An Introduction to Alternative Methods in Program Impact Evaluation

Nguyen Viet, Cuong

Development Economics Group, Wageningen University, Netherlands

1 December 2006

WORKING PAPER
MANSHOLT GRADUATE SCHOOL

An Introduction to Alternative Methods in Program Impact Evaluation

Nguyen Viet Cuong

DISCUSSION PAPER No. 33
2007

Mansholt Graduate School of Social Sciences

Hollandseweg 1, 6706 KN Wageningen,
The Netherlands
Phone: +31 317 48 41 26
Fax: +31 317 48 47 63
Internet: http://www.mansholt.wur.nl/
e-mail: office.mansholt@wur.nl

Working Papers are interim reports on work of Mansholt Graduate School (MGS) and have received only limited reviews[1]. Each paper is refereed by one member of the Editorial Board and one member outside the board. Views or opinions expressed in them do not necessarily represent those of the Mansholt Graduate School.

The Mansholt Graduate School's researchers are based in three departments: 'Social Sciences', 'Environmental Sciences' and 'Agrotechnology and Food sciences' and two institutes: 'LEI, Agricultural Economics Research Institute' and 'Alterra, Research Institute for the Green World'. In total Mansholt Graduate School comprises about 250 researchers.

Mansholt Graduate School is specialised in social scientific analyses of the rural areas and the agri- and food chains. The Graduate School is known for its disciplinary and interdisciplinary work on theoretical and empirical issues concerning the transformation of agriculture, rural areas and chains towards multifunctionality and sustainability.

Comments on the Working Papers are welcome and should be addressed directly to the author(s).

Nguyen Viet Cuong — Development Economics Group
De Leeuwenborch, Hollandseweg 1, 6706 KN Wageningen,
The Netherlands

[1] Working papers may have been submitted to other journals and have entered a journal's review process. Should the journal decide to publish the article the paper no longer will have the status of a Mansholt Working Paper and will be withdrawn from the Mansholt Graduate School's website. From then on a link will be made to the journal in question referring to the published work and its proper citation.

An Introduction to Alternative Methods in Program Impact Evaluation

Nguyen Viet Cuong[1]

December 2006[2]

Abstract

This paper presents an overview of several widely-used methods in program impact evaluation. In addition to a randomization-based method, these methods are categorized into: (i) methods assuming "selection on observable" and (ii) methods assuming "selection on unobservable".[3] The paper discusses each method under identification assumptions and estimation strategy. Identification assumptions are presented in a unified framework of counterfactual and two equation model. Finally, the paper uses simulated data to illustrate how these methods work under different identification assumptions.

Keywords: Program impact evaluation, treatment effect, counterfactual, potential outcomes, selection on observable, selection on unobservable.

[1] Ph.D. student in Development Economics Group of Wageningen University, the Netherlands. E-mail: c_nguyenviet@yahoo.com

[2] This paper is included in Ph.D. dissertation of Nguyen Viet Cuong under the supervision of Prof. David Bigman, Dr. Marrit Van Den Berg, and Prof. Vu Thieu. I am very grateful for their comments and correction on this paper.

[3] The names of "selection on observable" and "selection on unobservable"

1. Introduction

Impact evaluation of a program provides very helpful information for decisions as to whether the program should be terminated or expanded. If a program has no impact on its participants, it needs to be terminated or modified. Impact of a program on a subject is defined as the difference between its outcome with the program and its outcome without the program. However, for participants of the program, we can observe only their outcome in the program state, but not their outcome if they had not participated in the program. Similarly, for non-participants we can observe their outcome in the no-program state, but not the outcome in the program state. Outcomes that cannot be observed are called counterfactual. This problem is sometimes referred as a missing data problem, which causes the impact evaluation become difficult. Although it is impossible to measure the program impact for each subject (Heckman, et al., 1999), we can estimate an average of impact for a group of subjects. The main task is to estimate the average counterfactual outcomes. Bias in impact estimation is difference between counterfactual outcomes and their estimate. These biases can arise if there are concurrent factors that can affect the outcomes of participants and non-participants, and we are unable to net out impact of these factors from program impact.

This paper presents an overview of several widely-used methods in program impact evaluation. In addition to a randomization-based method in which participants are selected randomly, these methods are categorized into: (1) methods assuming "selection on observable", and (2) methods assuming "selection on unobservable". If impact of factors that can affect subjects is correlated with impact of a program of interest, we need to separate the program impact. "Selection on observable" methods are based on an assumption that we can observe all these correlated factors. In contrast, if we are not able to observe all the correlated factors, we need to resort "selection on unobservable" methods. The paper discusses each method under identification assumptions and estimation strategy. Identification assumptions are presented in a unified framework of counterfactual and two equation models. These assumptions will also be discussed in the context of traditional econometrics to illustrate how to estimate the program impact. In doing so, the paper aims to present the impact evaluation methods in a consistent framework of counterfactual, two equations model with a link to traditional econometrics regression.

The paper is structured into seven sections. Section 2 states problems in program impact evaluation including definition and difficulties in measuring the program impact. Section 3 illustrates how a method that is based on random selection solves these problems. Next, sections 4 and 5 introduce methods relying on selection of observables, and methods relying on selection of unobservable, respectively. Section 6 illustrates the methods using simulated data. Finally, section 7 presents some conclusions.

2. Problems in program impact evaluation

2.1. Counterfactual framework of program impact evaluation

The main objective of impact evaluation of a program is to assess the extent to which the program has changed outcomes for subjects. In other words, impact of the program on the subjects

is measured by the change in welfare outcome that is attributed only to the program. In the literature on impact evaluation, a broader term "treatment" is sometimes used instead of program/project to refer to intervention whose impact is evaluated.

To make the definition of impact evaluation more explicit, suppose that there is a program assigned to some people in a population P. For simplicity, let's assume that there is a single program, and denote by D the binary variable of participation in the program, i.e. $D = 1$ if she/he participates in the program, and $D = 0$ otherwise. D is also called the variable of treatment status. Further let Y denote the observed value of the outcome. This variable can receive two values depending on the participation variable, i.e. $Y = Y_1$ if $D = 1$, and $Y = Y_0$ if $D = 0$.[4] These outcomes are considered at a point or over a period of time after the program is implemented.

The impact of the program on the outcome of person i is measured by:

$$\Delta_i = Y_{i1} - Y_{i0}. \qquad (2.1)$$

It is equal to the difference in the outcome between the program state and the no-program state. The problem is that we cannot observe both terms in equation (2.1) for the same person. For those who participated in the program, we can observe only Y_1, and for those who did not participate in the program we can observe only Y_0. Outcomes that cannot be observed are called counterfactual.

It is practically impossible to estimate the program impact for each person (Heckman, et al., 1999), because we cannot know exactly the counterfactual outcome. If we do construct an estimator for individual effects, the associated standard error would be very large. In fact, program impact can be estimated for a group of people. In the literature on program impact evaluation, two popular parameters are the Average Treatment Effect (ATE), and the Average Treatment Effect on the Treated (ATT) [5].

ATT is the expected impact of the program on a person who is randomly selected and assigned to the program. It is defined as:

$$ATE = E(\Delta) = E(Y_1 - Y_0) = E(Y_1) - E(Y_0). \qquad (2.2)$$

This is the traditional average partial effect (APE) in econometrics. To see this, let's write the observed outcome in a switching model (Quandt, 1972):

$$Y = DY_1 + (1 - D)Y_0, \qquad (2.3)$$

where Y is observed outcome, which is equal to Y_1 and Y_0 for participants and non-participants, respectively.

Then,

[4] Y can be a vector of outcomes, but for simplicity let's consider a single outcome of interest.
[5] There are other parameters such as local average treatment effect, marginal treatment effect, or even effect of "non-treatment on non-treated" which measures what impact the program would have on the non-participants if they had participated in the program, etc.

$$APE = E(Y \mid D=1) - E(Y \mid D=0) = E(Y_1) - E(Y_0) = ATE. \tag{2.4}$$

Most programs are targeted to certain subjects. The important question is the program impact on those who participated in the program. If the program has positive impact, policy makers would be interested in expanding the program for similar groups. The expected treatment effect on the participants is equal to:

$$ATT = E(\Delta \mid D=1) = E(Y_1 - Y_0 \mid D=1) = E(Y_1 \mid D=1) - E(Y_0 \mid D=1). \tag{2.5}$$

Except for the case of randomized programs that is discussed in section 3, ATE and ATT are, in general, different from each other, since program participation often depends on the potential outcomes, and as a result $E(Y_1) \neq E(Y_1 \mid D=1)$, and $E(Y_0) \neq E(Y_0 \mid D=1)$.

Estimation of the two parameters is not straightforward, since there are some components that cannot be observed directly. Equation (2.2) can be rewritten as:

$$\begin{aligned} ATE = E(Y_1) - E(Y_0) &= \left[E(Y_1 \mid D=1)\,Pr(D=1) + E(Y_1 \mid D=0)\,Pr(D=0)\right] \\ &\quad - \left[E(Y_0 \mid D=1)\,Pr(D=1) + E(Y_0 \mid D=0)\,Pr(D=0)\right] \\ &= \left\{\left[E(Y_1 \mid D=1) - E(Y_0 \mid D=1)\right]Pr(D=1)\right\} \\ &\quad + \left\{\left[E(Y_1 \mid D=0) - E(Y_0 \mid D=0)\right]Pr(D=0)\right\}, \end{aligned} \tag{2.6}$$

where $Pr(D=1)$ and $Pr(D=0)$ are the proportions of participants and non-participants of the program, respectively.

Define the average treatment effect on the non-treated (ATNT) as:

$$ANTT = E(Y_1 \mid D=0) - E(Y_0 \mid D=0). \tag{2.7}$$

This parameter can be explained as the effect that the non-participants would have gained if they had participated in the program.

Then, ATE can be written as follows:

$$ATE = ATT\,Pr(D=1) + ATNT\,Pr(D=0). \tag{2.8}$$

In (2.6) what we can observe are the mean outcomes of participants and non-participants. As a result, the terms $E(Y_1 \mid D=1)$ and $E(Y_0 \mid D=0)$ can be estimated directly. However, the counterfactual terms $E(Y_1 \mid D=0)$ and $E(Y_0 \mid D=1)$ are not observed and cannot be estimated directly. $E(Y_1 \mid D=0)$ is the expected outcome of the participants had they not participated in the program, while $E(Y_0 \mid D=1)$ is the expected outcome of non-participants had they participated in the program. Thus the estimation of ATE and ATT is not straightforward, and different methods which are discussed in this study estimate ATE and ATT under certain assumptions on how the program is assigned to the population and how the outcome is determined.

Note that we can allow program impact to vary across a vector of observed variables, X, since we might be interested in the program impact on certain groups that are specified by the characteristics, X. The so-called conditional parameters are expressed as follows:

$$ATE_{(X)} = E(\Delta \mid X) = E(Y_1 \mid X) - E(Y_0 \mid X), \tag{2.9}$$

and

$$ATT_{(X)} = E(\Delta \mid X, D = 1) = E(Y_1 \mid X, D = 1) - E(Y_0 \mid X, D = 1). \tag{2.10}$$

If we denote by $ATNT_{(X)}$ the ATNT conditional on X:

$$ATT_{(X)} = E(\Delta \mid X, D = 0) = E(Y_1 \mid X, D = 0) - E(Y_0 \mid X, D = 0), \tag{2.11}$$

then, similar to (2.8):

$$ATE_{(X)} = ATT_{(X)} Pr(D = 1 \mid X) + ATNT_{(X)} Pr(D = 0 \mid X), \tag{2.12}$$

where $Pr(D = 1 \mid X)$ and $Pr(D = 0 \mid X)$ are the proportion of the participants and non-participants given the X variables, respectively.

