estimate of \( \gamma \), the scaling factor in the discrete choice problem can also be obtained.

\(^{53}\) Here, we can use the fact that the conditional expectation of \( s \), 

\[ E(s \mid z\gamma - \alpha + \delta k = \tau) = \Pr(s = 1 \mid z\gamma - \alpha + \delta k = \tau). \]
adopt in any particular circumstance will generally be context-specific and will also depend on the quality of the available data. For example, matching methods impose weak assumptions on the form of the conditional mean of the outcome equation, but 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. Difference-in-difference matching estimators overcome this strong assumption to a degree by allowing permanent, time-invariant unobservables to affect participation decisions. 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 is relatively straightforward, but a drawback to them is that they usually assume that error terms are normally distributed. Semiparametric control function estimators are more flexible, but usually require additional assumptions to achieve identification.

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 invoke weak assumptions, but can only provide information on treatment effects at the points of discontinuity, which may not be sufficient to evaluate the efficacy of a program.

Lastly, the evaluator also has access to a class of instrumental variables estimators, that can be applied in situations where there is a variable that affects the program participation decision but does not affect outcomes. When there is an exclusion restriction, one option is to apply the Wald IV estimator, which recovers the local average treatment effect (LATE) parameter (see Imbens and Angrist, 1994). A potential drawback of LATE is that it is instrument dependent and therefore may not correspond to the main parameter of interest. The most recent of the evaluation methods considered in this chapter are Local Instrumental Variable (LIV) estimators of Heckman and Vytlacil (2005) that can be applied in situations where there is a continuous instrumental variable (or at least one that is continuous over a range). LIV estimators provide a means of learning about the full distribution of treatment effects and of building up the various other parameters of interest, including LATE, treatment on the treated (TT), and average treatment effect (ATE).

The many recent developments in the evaluation literature have greatly increased the set of tools available for empirical evaluation researcher and have furthered the understanding of how different estimation methods relate to one another. It is hoped that by discussing the assumptions, data requirements, and implementation issues, this chapter has made the relative strengths and weaknesses of various approaches more transparent and thereby made the choice among estimators in any particular context easier.
Ch. 60: Evaluating Social Programs

References