In the following discussion, we will focus on the conditional parameters, $ATE_{(X)}$ and $ATT_{(X)}$, since if they are identified the unconditional parameters, ATE and ATT can be also identified as a results:

$$ATE = \int_X ATE_{(X)} dF(X), \tag{2.13}$$

$$ATT = \int_{X|D=1} ATT_{(X)} dF(X \mid D = 1). \tag{2.14}$$

2.2. Econometric framework of program impact evaluation

As mentioned, the selection of a method to estimate ATE and ATT for a program depends crucially on assumptions on how people are selected in the program as well as how the potential outcomes are affected by the program and other factors. Although assumptions are often not tested, they need to be stated explicitly so that one can know when impact evaluation results using a method are valid and robust. A popular way to discuss assumptions on the program in impact evaluation is to use a model of two outcome equations of Roy (1951) or Rubin (1974), in which potential outcomes Y_0 and Y_1 are expressed as functions of individual characteristics (conditioning variables), X:[6]

$$Y_0 = \alpha_0 + X\beta_0 + \varepsilon_0 \tag{2.15}$$

$$Y_1 = \alpha_1 + X\beta_1 + \varepsilon_1 \tag{2.16}$$

[6] For simplicity, subscript i is dropped.

Y_0 and Y_1 can be any functions of X, not necessarily linearly or parametrically specified, and all the identification strategies presented in this paper are still valid. However, to illustrate ideas and links with the traditional linear regression framework, we assume this linearity. The reason for writing this two equation model is to discuss assumptions that different methods rely on in terms of behavior of participants versus non-participants.

For simplicity and identification of program impact in some parametric regressions, we require X to be exogenous in the potential outcome equations.

Assumption 2.1: $E(\varepsilon_0 \mid X) = E(\varepsilon_1 \mid X) = 0$ (A.2.1)

To this end, two additional assumptions are needed for the validity of the micro-approach of program impact evaluation. The first assumption is common in partial equilibrium approach, and required in the literature on program impact evaluation. This assumption is called the stable unit treatment assumption (SUTVA).

Assumption 2.2: $Y_i \perp D_j \ \forall i, j$, i.e., realized (observed) outcome of individual i, Y_i, is independent of program status of individual j, D_j. (A.2.2)

This assumption implies that there is no spill-over effect of the program. In other words, an individual's participation in the program does not affect the outcome of other people. For programs that cover a large proportion of population, this assumption can be violated. For example, if a large number of farmers receive preferential credit, they can reduce production costs and increase their market share, which can affect the revenue of farmers who do not receive a similar credit.[7] When the assumption does not hold, one needs to use general equilibrium analysis.[8]

The second assumption is implied in the two equation model. Writing the same X variables in two equations (2.15) and (2.16) means that for each person the status of program participation (treatment status) does not affect X. Formally speaking, once conditional on potential outcomes, X are independent of D.

Assumption 2.3:[9] $X \perp D \mid Y_0, Y_1$ (A.2.3)

The assumption does not mean that X is uncorrelated with D, but means that X is uncorrelated with D given the potential outcomes. Under this assumption D does not affect X once conditioning on the potential outcomes. Although this assumption is not an indispensable condition to identify the program impact, it is maintained for simplicity. If D affects X, it will be much more complex to capture truly the impact of program. The program impact on the outcome can go through the conditioning variables, and we need to model the program impact on X to get the overall impact of the program on the outcome. As a result, we need to solve the problem of program impact not only

[7] For other examples on the violation of this assumption, see, e.g., Heckman, et al. (1999), and Rubin (1978)
[8] For more detailed discussion on general equilibrium approach in impact evaluation, see, e.g., Heckman, et al. (1999), and Heckman, et al. (1998b)
[9] Another expression for conditional independence $f(X \mid Y_0, Y_1) = f(X \mid D, Y_0, Y_1)$, where f(.) is conditional density of X. For discussion on conditional independence, see, e.g., Dawid (1979).

on outcome but also on X.[10] Thus assumption 2.3 is often made in the literature on impact evaluation. In the following discussions of different methods in impact evaluation, assumptions 2.2 and 2.3 are implicitly assumed to hold.

In this framework, the interested parameters in impact evaluation are expressed as follows:

$$\begin{aligned} ATE_{(X)} &= E(Y_1 \mid X) - E(Y_0 \mid X) \\ &= E[\alpha_1 + X\beta_1 + \varepsilon_1 \mid X] - E[\alpha_0 + X\beta_0 + \varepsilon_0 \mid X] \\ &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) \end{aligned} \tag{2.17}$$

and,

$$\begin{aligned} ATT_{(X)} &= E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=1) \\ &= E[\alpha_1 + X\beta_1 + \varepsilon_1 \mid X, D=1] - E[\alpha_0 + X\beta_0 + \varepsilon_0 \mid X, D=1] \\ &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) + E(\varepsilon_1 - \varepsilon_0 \mid X, D=1). \end{aligned} \tag{2.18}$$

It should be noted that even if coefficients $\alpha_0, \alpha_1, \beta_0, \beta_1$ can be estimated, $ATT_{(X)}$ still includes an unobservable term $E(\varepsilon_1 - \varepsilon_0 \mid X, D=1)$, while $ATE_{(X)}$ does not. To identify $ATT_{(X)}$ we need the following additional assumption:

Assumption 2.4: $E(\varepsilon_0 \mid X, D=1) = E(\varepsilon_1 \mid X, D=1)$ (A.2.4)

This assumption states that given X, the expectation of the unobserved variables for the participants is the same regardless of the program so that the unobserved term in (2.18) vanishes. It is worth noting that assumption (A.2.4) does not mean the expectation of the error terms conditional on all the X variables. Instead, this assumption is required for some variables of X that we are interested in the conditional parameters. There might be many explanatory variables X, but we are often interested in $ATE_{(X)}$ and $ATT_{(X)}$ conditional on a certain number of variables in X, not all X.

To link the counterfactual data with the observed data, substitute (2.15) and (2.16) into the switching model in (2.3). This results in:

$$\begin{aligned} Y &= D(\alpha_1 + X\beta_1 + \varepsilon_1) + (1-D)(\alpha_0 + X\beta_0 + \varepsilon_0) \\ &= \alpha_0 + \beta_0 X + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) + (\varepsilon_1 - \varepsilon_0)] + \varepsilon_0. \end{aligned} \tag{2.19}$$

Equation (2.19) is a very general model of program impact, in which the program impact measured by the coefficient of variable D varies across subjects. This coefficient depends on both

[10] Specifically, let's assume the linear equations:
$Y = X\beta + D\gamma + \varepsilon,$
$X = X\delta + v,$
then $Y = D(\beta\delta + \gamma) + (\varepsilon + \beta v).$
To identify the program impact, $\beta\delta + \gamma$, using these equations, D must be exogenous in both equations of Y and X.

observable and unobservable variables, X and ε. It can also be correlated with D if D is correlated with X and ε. This is a random coefficient model in which the coefficient is correlated with observed and unobserved characteristics variables.

Since we are unable to estimate the unobserved term in (2.19), we often invoke assumption (A.2.4) to identify $ATE_{(X)}$ and $ATE_{(X)}$. The remaining problem is how to estimate $\alpha_0, \alpha_1, \beta_0, \beta_1$ without bias. The error term in (2.19) is required to have conventional property:

$$E[(\varepsilon_1 - \varepsilon_0)D + \varepsilon_0 \mid X, D] = 0. \quad (2.20)$$

To complete this section, a model of program participation is introduced. The participation of a person in the program can depend on selection criteria of the program and own decisions of the person. The program participation model is often expressed in a latent index framework:

$$D^* = \theta W + v, \quad (2.21)$$

$$D = 1 \; if \; D^* > 0,$$

$$D = 0 \text{ otherwise,}$$

where D^* is the latent index of the program selection that is correlated with observable variables, W and unobservable term, v. W and v are all the variables that can affect the program participation of subjects.

2.3. Determinations of program impact

The objective of the impact evaluation of a program is to measure the size of program impact on outcome. The magnitude of a program impact on a subject's outcome is given in equation (2.1):

$$\Delta_i = Y_{i1} - Y_{i0}.$$

This realized magnitude depends on many factors, but in general these factors can be grouped into 3 groups: intervention of the program, time to conduct program evaluation, and the characteristics of the subject i.

Obviously, magnitude of program impact depends on what the program offers to the subject. Any change in the design of intervention can lead to a change in the program impact. For example, a vocational training provides courses in two ways: courses in morning and course in evening. For a given person, the impact of participating in morning courses can be higher than the impact of participating in evening courses, since her learning ability is better in mornings. Another example is a program of micro-credit that provides a small amount of credit for a targeted group of people. Eligible people, who meet the conditions for borrowing, can receive two specific amounts of credit, say C_1 and C_2, depending on their demand for credit. It is obvious that impact of receiving C_1 credit is different from impact of receiving C_2 credit. It should be noted that a program that provides different designs of intervention is sometimes understood as different mutually exclusive programs. For the example of the micro-credit program, a person would have

three potential outcomes $(Y_0, Y_{1C_1}, Y_{1C_2})$ that correspond to the state of no credit, the state of receiving C_1 credit, and the state of receiving C_2 credit, respectively. In this case, impact of the program on a person's outcome needs to be defined explicitly as impact of receiving C_1 credit or impact of receiving C_2 credit.

The second factor that affects the measured impact of a program is when the data on outcome are collected[11]. The definition of potential outcomes is made after the program implementation. Suppose that the program started at time t = 0. After that, the potential outcomes for a subject i at any point of time t (t>0) are Y_{i0t}, Y_{i1t}, and the program impact is:

$$\Delta_{it} = Y_{i1t} - Y_{i0t}.$$

The impact is now a function of t, thereby depending on time t. The result from impact evaluation conducted one year after program implementation can be different from the results of impact evaluation conducted two years after implementation. It is possible to assume that the program impact can be stable after a period of time, i.e., the program can move the outcome level (in the no-program state) to a new level of outcome (in the program state) in the long term. In the literature on impact evaluation, the argument "t" is often dropped for simplicity.

Thirdly, the impact of a program on a subject depends on her own characteristics. Different people will gain different benefits from a program. If the two equations of potential outcomes (2.15) and (2.16) are used, program impact can be expressed as a function of observable and unobservable variables, X and ε, respectively:

$$\Delta_i = Y_{1i} - Y_{0i} = (\alpha_1 - \alpha_0) + X_i(\beta_1 - \beta_0) + (\varepsilon_{1i} - \varepsilon_{0i}). \quad (2.22)$$

Unless two subjects have the same characteristics, i.e., the X variables and ε in the potential outcome equations, the program impact differs between them.

Similarly, the magnitude of ATE and ATT also depends on the three above factors. In the case of ATT, its magnitude depends on characteristics of subjects who actually take part in the program. The program design plays a very important role in affecting the magnitude of ATT, since different designs of a program mean different program selections, thereby different groups of participants. The magnitude of ATT will differ for various groups of participants because of different characteristics.

3. Method based on randomization design

3.1. Impact measurement of randomized program

In the impact evaluation literature, the ideal situation is that a program is assigned randomly to subjects, and those who are assigned the program are willing to participate in the program. The non-participants will form the control group, and do not participate in similar

[11] Or similarly, the measured program impact on a subject depends on when the subject participates in the program.

programs. In this case, the program assignment D is said to be independent of the potential outcomes Y_0 and Y_1. We can state this condition as an assumption.

Assumption 3.1: $Y_0, Y_1 \perp D$ (A.3.1)

Proposition 3.1: $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified under assumption (A.3.1).

Proof: As a result of assumption (A.3.1), conditional on D the value of the potential outcomes does not alter.

$$E(Y_1 \mid X) = E(Y_1 \mid X, D=1) = E(Y_1 \mid X, D=0), \tag{3.1}$$

$$E(Y_0 \mid X) = E(Y_0 \mid X, D=1) = E(Y_0 \mid X, D=0). \tag{3.2}$$

Hence:

$$\begin{aligned} ATE_{(X)} &= E(Y_1 \mid X) - E(Y_0 \mid X) \\ &= E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=0), \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} ATT_{(X)} &= E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=1) \\ &= E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=0). \end{aligned} \tag{3.4}$$

Thus, $ATE_{(X)} = ATT_{(X)}$. Since it is possible to observe all terms in $ATE_{(X)}$ and $ATT_{(X)}$, these parameters are identified in the case of randomization. ■

The program impact is estimated simply by comparing the mean outcome between the participants and non-participants.

When we have post-program data from a representative sample on participants and non-participants in a randomized program, we can use sample mean of outcomes for treatment and control group to estimate ATE, ATT, and their conditional version $ATE_{(X)}$ and $ATT_{(X)}$. Another way to estimate the program impact is to use the regression model. In the framework of two equation models, assumption (A.3.1) leads to:

$$D \perp \varepsilon_0, \varepsilon_1. \tag{3.5}$$

Note that assumption (A.3.1) results in (3.5), but the reverse does not hold if D is correlated with X. In general, (A.3.1) is stronger than (3.5).

In order to get unbiased estimators of $ATE_{(X)}$ and $ATT_{(X)}$ using regression, we need the assumption on exogeneity of X, i.e. (A.2.1).

Proposition 3.2: Under assumptions (A.3.1) and (A.2.1), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT can be estimated unbiasedly by OLS regression.

Proof: Under (A.3.1) and (A.2.1), $ATE_{(X)}$ and $ATT_{(X)}$ are the same:

$$ATE_{(X)} = ATT_{(X)} = (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0). \quad (3.6)$$

In which, coefficients can be estimated without bias from the equation:

$$Y = \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] + [D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0]. \quad (3.7)$$

Since the error term has the following property:

$$E[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0 \mid X, D] = DE(\varepsilon_1 - \varepsilon_0 \mid X, D) + E(\varepsilon_0 \mid X, D) = E(\varepsilon_0 \mid X) = 0. \quad (3.8)$$

Thus the estimator of the parameters is:

$$\hat{\text{ATE}}_{(X)} = \hat{\text{ATT}}_{(X)} = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)X \quad (3.9)$$

Once the conditional parameters are identified, the unconditional parameters are also identified because of (2.13) and (2.14).■

In short, when a program is assigned randomly to people, the program impacts are estimated directly using sample means of outcomes without further assumption.[12] However if the linear regression model is used, we need assumption on exogeneity of X to estimate $ATE_{(X)}$ and $ATT_{(X)}$.

3.2. Program impact evaluation under experiment

In reality, we are often interested in impact of a program that is targeted at specific subjects. For example, poverty reduction programs aim to provide the poor with support to get rid of poverty. Vocational training programs are targeted at the unemployed. The program is not assigned randomly to people in the population. In this case, experimental designs can be used to evaluate the impact of the targeted program.

A randomization design or experiment is conducted by choosing a group of people who are willing to participate in the experiment. Denote by D^* the variable indicating the experiment participation. $D^* = 1$ for those in the experiment, and $D^* = 0$ otherwise. Among people with $D^* = 1$, we randomly select people for program participation. Denote R as a variable that $R = 1$ for the participants, and $R = 0$ for non-participants in the experiment. The participants are called the treatment group, while the non-participants (among those in the experiment) are called the control group (or comparison group).

The randomization of program among those in the experiment is stated formally as follows:

Assumption 3.2:[13] $Y_0, Y_1 \perp R \mid D^* = 1$ (A.3.2)

[12] Assumption (A.2.1) is made for all methods in impact evaluation.

[13] Assumption (A.3.2) states that the selection of participants among the experimental people is independent of the potential outcomes. In fact we only need a weaker version to identify ATT:
$E(Y_1 \mid D^* = 1, R = 1) = E(Y_1 \mid D^* = 1, R = 1)$ and $E(Y_0 \mid D^* = 1) = E(Y_0 \mid D^* = 1, R = 0)$.

To estimate both $ATE_{(X)}$ and $ATT_{(X)}$, we need an additional assumption:

Assumption 3.3: $$E(Y_1 \mid X, D=0) = E(Y_1 \mid X, D=1) = E(Y_1 \mid X, D^* = 1)$$

$$E(Y_0 \mid X, D=0) = E(Y_0 \mid X, D=1) = E(Y_0 \mid X, D^* = 1) \qquad \text{(A.3.3)}$$

That is, once conditional on X, the expected outcome of those in the experiment is the same as the expected outcome of those not participating in the experiment. It is implied that people who participate in the experiment are similar to those in the reality once conditional on X.

Proposition 3.3: $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified under assumptions (A.3.2) and (A.3.3).

Proof:

Under (A.3.2) and (A.3.3), ATT is identified:

$$\begin{aligned} ATT_{(X)} &= E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=1) \\ &= E(Y_1 \mid X, D^* = 1) - E(Y_0 \mid X, D^* = 1) \\ &= E(Y_1 \mid X, D^* = 1, R=1) - E(Y_0 \mid X, D^* = 1, R=0), \end{aligned} \qquad (3.10)$$

and similarly, the average treatment effect on the non-treated (ATNT) is the same:

$$\begin{aligned} ATNT_{(X)} &= E(Y_1 \mid X, D=0) - E(Y_0 \mid X, D=0) \\ &= E(Y_1 \mid X, D^* = 1) - E(Y_0 \mid X, D^* = 1) \\ &= E(Y_1 \mid X, D^* = 1, R=1) - E(Y_0 \mid X, D^* = 1, R=0). \end{aligned} \qquad (3.11)$$

Thus, the $ATE_{(X)}$ is identified and the same as $ATT_{(X)}$ due to (2.12).■

As a result, (3.10) is the unbiased estimator of $ATT_{(X)}$ and $ATE_{(X)}$. We simple calculate the difference in the mean outcome between the participants and non-participants of the program among those attending the experiment. Once the conditional parameters are identified, the conditional parameters are also identified because of (2.13) and (2.14).

3.3. Advantages and disadvantages of the method based on randomization

There is no controversy that among methods of program impact evaluation, the method that is based on randomization of the program produces the most reliable results. When the randomization of a program is properly conducted, the average impact of the program is identified without any further assumptions. The randomization of programs ensures that there is no systematic difference in both observable and unobservable characteristics between treatment and control groups. As a result, any difference in the average outcome between these groups can be attributed to the program effect. The estimation of the program is also very simple given available data on the control and treatment groups. An another advantage of the method is the easiness in explaining its results of impact evaluation to program designers and policy makers, who often

---

However this assumption is difficult to interpreter. Thus we mention the assumption (A.3.2) in discussing the identification of the program impact.

have not much knowledge on statistics and econometrics. The idea of the method is very straightforward. Once policymakers understand the method, they will believe the results of the impact evaluation. Meanwhile, it is much more difficult for them to understand other methods that rely on complicated mathematics. The difficulty can lead them to doubt the results of impact evaluation.

The method based on randomized program, however, suffers from several drawbacks. Firstly, it is hardly to randomize a program which is targeted at a specific group due to issues of ethics and politics. Randomization of a program means exclusion of some eligible people from the program. It is unfair to deny (or delay) a program that provides supports such as health care or education for some eligible people. Policy makers will be criticized if they cannot explain why some eligible people are not allowed to participate in programs. Nevertheless, the randomization of a program can be conducted if the fund for the program is not sufficient to cover all eligible people. Some people have to participate in later periods, and they can serve as the control group for those who participate at the beginning.

Secondly, the implementation of impact evaluation of a socioeconomic program that is based on randomization design is often expensive. Subjects are scattered in the population, which increases the cost of program administration and data collection for impact evaluation.

Thirdly, there can be some factors in addition to the program that might bias estimates from the randomization-based evaluation. These factors can invalidate the key identification assumption (A.3.1), $D \perp Y_0, Y_1$. Two problems that are widely mentioned in a randomized program are attrition and substitution effects. Attrition means that some people in the treatment group quit the program during implementation. As a result, their observed outcome is not the potential outcome in the presence of the program, Y_1. If this drop-out is random, there is no concern about this problem since the randomization feature remains preserved.[14] If the attrition is not random but correlated with some characteristics of the drop-outs, the remaining subjects in the treatment group who actually take the program will be systematically different from the subjects in the control group. In other words, there is self-selection into the program of the participants. Estimation of program impact is not straightforward, and alternative methods that will be discussed in next sections are devised to deal with this self-selection. In this case, the mean difference in outcome between the treatment and control group is not an estimator of the program impact, but an estimator of "the mean effect of the offer of treatment" (Heckman, et al., 1999). However, if we expect that program impact is negligible for the drop-outs, we can measure the ATE by this mean difference. This is because the drop-outs are not interested in the program, and there would be no impact on them if we force them to follow the whole program.[15]

Substitution effect means that some people in the control group might try to get substituted programs that are similar to the program to be evaluated. The substituted programs can

[14] Of course, the remaining subjects in the control group should be statistical representative for the studied population.

[15] For example, a program that provides households with a certain amount of no-interest credit for a period is carried out. For those who refuse to receive credit, if we do force them to accept credit they can keep the credit unused and return it at the end of the period. In this case, the impact of program on those people is zero.

contaminate the outcome of the control group. It is implied that if the program had not been implemented, the participants would have taken other similar programs. The mean difference in outcome between the control and treatment groups reflects "the mean incremental effect of the program relative to the world in which it does not exist" (Heckman, et al., 1999). To capture truly the program impact, we need to know information on impacts of the substituted programs, and subtract them from the outcome of the control group to estimate the potential outcome of the treatment group in the absence of the program.[16]

Finally, a randomized program that is used for impact-evaluation purposes is often a pilot one, and impact of the pilot program can be far from the impact of the program when it is implemented in reality. A pilot program is often smaller and more easily administered. In addition, people involved in a pilot program including the administrators, control and treatment group, may follow the program rules more strictly if they know the program is pilot.

4. Methods assuming selection on observable

4.1. Selection bias and conditional independence assumption

As mentioned, randomization of a program is not always the case in reality. When a program is not assigned randomly, the potential outcomes of the participants will be different from those of non-participants. Assumption (A.3.1) no longer holds, and simple comparison of mean outcomes between participants and non-participants does not produce unbiased estimators of the program impacts. Bias in these estimators is called the selection bias in the literature on impact evaluation.

To see the selection bias in estimating the average treatment effect $ATE_{(X)}$ conditioning on X , rewrite the formula of $ATE_{(X)}$:

$$\begin{aligned} ATE_{(X)} &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) + E(\varepsilon_1 - \varepsilon_0 \mid X) \\ &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) \\ &+ \{[Pr(D=1 \mid X)E(\varepsilon_1 \mid X, D=1) + Pr(D=0 \mid X)E(\varepsilon_1 \mid X, D=0)] \\ &\quad - [Pr(D=1 \mid X)E(\varepsilon_0 \mid X, D=1) + Pr(D=0 \mid X)E(\varepsilon_0 \mid X, D=0)]\}. \end{aligned} \tag{4.1}$$

We use the following estimator:

$$\begin{aligned} \hat{ATE}_{(X)} &= E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=0) \\ &= E(\alpha_1 + X\beta_1 + \varepsilon_1 \mid X, D=1) - E(\alpha_0 + X\beta_0 + \varepsilon_0 \mid X, D=0) \\ &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) + E(\varepsilon_1 \mid X, D=1) - E(\varepsilon_0 \mid X, D=0). \end{aligned} \tag{4.2}$$

Thus the bias is equal to:

[16] For more discussion on the drop-out and substitution, see, e.g., Heckman et al. (2000).

$$
\begin{aligned}
\hat{ATE}_{(X)} - ATE_{(X)} &= \left[E(\varepsilon_1 \mid X, D=1) - E(\varepsilon_0 \mid X, D=0)\right] \\
&\quad - \{\left[Pr(D=1 \mid X)E(\varepsilon_1 \mid X, D=1) + Pr(D=0 \mid X)E(\varepsilon_1 \mid X, D=0)\right] \\
&\quad - \left[Pr(D=1 \mid X)E(\varepsilon_0 \mid X, D=1) + Pr(D=0 \mid X)E(\varepsilon_0 \mid X, D=0)\right]\} \qquad (4.3) \\
&= Pr(D=0 \mid X)\left[E(\varepsilon_1 \mid X, D=1) - E(\varepsilon_1 \mid X, D=0)\right] \\
&\quad + Pr(D=1 \mid X)\left[E(\varepsilon_0 \mid X, D=1) - E(\varepsilon_0 \mid X, D=0)\right].
\end{aligned}
$$

Even though X are controlled for, the selection bias in estimating $ATE_{(X)}$ can arise if the conditional expectation of unobserved variables in potential outcomes, ε_0 and ε_1, is different for the participants and non-participants.

Similarly, if we use the same estimator in (4.2) for $ATT_{(X)}$, the selection bias will be:

$$
\hat{ATT}_{(X)} - ATT_{(X)} = E(\varepsilon_0 \mid X, D=1) - E(\varepsilon_0 \mid X, D=0) \qquad (4.4)
$$

The selection bias stems from the difference in the conditional expectation of unobserved variables, ε_0, between the participants and non-participants.[17]

One intuitive way to avoid the selection biases, (4.3) and (4.4), in estimating $ATE_{(X)}$ and $ATE_{(X)}$ is to invoke assumptions so that the selection biases are equal to zero. The assumption on "selection on observable" assumes that one is able to observe all variables that affect both the program selection and potential outcomes so that once conditioned on these variables, the potential outcomes Y_0 and Y_1 are independent of the program assignment. In Rosenbaum and Rubin (1983), this assumption is called ignorability of treatment or conditional independence. Formally, it is written as:

Assumption 4.1: $Y_0, Y_1 \perp D \mid X$ (A.4.1)

Assumption (A.4.1) can be considered as a conditional version of assumption (A.3.1). Once we have control for X, the assignment of the program becomes randomized. Actually, we just need a weaker form of (A.4.1) in order to identify the program impact parameters.

Assumption 4.1': $E(Y_0 \mid X, D) = E(Y_0 \mid X)$

$$
E(Y_1 \mid X, D) = E(Y_1 \mid X) \qquad (A.4.1')
$$

This is called the conditional mean independence assumption. It is weaker than (A.4.1) in sense that (A.4.1) implies (A.4.1') but the reverse is not correct. Although assumption (A.4.1') is weaker and enough to identify the program impacts, it seems difficult to think whether it holds in

[17] If one has data before and after a program, they sometimes use the before and after estimator to estimate the program impact. The bias is equal to $E(Y_{0B} \mid D=1) - E(Y_{0A} \mid D=1)$, where $E(Y_{0B} \mid D=1)$ and $E(Y_{0A} \mid D=1)$ are the expectation of participants' outcome in the state of no program before and after the program, respectively. The assumption is valid if there is no change in the participants' outcome during the program implementation if they had not participated. Intuitively, this assumption seems plausible in short time, but might be unreasonable in long time.

reality since it involves the expectation terms. Thus, we will use assumption (A.4.1) in discussion of program impact evaluation.

A corollary of assumption (A.4.1) is that the error terms in the potential outcomes is also independent of D given X, i.e.:

$$\varepsilon_0, \varepsilon_1 \perp D \mid X . \tag{4.5}$$

Under condition (4.5), we have (Dawid, 1979):

$$E(\varepsilon_0 \mid X, D = 0) = E(\varepsilon_0 \mid X, D = 1), \tag{4.6}$$

$$E(\varepsilon_1 \mid X, D = 0) = E(\varepsilon_1 \mid X, D = 1). \tag{4.7}$$

As a result of equation (4.6) and (4.7), the selection biases given in (4.3) and (4.4) are equal to zero. $ATE_{(X)}$ and $ATT_{(X)}$ are identified, and so are ATE and ATT.

In addition, assumption (4.5) results in:

$$E(\varepsilon_1 - \varepsilon_0 \mid X, D = 1) = E(\varepsilon_1 - \varepsilon_0 \mid X) \tag{4.8}$$

Due to (4.8), $ATE_{(X)}$ is equal to $ATT_{(X)}$.

Finally, it should be noted that writing the same X variables in assumption (A.4.1) as in the potential outcome equations is just for the purpose of convenience. Actually, we should write variables Z. Conditional on these variables program assignment is independent of the potential outcomes, *i.e.*:

$$Y_0, Y_1 \perp D \mid Z . \tag{4.9}$$

To identify the program impact by this approach, Z must be observed and included in X . In terms of equation (2.20), $D^* = \gamma W + v$, Z must be included in W. The terminology "selection on observable" does not mean that we have to observe all information on the program selection, i.e. D is deterministic, but it implies that all the Z variables that make D correlated with Y_0 and Y_1 are observed. Put differently, the unobserved variables v are required uncorrelated with Y_0 and Y_1 given Z.

Assumption (A.4.1) is the key assumption for identifying program impacts that several methods rely on. Thus the methods are called methods based on selection on observables. Three widely-used methods are presented in this paper, namely regression, matching method, regression discontinuity.[18] All these methods can be conducted using single cross section data.

4.2. Regression methods

4.2.1. Linear regression

[18] In this section, "regression methods" mean those based on the assumption "selection on unobservables".

For simplicity we maintain the assumption of linearity in outcome equations for this section. Next we will discuss the case of nonlinear functions of potential outcomes.

As mentioned, under assumption (A.4.1) $ATE_{(X)}$ and $ATT_{(X)}$ are the same. Equation (2.17) becomes equation (2.16) as follows:

$$Y = \alpha_0 + X\beta_0 + D\left[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) + E(\varepsilon_1 - \varepsilon_0 \mid X)\right] + \left\{D\left[(\varepsilon_1 - \varepsilon_0) - E(\varepsilon_1 - \varepsilon_0 \mid X)\right] + \varepsilon_0\right\}.$$

There is still an unobservable element in the coefficient of variable D in the above equation. To identify $ATE_{(X)}$ and $ATT_{(X)}$, we need assumption (A.2.1) that the conditional expectation of the error terms is the same with and without the program. It should be noted again that we do not require this assumption for all X in the potential outcome equations, but only for variables that we want to estimate the parameters conditional on.

Proposition 4.1: Given assumptions (A.4.1) and (A.2.1), OLS regression produces unbiased estimators of $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT.

Proof: Under assumption (A.2.3) equation (4.10) becomes:

$$Y = \alpha_0 + X\beta_0 + D\left[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)\right] + \left[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0\right] \tag{4.10}$$

The proof is now similar to the proof of Proposition 3.2. The error term has the following property:

$$E\left[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0 \mid X, D\right] = DE(\varepsilon_1 - \varepsilon_0 \mid X, D) + E(\varepsilon_0 \mid X, D) = E(\varepsilon_0 \mid X) = 0. \tag{4.11}$$

Thus the estimator of the conditional parameters is:

$$\hat{ATE}_{(X)} = \hat{ATT}_{(X)} = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)X. \tag{4.12}$$

ATE and ATT are identified simply by taking the expectation of $ATE_{(X)}$ and $ATT_{(X)}$ over the distribution of X for the whole population, and the distribution of X for the participant population, respectively. ■

To this end, there is no assumption on homogenous impact required to identify ATE and ATT. Program impact is allowed to differ for different subjects. If the assumption on homogenous impact is imposed, then:

$$\varepsilon_0 = \varepsilon_1. \tag{4.13}$$

The equation of observed outcome with the coefficient of D as $ATE_{(X)}$ and $ATT_{(X)}$ is equal to:

$$Y = \alpha_0 + X\beta_0 + D\left[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)\right] + \varepsilon_0. \tag{4.14}$$

If we further assume that the pattern of two potential outcomes is the same so that $\beta_1 = \beta_0$, equation (4.14) is reduced to:

$$Y = \alpha_0 + X\beta_0 + D(\alpha_1 - \alpha_0) + \varepsilon_0 . \quad (4.15)$$

Equation (4.15) is popular in traditional regression. However, the common effect assumption is very strong, and often does not hold in reality.

Finally, instead of running one regression for the whole sample on the participants and non-participants, it is possible to run two separate regressions for the sub-samples of the participants and non-participants, respectively. Under assumption (A.4.1), we can write:

$$E(Y_0 \mid X, D = 0) = E(Y_0 \mid X) = \alpha_0 + \beta_0 X + E(\varepsilon_0 \mid X), \quad (4.16)$$

$$E(Y_1 \mid X, D = 1) = E(Y_1 \mid X) = \alpha_1 + \beta_1 X + E(\varepsilon_1 \mid X). \quad (4.17)$$

Together with assumption (A.2.1), $\hat{\alpha}_1, \hat{\alpha}_0, \hat{\beta}_1, \hat{\beta}_0$ are obtained by running two regressions on (4.16) and (4.17): one for the participants and the another for the non-participants.

4.2.2. Nonlinear regression

In some cases, the assumption on linearity of the potential outcome function is not reasonable. One important case is that the outcome variable is binary, e.g., one can be interested in the impact of a vocational training program on the probability of getting a job. The outcome variable equals 1 if a person is employed, and 0 otherwise. As we know, the widely-used models are logit or probit instead of the linear model.

In general, we write the potential outcome equations as follows:

$$Y_0 = g(X, \beta_0) + \varepsilon_0, \quad (4.18)$$

$$Y_1 = g(X, \beta_1) + \varepsilon_1, \quad (4.19)$$

where g(X) is an any function of X which can be linear or non-linear in X and parameters β_0 and β_1. Under assumption (A.4.1) we can estimate the two equations separately using the sub-samples of non-participants and participants, respectively. Together with assumption (A.2.3), $ATE_{(X)}$ and $ATT_{(X)}$ are identified:

$$ATE_{(X)} - ATT_{(X)} = g(X, \beta_1) - g(X, \beta_0). \quad (4.20)$$

ATE and ATT are then identified simply by taking the expectation of $ATE_{(X)}$ and $ATT_{(X)}$ over the distribution of X for the whole population, and the distribution of X for the participant population, respectively.

As a matter of estimation, equation (4.20) can suggest the following general estimators for the treatment parameters:

$$\hat{ATE}_{(X)} - \hat{ATT}_{(X)} = g(X, \hat{\beta}_1) - g(X, \hat{\beta}_0), \quad (4.21)$$

$$\hat{ATE} = \frac{1}{n}\sum_{X}\left[g(X,\hat{\beta}_1) - g(X,\hat{\beta}_0)\right], \tag{4.22}$$

$$\hat{ATT} = \frac{1}{n_1}\sum_{X,D} D\left[g(X,\hat{\beta}_1) - g(X,\hat{\beta}_0)\right], \tag{4.23}$$

where n is the number of the total observations (include participants and non-participants), and n_1 is the number of the participants in sample data.

4.2.3. Advantages and disadvantages of regression method

The above described regression method has the advantage of simple implementation, but has with three main drawbacks. Firstly, it imposes a specific functional form on the relation between outcome and conditioning variables and the program participation variable, e.g., a linear function in the above illustration. Secondly, because of the functional form, the OLS regression can have problems of multicollinearity and heteroscedasticity, making the estimator of the program impact inefficient. This might be the case, since the conditioning variables can have effect on D, and the error term can have nonconstant variance. Finally, the method relies on the assumption of program selection based on the observable variables. This assumption is strong unless a rich data set on program selection of participants and non-participants is available.

4.3. Matching methods

4.3.1. Identification assumptions

There is a large amount of literature on matching methods of impact evaluation. Important contributions in this areas can be found in researches such as (Rubin, 1977, 1979, 1980), (Rosenbaum and Rubin, 1983, 1985a), and (Heckman, et al., 1997b). The matching method can be used to estimate the two program impact parameters, ATE and ATT under conditional independence assumption (A.4.1). The basic idea of the matching method is to find a control group (also called comparison group) that has the same (or at least similar) distribution of X as the treatment group. By doing so, we have controlled for the difference in X between the participants and non-participants. The potential outcomes of the control and treatment group are now independent of the program selection. The difference in outcome of the control group and the treatment group then can be attributed to the program impact. However for the matching method to be implemented, we must find a control group who is similar to the treatment group but does not participate in the program. This assumption is called common support in the literature on the matching method. If we denote p(X) as the probability of participating in the program for each subject, i.e. $p(X) = P(D = 1 \mid X)$, this assumption can be stated formally as follows: [19]

Assumption 4.2: $0 < p(X) < 1$ (A.4.2)

[19] As mentioned, this assumption is actually required for variables Z (not for all explanatory variables X in the potential outcome equations) on which the conditional independence assumption is satisfied.

As compared with the regression method, the method of matching has an advantage that it does not require assumption (A.2.3) to identify $ATE_{(X)}$ and $ATT_{(X)}$, but it also suffers from a disadvantage of the common support requirement.

Proposition 4.2: Under assumptions (A.4.1) and (A.4.2), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified by the matching method.

Proof: the proof is straightforward using the conditional independence assumption.

$$ATE_{(X)} = ATT_{(X)} = E(Y_1 \mid X) - E(Y_0 \mid X) = E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=0) \ . \qquad (4.24)$$

Both terms in (4.24) can be observed. In addition, assumption (A.4.2) ensures that there are some participants and non-participants whose values of X are the same so that we are able to use sample information to estimate (4.24).

ATE and ATT are identified as in (2.13) and (2.14).■

4.3.2. Alternative matching methods

Construction of a comparison group

To implement the matching method, we need to find a comparison group who variables are identical to those of the treatment group. The comparison group is constructed by matching each participant i in the treatment group with one or more non-participants j's whose variables X_j are closest to X_i of the participant i. The weighted average outcome of non-participants who are matched with an individual participant i will form the counterfactual outcome for the participant i.

For a participant i, denote n_{ic} as the number of non-participants j who are matched with this participant, and w(i,j) the weight attached to the outcome of each non-participant. These weights are defined non-negative and sum up to 1, i.e.:

$$\sum_{j=1}^{n_{ic}} w(i,j) = 1 \ . \qquad (4.25)$$

The estimator of the conditional program parameters is then equal to:

$$\hat{ATE}_{(X)} = \hat{ATT}_{(X)} = \frac{1}{\sum_{X_i = X} D_i} \left\{ \sum_{X_i = X} \left[Y_{1i} - \sum_{j=1}^{n_{ic}} w(i,j) Y_{0j} \right] \right\} \qquad (4.26)$$

where Y_{1i} and Y_{0j} are the observed outcomes of participant i and non-participant j. In practice, when there are many variables X, it is difficult to have a large number of observations who have the same variables X in a sample. Estimates of program impact conditional on a large number of X will be associated with a huge standard error. Thus formula (4.26) should be used to estimate the program impact for several subgroups defined by one or only a few binary or discrete variable X.

ATT is simply the average of differences in outcome between the treatment and comparison group:

$$\hat{ATT} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left[Y_{1i} - \sum_{j=1}^{n_{ic}} w(i,j) Y_{0j} \right] \quad (4.27)$$

where n_1 is the number of the participants in the data sample.

To estimate the ATE, we use formula (2.8) in which there remains component $E(Y_1 \mid D = 0)$ that requires an estimator. The similar matching procedure is applied to estimate this term. Each non-participant will be matched with one or more participants who have the closest value of X . Put differently we are going to estimate the effect of non-treatment on the non-treated:

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

using the estimator which is similar to (4.27):

$$\hat{ANTT} = \frac{1}{n_2}\sum_{i=1}^{n_2}\left[Y_{0j} - \sum_{i=1}^{n_{jt}} w(j,i) Y_{1i} \right] \quad (4.28)$$

where n_2 is the number of the non-participants in the sample. n_{jt} is the number of participants is matched with a non-participant j, and w(j,i) are weights attached to each participant i in this matching.

Thus using (2.8) the estimator of ATE is expressed as follows:

$$\hat{ATE} = \frac{1}{n_1 + n_2}\left\{ \sum_{i=1}^{n_1}\left[Y_{1i} - \sum_{j=1}^{n_{ic}} w(i,j) Y_{0j} \right] + \sum_{i=1}^{n_2}\left[Y_{0j} - \sum_{i=1}^{n_{jt}} w(j,i) Y_{1i} \right] \right\} \quad (4.29)$$

To this end, there are still two essential issues that have not been discussed. The first is how to select non-participants and participants for matching. The second is how to determine weights w(i,j) among these matched people .

Methods to find a matched sample

Clearly, matched non-participants should have X that is closest to X of participants. There will be no problem if there is a single conditioning variable X. However X is often a vector of variables, and finding "close" non-participants to match with a participant is not straightforward. In the literature on impact evaluation, there are three widely-used methods to find matched non-participants for a participant (and vice versa matched participants for a non-participant).

The first method is called subclassification of the treatment and control group based on X (see, e.g., Cochran and Chambers, 1965; Cochran, 1968). All participants and non-participants will be classified into blocks according to the value of X. It means that subjects in a block have the same value of X. Then non-participants will be matched with participants in each block. However the subclassification becomes difficult when there are many variables X or when some variables of X are continuous or discrete with many values.

The second method is called covariate matching which matches participants with non-participants based on their distance of variables defined on some metric (Rubin, 1979, 1980). Since X can be considered as a vector in a space, the closeness between two sets of X can be defined by a distance metric. A non-participant j will be matched with a participant i if the distance from X_j to X_i is smallest as compared with other non-participants. A quickly emerging metric in space is the traditional Euclidean metric:

$$d_E(i,j) = \left\| X_i - X_j \right\|_E = \left(X_i - X_j\right)' \left(X_i - X_j\right) \tag{4.30}$$

However this metric is sensitive to the measure unit of X . To get an unit-free metric distance, a natural way is to standardize the Euclidean metric by multiplying it with the inversed covariance matrix of X to get the Mahalanobis metric[20] (Rubin, 1979, 1980) or the inversed variance matrix of X (Abadie and Imbens, 2002):

$$d_M(i,j) = \left\| X_i - X_j \right\|_M = \left(X_i - X_j\right)' S_X^{-1} \left(X_i - X_j\right) \tag{4.31}$$

$$d_V(i,j) = \left\| X_i - X_j \right\|_V = \left(X_i - X_j\right)' V_X^{-1} \left(X_i - X_j\right) \tag{4.32}$$

where S_X and V_X are the covariance and variance of X in the sample.

The third way to find the matched sample is the propensity score matching. Since a paper by Rosenbaum and Rubin (1983), matching is often conducted based on the probability of being assigned to the program, which is called the propensity score. Rosenbaum and Rubin (1983) show that if the potential outcomes are independent of the program assignment given X , then they are also independent of the program assignment given the balance score. The balance score is any function of X but finer than p(X), which is the probability of participating in the program (the so-called propensity score).

Proposition 4.3 (Rosenbaum and Rubin, 1983)**:** $Y_0, Y_1 \perp D \mid X \Rightarrow (Y_0, Y_1) \perp D \mid b(X)$, where $b(X)$ is any function such that $p(X) = f[b(X)]$ and $p(X) = Pr(D = 1 \mid X) = E(D \mid X)$.

Proof: It is sufficient to show that:

$$P[D = 1 \mid Y_0, Y_1, b(X)] = P[D = 1 \mid b(X)]. \tag{4.33}$$

Using the law of iterated expectation and noting that $p(X) = f[b(X)]$, we have the following equations:

[20] This metric is presented in Mahalanobis (1936).

$$
\begin{aligned}
P[D=1 \mid Y_0, Y_1, b(X)] &= E[D \mid Y_0, Y_1, b(X)] \\
&= E\{E[D \mid Y_0, Y_1, X, b(X)] \mid Y_0, Y_1, b(X)\} \\
&= E\{E[D \mid Y_0, Y_1, X] \mid Y_0, Y_1, b(X)\} \\
&= E\{E[D \mid X] \mid Y_0, Y_1, b(X)\} \\
&= E\{p(X) \mid Y_0, Y_1, b(X)\} \\
&= E\{p(X) \mid b(X)\} \\
&= P[D=1 \mid b(X)].
\end{aligned}
\tag{4.34}
$$

Using the results of Proposition 4.3, the program impacts can be identified as follows:

$$
\begin{aligned}
ATE_{(X)} = ATT_{(X)} &= E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=0) \\
&= E(Y_1 \mid b(X), D=1) - E(Y_0 \mid b(X), D=0).
\end{aligned}
\tag{4.35}
$$

In fact, the propensity score is often selected as the balance score in estimating the program impacts. The propensity score can estimated parametrically or non-parametrically by running regression of the treatment variable D on the conditioning variables X. Since D is a binary variable, a logit or probit model is often used. Once the propensity score is obtained for all subjects in the sample, non-participants can be matched with participants based on the closeness of the propensity scores.

Researchers can combine the three above methods, subclassification, covariate matching and propensity score matching in finding the matches (Rosenbaum and Rubin, 1984, 1985a). Subclassification can be performed for certain important variables X to ensure that participants and matched non-participants have the same value of these variables. The propensity score can also be used in linear regression to estimate the program impact (see, e.g., Wooldridge, 2001; Rosenbaum and Rubin, 1985a). The program impact parameters can be estimated by running OLS regressions of observed outcome on the propensity score p(X) instead of X (as presented in section (4.2)).

Weighting methods of matched comparisons

Once a metric distance, d(i,j), between a participant i and a non-participant j is defined, one can select methods to weight their outcomes. If one non-participant is matched with one participant who has the minimum value of d(i,j), the weight w(i,j) is equal to 1 for all pairs of matches. This is one nearest neighbor matching. When there are more than one non-participants who are matched with one participant (or vice versa), we need some ways to define weights that are attached to each non-participant.

The n-nearest neighbor matching is to match each participant with n non-participants whose have closest distances d(i,j). Each matched non-participant will receive an equal weight, i.e. $w(i,j) = 1/n$. Another version of equal weights can be caliper matching (see, e.g., Dehejia and Wahba, 1998; Smith and Todd, 2005), which allows for the matching between two subjects if their distance d(i,j) is smaller than a specific value, say 0.05 or 0.1. This is aimed to ensure the quality of matched subjects. Stratification (interval) matching also shares this feature (see, e.g., Dehejia

and Wahba, 1998); Smith and Todd, 2005). It divides the range of estimated distances into several strata (blocks) of equal ranges. Within each stratum, a participant is matched with all non-participants with equal weights.

However, it would be reasonable to assign different weights to different non-participants depending on metric distances between their covariates and the covariates of the matched participant. This argument motivates some others matching schemes such as kernel, local linear matching (see, e.g., Heckman, et al., 1997b; Smith and Todd, 2005), and matching using weights of inversed propensity score (see, e.g., Hahn, 1998; Hirano, et al., 2002).

Kernel matching method matches a participant with one or many non-participants depending a kernel function G and a selected bandwidth h. The weight is defined as:

$$w(i,j) = \frac{G\left[\frac{d(i,j)}{h}\right]}{\sum_{k=1}^{n_2} G\left[\frac{d(i,k)}{h}\right]}. \tag{4.36}$$

Kernel matching can be explained as kernel (non-parametric) estimation of counterfactual $E(Y_0 \mid X, D=1)$ using sample information on non-participants. However, kernel function results in biased estimation if the true regression line is linear. Fan (1992) shows that a so-called method of local linear regression is more flexible and robust to different types of outcome function. This method estimates the regression curve by series of local linear regression lines. Weights estimated from the local linear regression are as follows (Smith and Todd, 2005):

$$w(i,j) = \frac{G_{ij}\sum_{k=1}^{n_2} G_{ik}\left[d(k,i)\right]^2 - \left[G_{ij}d(j,i)\right]\left[\sum_{k=1}^{n_2} G_{ik}d(k,i)\right]}{\sum_{j=1}^{n_2} G_{ij}\sum_{k=1}^{n_2} G_{ik}\left[d(k,i)\right]^2 - \left[\sum_{k=1}^{n_2} G_{ik}d(k,i)\right]^2}. \tag{4.37}$$

Finally, Hahn (1998) and Hirano, et al. (2002) use weights of inversed propensity score to estimate the potential outcomes as follows:

$$E(Y_1 \mid X) = \frac{YD}{p(X)}, \tag{4.38}$$

$$E(Y_0 \mid X) = \frac{Y(1-D)}{1-p(X)}. \tag{4.39}$$

As a result, the conditional program impact is equal to:

$$ATE_{(X)} = ATT_{(X)} = \frac{Y\left[D - p(X)\right]}{p(X)\left[1-p(X)\right]}. \tag{4.40}$$

Thus, instead of using (4.31), these conditional impacts are estimated by:

$$\hat{ATE}_{(X)} = \hat{ATT}_{(X)} = \sum_{X_i = X} \frac{1}{D_i} \left\{ \frac{Y_i \left[D_i - \hat{p}(X_i) \right]}{\hat{p}(X_i)\left[1 - \hat{p}(X_i)\right]} \right\}, \tag{4.41}$$

and the unconditional versions:

$$\hat{ATE} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i \left[D_i - \hat{p}(X_i) \right]}{\hat{p}(X_i)\left[1 - \hat{p}(X_i)\right]}, \tag{4.42}$$

$$\hat{\mathrm{ATT}} = \frac{1}{\mathrm{n}_1} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\mathrm{Y_i} \left[\mathrm{D_i} - \hat{\mathrm{p}}(\mathrm{X_i}) \right]}{\hat{\mathrm{p}}(\mathrm{X_i})\left[1 - \hat{\mathrm{p}}(\mathrm{X_i})\right]}. \tag{4.43}$$

4.3.3. Advantages and disadvantages of matching method

The main advantage of the matching method is that it does not rely on a specific function form of the outcome, thereby avoiding assumptions on functional form, e.g., linearity imposition, multicollinearity and heteroscedasticity issues. Compared with the linear regression, the matching method does not require assumption (A.2.1) on exogeneity of X. In addition, the matching method emphasizes the problem of common support, thereby avoiding the bias due to extrapolation to non-data region. Results from the matching method are easy to explain to policy makers, since the idea of comparison of similar group is quite intuitive.

However, the matching method has several limitations. It relies on the assumption of conditional mean independence. It also requires the assumption of common support. If this assumption does not hold, one can use a method of regression discontinuity, which will be discussed in the next section. Finally, the matching estimators can work very poorly in small samples if the quality of matches is not good, i.e., participants are matched with non-participants who have very different conditioning variables X.[21]

4.4. Regression discontinuity

For the matching method, the assumption on the common support is required to identify the program impacts. When the conditioning variables X are different for participants and non-participants, we cannot implement matching methods. In other words, if there are some variables X that predict the treatment variable D perfectly, the assumption of common support no longer holds. In Van der Klaauw (2002), it means that there is a conditioning variable S belonging to X such that D equals 1 if and only if S is larger than a specific value $\overline{S}$.[22] The assignment of the program is called deterministic. To make this assumption consistent with notation in this paper, we assume that $D = 1$ if and only if $X \geq \tilde{X}$. Then we have:

$$P(D = 1 \mid X \geq \tilde{X}) = 1, \tag{4.44}$$

$$P(D = 1 \mid X < \tilde{X}) = 0. \tag{4.45}$$

[21] For potential in matching estimators, see, e.g., Rosenbaum and Rubin (1985b), and Heckman, et al. (1998a)

[22] Heckman, et al. (1999) presents the case in which $D = 1$ only if $S < \overline{S}$. These two cases are similar.

Which means that the common support assumption $0 < P(D = 1 \mid X) < 1$ is not valid.

We know that the regression method does not require a common support. As a result it can be applied in this context taking into account some important notes. Under the assumption on conditional mean independence, the conditional and unconditional program impact parameters are the same because of:

$$E(Y_0 \mid X, D = 1) = E(Y_0 \mid X, D = 0), \tag{4.46}$$

$$E(Y_1 \mid X, D = 1) = E(Y_1 \mid X, D = 0), \tag{4.47}$$

which can be expressed as follows due to (4.44) and (4.45):

$$E(Y_0 \mid X, X \geq \tilde{X}) = E(Y_0 \mid X, X < \tilde{X}), \tag{4.48}$$

$$E(Y_1 \mid X, X \geq \tilde{X}) = E(Y_1 \mid X, X < \tilde{X}). \tag{4.49}$$

If the potential outcomes are monotonous (as in case of linear function with first-order variables X), (4.48) and (4.49) are obtained only at the point $X = \tilde{X}$ under a condition that the potential outcome are continuous at this point. Since the potential outcomes are functions of the error terms, we can state this assumption with respect to the error terms.

Assumption 4.3: The conditional means of the error terms $E(\varepsilon_0 \mid X)$, and $E(\varepsilon_1 \mid X)$ are continuous at $\tilde{X}$. (A.4.3)

Under assumption (A.4.3) the matching method and other non-parametric estimation methods can be used to estimate the program impacts at the mass of $\tilde{X}$. This is called local treatment effect at $\tilde{X}$ (see, e.g.,Van der Klaauw, 2002; Hahn, et al., 2001).

The parametric approach can identify the program impact at all the range of X . Thus the regression method presented in section 4.2 can be used to estimate the program impact parameters. But we need to make an assumption that the parameters, i.e., $\alpha_0, \beta_0, \alpha_1, \beta_1$, in the potential outcomes are the same in the range $X \geq \tilde{X}$, and $X < \tilde{X}$ as well as the whole range of X. By running regression of the potential outcomes (or observed outcomes), we use data on outcome of participants Y_1 with $X \geq \tilde{X}$ to extrapolate the value of potential outcome Y_1 for non-participants with $X < \tilde{X}$. Similarly, data on outcome of non-participants Y_0 with $X < \tilde{X}$ are used to extrapolate the value of potential outcome Y_0 for participants with $X \geq \tilde{X}$. This method might lead to a so-call extrapolation bias since we predict outcome values in regions of no observations.

When the program participation is not absolutely deterministic, i.e. there are some subjects who have $X \geq \tilde{X}$ but do not participate in the program, or some other subjects who have $X < \tilde{X}$ but do participate in the program, the problem becomes similar to other contexts. The actual program assignment still depends on other factors, and we need to use methods based on selection on unobservable to estimate the program impact.

In short, the method of regression discontinuity is a version of regression method. Thus it has advantages and disadvantages similar to the regression method. It is worth noting that this method can overcome the requirement of common support in the matching method at the expense of a potential bias in extrapolation into no data regions.

5. Methods assuming selection on unobservable

As discussed, the main assumption that the methods of selection on observable rely on is the conditional independence between the potential outcomes and program assignment (or a weaker version of conditional mean independence). This assumption no longer holds if there is an unobserved variable affecting both the potential outcome and the program participation. This section presents three methods that are widely-used in dealing with the problem of "selection on unobservable". The methods include instrumental variables, sample selection model, and panel data model.

5.1. Instrumental variables

5.1.2. Program impact identification

A standard solution to the problem of an endogenous variable in parametric regression is to use an instrumental variable for the program assignment variable D. An instrument variable has two properties: (i) it is correlated with the program assignment (ii) it is uncorrelated with the error term in the potential outcomes. [23]

To illustrate how the instrumental variables method identifies the program impact, recall the equations of the observed outcome (2.16) and (2.17) in which the coefficients of D are expressed as $ATE_{(X)}$ and $ATT_{(X)}$. In these parameters, there remains an unobserved component:

$E(\varepsilon_1 - \varepsilon_0 \mid X, D = 1)$.

To identify $ATT_{(X)}$, we assume that the expectation of error terms conditional on X for the participants is the same in the state of program and the state of no-program. That is:

Assumption 5.1: $E(\varepsilon_1 \mid X, D = 1) = E(\varepsilon_0 \mid X, D = 1)$. (A.5.1)

Under this assumption and assumption (A.2.1), $ATE_{(X)}$ and $ATT_{(X)}$ are the same, and they can be estimated from the following equation:

$$Y = \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] + [D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0]. \quad (5.1)$$

The remaining problem is to solve the endogeneity of D in (5.1). Thus we need an instrumental variable for D to estimate the program impacts.

Assumption 5.2: There is at least an instrumental variable Z such that:

[23] Examples of instrumental variables can be seen in econometrics textbooks such as Wooldridge (2001), Greene (2003) or papers on review of impact evaluation such as Moffitt (1991).

$Cov(D,Z) \neq 0$,

$$E(\varepsilon_0 \mid Z) = E(\varepsilon_0), \tag{A.5.2}$$

$E(\varepsilon_1 \mid Z) = E(\varepsilon_1)$.

Proposition 5.1: Under assumptions (A.5.1), (A.5.2) and (A.2.1), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified and estimated by the instrumental variables method.

Proof:

Firstly we show that:

$$Cov\left(\left[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0\right], Z\right) = 0. \tag{5.2}$$

Note that $E(\varepsilon_1 - \varepsilon_0 \mid D,Z) = E(\varepsilon_1 - \varepsilon_0 \mid D) = 0$ because of (A.5.1) and (A.5.2), hence:

$$\begin{aligned} Cov\left(\left[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0\right], Z\right) &= Cov\left(\left[D(\varepsilon_1 - \varepsilon_0)\right], Z\right) + Cov\left(\varepsilon_0, Z\right) \\ &= E\left\{D(\varepsilon_1 - \varepsilon_0) - E\left[D(\varepsilon_1 - \varepsilon_0)\right]\right\}\left\{Z - E(Z)\right\} \\ &= E\left[DZ(\varepsilon_1 - \varepsilon_0)\right] \\ &= 0. \end{aligned}$$

Similar, we have:

$$Cov\left(\left[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0\right], X\right) = 0, \tag{5.3}$$

$$Cov\left(\left[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0\right], XZ\right) = 0. \tag{5.4}$$

Then we have the following covariance equations due to (5.2), (5.3) and (5.4):

$$\begin{aligned} Cov(Y,Z) &= Cov\left\{\alpha_0 + X\beta_0 + D\left[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)\right] + \left[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0\right], Z\right\} \\ &= Cov(X,Z)\beta_0 + Cov(D,Z)(\alpha_1 - \alpha_0) + Cov(XD,Z)(\beta_1 - \beta_0), \end{aligned} \tag{5.5}$$

$$\begin{aligned} Cov(Y,X) &= Cov\left\{\alpha_0 + X\beta_0 + D\left[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)\right] + \left[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0\right], X\right\} \\ &= Var(X)\beta_0 + Cov(D,X)(\alpha_1 - \alpha_0) + Cov(XD,X)(\beta_1 - \beta_0), \end{aligned} \tag{5.6}$$

$$\begin{aligned} Cov(Y,XZ) &= Cov\left\{\alpha_0 + X\beta_0 + D\left[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)\right] + \left[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0\right], XZ\right\} \\ &= Cov(X,XZ)\beta_0 + Cov(D,XZ)(\alpha_1 - \alpha_0) + Cov(XD,XZ)(\beta_1 - \beta_0) \end{aligned} \tag{5.7}$$

It is obvious that the number of unknown parameters is equal to the number of equations. Thus the parameters in (5.1) are estimated without bias, and so are the conditional and unconditional ATE and ATT.■

It should be noted that equation (5.1) includes the interaction between X and D, thus it is considered to include endogenous variables, D and XD, and we use instrumental variables Z and XZ to solve the endogeneity problem.

The model (5.1) allows for the program impact to be different across subjects, but it needs to impose assumption (A.5.1) on the conditional expectation of the error terms. If we are willing to invoke an assumption on homogenous impact given X, i.e. $\varepsilon_0 = \varepsilon_1$, which is stronger than assumption (A.5.1), then (5.1) becomes simpler:

$$Y = \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] + \varepsilon_0 . \quad (5.8)$$

There is no component ε_1 in (5.8), thus the condition $Cov(\varepsilon_1, Z) = 0$ in (A.5.2) can be dropped.

Furthermore, if we make an assumption on homogenous impact regardless of X, i.e., every subject gain the same impact when joining the program, the program impact can be identified by the simplest model:

$$Y = \alpha + \beta D + \varepsilon . \quad (5.9)$$

The condition is that at least an instrumental variable such that $Cov(D, Z) \neq 0$ and $Cov(\varepsilon, Z) \neq 0$ is found.

The instrumental variable method is presented above for just-identification, i.e., only one instrumental variable. The case of over-identification in which there are more than one instrumental variable for the treatment variable D can be solved easily by applying two-stage least square regression (see, e.g., Wooldridge (2001)).[24]

5.1.3. Local average treatment effect

The instrumental variable method presented in the above section is the standard one. It requires assumption (A.5.1) to identify the program impact. Imbens and Angrist (1994) proposes an another method of instrumental variables that does not rely on assumption (A.5.1) in identifying a so-called local average treatment effect (LATE). The LATE parameter measures the effect of the program on those who change program status due to a change in an instrumental variable Z. As Z is define as a policy or a set of policies, one would be interested in impact of a program on those who are included in the program as a result of policy changes.

To make the definition formally, suppose there is an instrument variable Z, whose value changed from $Z = z_0$ to $Z = z_1$. As a result, there are a number of subjects who changes their status from non-participation to participation in the program. Further denote D(z,X) is the treatment variable D but conditional on $Z = z$ for subjects with X . Then LATE is defines as[25]:

$$LATE_{(X, z_0, z_1)} = E[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1]. \quad (5.10)$$

[24] For example, in the first stage the propensity score is estimated using instrumental variables. Then in the second stage, the predicted propensity score is used as an instrumental variable in the outcome equation.

[25] (Heckman and Vytlacil, 1999) defines the limit form of LATE which called local instrument variable parameter:

$$LIV_{X, P(D=1 \mid X, Z=z)} = \frac{\partial[Y \mid X, P(D=1 \mid X, Z=z)]}{\partial P(D=1 \mid X, Z=z)}$$

In addition to the condition of instrumental variables (A.5.2), Imbens and Angrist (1994) impose an additional assumption to identify LATE.

Assumption 5.4: For all z and z' of Z, either $D(z, X) \geq D(z', X)$ or $D(z, X) \leq D(z', X)$ for all subjects. (A.5.4)

In other words, if D can be expressed in a latent variable context, in which D = 1 if D^* is greater than zero, and otherwise, then D^* is required to be monotonous in Z. Once conditional on X, any subject should prefer to participate (or quit) the program as the instrument Z changes its value from z to z'.

Proposition 5.1 (Imbens and Angrist, 1994)**:** Under assumption (A.5.2) and (A.5.4), LATE is identified as follows:

$$\begin{aligned} LATE_{(X, z_0, z_1)} &= E[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1] \\ &= \frac{E(Y \mid X, Z = z_1) - E(Y \mid X, Z = z_0)}{P(D = 1 \mid X, Z = z_1) - P(D = 1 \mid X, Z = z_0)}, \end{aligned} \tag{5.11}$$

where Y is the observed outcome, and the denominator is different from zero.

Proof: We have:

$$\begin{aligned} E(Y \mid X, Z = z_0) &= E\{(Y_1 D(z_0, X) + [(1 - D(z_0, X))Y_0] \mid X, Z = z_0\} \\ &= D(z_0, X)E(Y_1 \mid X) + [1 - D(z_0, X)]E(Y_0 \mid X), \end{aligned} \tag{5.12}$$

$$\begin{aligned} E(Y \mid X, Z = z_1) &= E\{(Y_1 D(z_1, X) + [(1 - D(z_1, X))Y_0] \mid X, Z = z_1\} \\ &= D(z_1, X)E(Y_1 \mid X) + [1 - D(z_1, X)]E(Y_0 \mid X). \end{aligned} \tag{5.13}$$

Subtract (5.12) from (5.13), we get:

$$\begin{aligned} &E(Y \mid X, Z = z_1) - E(Y \mid X, Z = z_0) \\ &= [D(z_1, X) - D(z_0, X)]E(Y_1 - Y_0 \mid X) \\ &= E[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1]P[D(z_1, X) - D(z_0, X) = 1] \\ &\quad + E[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = -1]P[D(z_1, X) - D(z_0, X) = -1] \\ &= E[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1]P[D(z_1, X) - D(z_0, X) = 1]. \end{aligned} \tag{5.14}$$

The last line results from assumption (A.5.4) that there is no person who quits the program due to the change in Z from z_0 to z_1.

Hence:

$$\begin{aligned} E[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1] &= \frac{E(Y \mid X, Z = z_1) - E(Y \mid X, Z = z_0)}{P[D(z_1, X) - D(z_0, X) = 1]} \\ &= \frac{E(Y \mid X, Z = z_1) - E(Y \mid X, Z = z_0)}{P(D = 1 \mid X, Z = z_1) - P(D = 1 \mid X, Z = z_0)}. \end{aligned} \tag{5.15}$$

The unconditional LATE is identified by taking the expectation of (5.11) over X . The parameters can be estimated non-parametrically since all variables in (5.11) are observed in sample data.■

Finally, it should be noted that Z can be a vector of instrumental variables, and LATE is defined as the program impact on those whose participate in the program due to a change in a set of program policies.

5.1.4. Advantages and disadvantages of instrumental variable method

The main advantage of the instrumental variable method is that it allows for the program selection based on unobservable. In addition, LATE can be identified by this method under very general conditions. However, the main problem in this method is to find good instrumental variables. A variable that is correlated with the program selection is often correlated with outcomes and error terms in the potential outcome equations. Using an invalid instrumental variable that does not satisfy the instrument conditions will lead to biased and inconsistent estimates of the program impacts. In contrast, a variable that is uncorrelated with the error terms can be very weakly correlated with the program selection. Estimation with weak instruments can have large standard errors in small samples. In addition, explanation of this method to policy makers is not straightforward.

5.2. Sample selection model

5.2.1. Program impact identification

Impacts of a program can be identified using a sample selection model (Heckman, 1978). Recall that we cannot run regression of the potential outcomes using sample data in the presence of the selection bias because of the non-random missing data. For example, in the equation of Y_0 there is no data on the dependent variable for those who participated in the program. This is similar to the case of the censored dependent variable model, in which the dependent variables is censored according a selection mechanism. Under assumptions on distribution between the error term in the program selection and the error terms in the potential outcome equations, we can estimate coefficients in the potential outcomes without any bias.

Let's write the impact evaluation model again:

The potential outcomes:

$$Y_0 = \alpha_0 + X\beta_0 + \varepsilon_0 ,$$

$$Y_1 = \alpha_1 + X\beta_1 + \varepsilon_1 ,$$

and the outcome that we observe is:

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

where D is determined by the following framework:

$$D^* = \theta W + v ,$$

$D = 1$ if $D^* > 0$,

$D = 0$ otherwise.

As in (2.19), the equation of the observed outcome is:

$$Y = \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] + [D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0]$$

$ATE_{(X)}$ and $ATT_{(X)}$ can be estimated without bias if we are able to get unbiased estimators of $(\alpha_1 - \alpha_0)$, and $(\beta_1 - \beta_0)$, and the term, $E(\varepsilon_1 - \varepsilon_0 \mid X, D = 1)$.

If we estimate coefficients in (2.19) directly, the term, $[(\varepsilon_1 - \varepsilon_0)D + \varepsilon_0]$ that is correlated with X and D will enter the error term. As a result, the coefficient estimators will be biased due to the endogeneitty of X and D. To avoid this problem, we need to model the term $[(\varepsilon_1 - \varepsilon_0)D + \varepsilon_0]$ under an assumption on the relation between error terms $v, \varepsilon_0, \varepsilon_1$.

Assumption 5.5: The error term v in the program participation equation and each of the error terms ε_0, ε_1 in the potential outcome equations follows the following bivariate normal distributions:

$$(v, \varepsilon_0) \sim N_2(0,0,1,\sigma_{\varepsilon_0}, \rho_0)$$

$$(v, \varepsilon_1) \sim N_2(0,0,1,\sigma_{\varepsilon_1}, \rho_1) \qquad \text{(A.5.5)}$$

To get the unbiased estimators of the conditional parameters, we need an assumption on the exogeneity of X in the potential outcome equations, i.e. assumption (A.2.1).

Proposition 5.2: Under assumptions (A.5.5) and (A.2.1), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified.

Proof:

We have the conditional expectation of the observed outcome in equation (2.15):

$$E(Y \mid X, D) = \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] + E\{D[(\varepsilon_1 - \varepsilon_0)] + \varepsilon_0 \mid X, D\}, \qquad (5.16)$$

in which:

$$\begin{aligned}
E\{D[(\varepsilon_1 - \varepsilon_0)] + \varepsilon_0 \mid X, D\} &= DE(\varepsilon_1 - \varepsilon_0 \mid X, D) + E(\varepsilon_0 \mid X, D) \\
&= E(\varepsilon_1 \mid X, D = 1)P(D = 1 \mid X) + E(\varepsilon_0 \mid X, D = 0)P(D = 0 \mid X) \\
&= E(\varepsilon_1 \mid X, v > -\theta W)P(D = 1 \mid X) + E(\varepsilon_0 \mid X, v \le -\theta W)P(D = 0 \mid X) \\
&= \left\{E(\varepsilon_1 \mid X) + \rho_1\sigma_{\varepsilon_1}\frac{\phi(\theta W)}{\Phi(\theta W)}\right\}P(D = 1 \mid X) + \left\{E(\varepsilon_0 \mid X) + \rho_0\sigma_{\varepsilon_0}\frac{-\phi(\theta W)}{1 - \Phi(\theta W)}\right\}P(D = 0 \mid X) \\
&= \left[\rho_1\sigma_{\varepsilon_1}\frac{\phi(\theta W)}{\Phi(\theta W)}\right]P(D = 1 \mid X) - \left[\rho_0\sigma_{\varepsilon_0}\frac{\varphi(\theta W)}{1 - \Phi(\theta W)}\right]P(D = 0 \mid X),
\end{aligned} \qquad (5.17)$$

where the fourth lines results from the definition of the truncated distribution (see, e.g., Greene (2003)). $\phi(.)$ and $\Phi(.)$ are the probability density function and the cumulative probability function of the standard normal distribution, respectively.

Hence (5.16) has the form:

$$Y = \alpha_0 + X\beta_0 + D\left[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)\right] + \left[\rho_1\sigma_{\varepsilon_1}\frac{\phi(\theta W)}{\Phi(\theta W)}\right]P(D=1 \mid X) - \left[\rho_0\sigma_{\varepsilon_0}\frac{\varphi(\theta W)}{1-\Phi(\theta W)}\right]\left[1 - P(D=1 \mid X)\right] + u, \quad (5.18)$$

where u is an error term. (5.18) can be estimated by OLS or maximum likelihood methods. Estimates of θ are obtained from estimation of the program selection equation, while $P(D=1 \mid X)$ is the propensity score that can be estimated parametrically or non-parametrically.

To identify $ATT_{(X)}$, we need the estimation the term $E(\varepsilon_1 - \varepsilon_0 \mid X, D=1)$, which is equal to:

$$E(\varepsilon_1 - \varepsilon_0 \mid X, D=1) = E(\varepsilon_1 \mid X, v > -\theta W) - E(\varepsilon_0 \mid X, v > -\theta W) = \left(\rho_1\sigma_{\varepsilon_1} - \rho_0\sigma_{\varepsilon_0}\right)\frac{\phi(\theta W)}{\Phi(\theta W)}, \quad (5.19)$$

in which $\rho_1\sigma_{\varepsilon_1}$ and $\rho_0\sigma_{\varepsilon_0}$ are estimated from (5.18).■

Although there is no strict requirement of exclusion restriction, i.e. at least an instrumental variable included in W, such an instrumental variable should be included in W to avoid high multicollearity in (5.18). In addition, if we are able to find instrumental variables in W, the expectation of the error terms conditional on X and D can be estimated semi-parametrically or non-parametrically without assumption on the bivariate normal distribution of the error terms (see, e.g., Heckman, 1990; Powell, 1994).

5.2.2. Advantages and disadvantages

Similar to the method of instrumental variables, the main advantage of the sample selection method is that it allows for selection of a program based on unobservable. In addition, it is robust to heterogeneous impacts of the program. However, the main problem in this method is that it requires the assumption on the functional form of the join distribution of the error terms in the selection equation and the potential outcome equations. In addition, a good instrumental variable is often needed to get efficient estimators of the program impact. However, finding a good instrument is rather difficult. It is also difficult to explain the method to policy makers as well as the program administrators.

5.3. Panel data methods

When longitudinal data or panel data on the participants and non-participants in a program before and after the program implementation are available, we can get unbiased estimators of program impacts which allow for "selection on unobservable". Methods discussed here are based

on the panel data at two points of time, since this type of data are the most popular. Panel data with many repeated observations are rather rare in reality.

5.3.1. First-difference method

To illustrate how the method identifies the program impact, let's write the model of the outcome before the program implementation as follows:

$$Y_{0B} = \alpha_{0B} + X_B\beta_{0B} + \varepsilon_{0B} \tag{5.20}$$

where Y, X, and ε are outcome, conditioning variables, and error term, respectively. But they have the subscripts "0" and "B" that means "no program" and "before the program", respectively. Before the program, all people are in status of no program, and the observed outcome is the outcome in the absence of the program.

After the program, the denotation of the potential outcomes is similar to the case of single cross-section data, but has an additional subscript "A" that means "after the program":

$$Y_{0A} = \alpha_{0A} + X_A\beta_{0A} + \varepsilon_{0A} \tag{5.21}$$

$$Y_{1A} = \alpha_{1A} + X_A\beta_{1A} + \varepsilon_{1A} \tag{5.22}$$

Then, the conditional parameters of interest are expressed as follows:

$$ATE_{(X)} = (\alpha_{1A} - \alpha_{0A}) + X_A(\beta_{1A} - \beta_{0A}) + E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_A) \tag{5.23}$$

$$ATT_{(X)} = (\alpha_{1A} - \alpha_{0A}) + X_A(\beta_{1A} - \beta_{0A}) + E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_A, D = 1) \tag{5.24}$$

The key assumption in the first-difference method is that the error term includes a time-invariant component and any correlation between D and the error is included in this component. The time-invariant component can be called the fixed and unobserved effect.

Assumption 5.6: Error terms in the potential outcome equations are decomposed to components with the following properties:

$\varepsilon_{0B} = \pi + \eta_{0B}$,

$\varepsilon_{0A} = \pi + \eta_{0A}$,

$\varepsilon_{1A} = \pi + \eta_{1A}$,

where:

$$\eta_{0B}, \eta_{0A}, \eta_{1A} \perp D \mid X_B, X_A \tag{A.5.6}$$[26]

[26] In some econometrics text, $\eta_{0B}, \eta_{0A}, \eta_{1A} \perp D \mid X_B, X_A$ is called strict exogeneity condition.

For identification of the program impact, we require a weaker assumption, in which the assumption (A.5.6) is stated in terms of expectation of errors.

Assumption 5.6': Error terms in the potential outcome equations are decomposed to components with the following properties:

$$E(\varepsilon_{0B} \mid X_{BA}, D) = E(\pi \mid X_{BA}, D) + E(\eta_{0B} \mid X_{BA}, D) = E(\pi \mid X_{BA}, D) + E(\eta_{0B} \mid X_{BA}) \quad (5.25)$$

$$E(\varepsilon_{0A} \mid X_{BA}, D) = E(\pi \mid X_{BA}, D) + E(\eta_{0A} \mid X_{BA}, D) = E(\pi \mid X_{BA}, D) + E(\eta_{0A} \mid X_{BA}) \quad (5.26)$$

$$E(\varepsilon_{1A} \mid X_{BA}, D) = E(\pi \mid X_{BA}, D) + E(\eta_{1A} \mid X_{BA}, D) = E(\pi \mid X_{BA}, D) + E(\eta_{1A} \mid X_{BA}) \quad (5.27)$$

where π is a component with the expectation unchanged during time for the state of no program. η is a component that is allowed to change over time, but its expectation is independent of D given the variables, $X_{BA} = \{X_B, X_A\}$. (A.5.6')

This assumption hold if the time-variant component of the error terms is independent of the program selection. However, assumption (A.5.6') requires only the conditional mean independence of this component with respect to the program selection.

In addition, to identify $ATE_{(X)}$ and $ATT_{(X)}$, we need assumptions on exogeneity of X, i.e., an assumption similar to (A.2.1):

Assumption 5.7: $E(\varepsilon_{0B} \mid X_B, X_A) = E(\varepsilon_{0A} \mid X_B, X_A) = E(\varepsilon_{1A} \mid X_B, X_A) = 0$ (A.5.7)

Proposition 5.3: Under assumptions (A.5.6) and (A.5.7), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified and can be estimated by OLS regression.

Proof:

Firstly, under assumption (A.5.6) and (A.5.7), $ATE_{(X)}$ and $ATT_{(X)}$ are identified and the same, since:

$$E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_A) = 0,$$

$$\begin{aligned} E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_B, X_A, D = 1) &= E(\eta_{1A} - \eta_{0A} \mid X_B, X_A, D = 1) \\ &= E(\eta_{1A} - \eta_{0A} \mid X_B, X_A) \\ &= E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_B, X_A) \\ &= 0, \end{aligned}$$

As a result, $E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_A, D = 1) = 0$.

The estimator of $ATE_{(X)}$ and $ATT_{(X)}$ is the coefficient of D in the following equation:

$$Y_A = \alpha_{0A} + X_A\beta_{0A} + D[(\alpha_{1A} - \alpha_{0A}) + X_A(\beta_{1A} - \beta_{0A})] + [D(\varepsilon_{1A} - \varepsilon_{0A}) + \varepsilon_{0A}], \quad (5.28)$$

To estimate $(\alpha_{1A} - \alpha_{0A})$ and $(\beta_{1A} - \beta_{0A})$, subtract (5.20) from (5.26) to obtain:

$$Y_A - Y_{0B} = (\alpha_{0A} - \alpha_{0B}) + (X_A\beta_{0A} - X_B\beta_{0B}) + D[(\alpha_{1A} - \alpha_{0A}) + X_A(\beta_{1A} - \beta_{0A})] + [D(\varepsilon_{1A} - \varepsilon_{0A}) + (\varepsilon_{0A} - \varepsilon_{0B})], \quad (5.29)$$

in which the error term has the traditional property due to the (A.5.7) and (A.5.6):

$$\begin{aligned} & E\{[D(\varepsilon_{1A} - \varepsilon_{0A}) + (\varepsilon_{0A} - \varepsilon_{0B})]X_B, X_A, D\} \\ &= DE(\varepsilon_{1A} - \varepsilon_{0A} \mid X_{BA}, D) + E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_{BA}, D) \\ &= DE(\eta_{1A} - \eta_{0A} \mid X_{BA}, D) + E(\eta_{0A} - \eta_{0B} \mid X_{BA}, D) \\ &= DE(\eta_{1A} - \eta_{0A} \mid X_{BA}) + E(\eta_{0A} - \eta_{0B} \mid X_{BA}) \\ &= 0 \end{aligned} \quad (5.30)$$

Thus, we can estimate all coefficients in (5.29) (also in (5.28)) without bias by running regression of the difference in observed outcome before and after the program on X_B and X_A, and the program selection variable D. Then, the estimates of these coefficients will be used to estimate the conditional and unconditional parameters of the program impact.■

5.3.2. Difference-in-difference with matching method

The method of difference-in-difference with matching can be regarded a non-parametric version of the first-difference method. It allows the program selection to be based on unobservable variables in sense that it does not require the conditional independence assumption (A.4.1). It allows for bias in using the conditional expectation of outcome of non-participants to predict the conditional expectation of outcome of participants if they had not participated in the program. However, it requires the bias be time-invariant. Compared with the first-difference method, it has an advantage that it does require the assumption on exogeneity of X to identify the program impact parameters.

Proposition 5.4: Under assumptions (A.5.6), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified and can be estimated non-parametrically by the matching method.

Proof:

From (A.5.6), we get:

$$\begin{aligned} E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_B, X_A, D) &= E(\eta_{0A} - \eta_{0B} \mid X_B, X_A, D) \\ &= E(\eta_{0A} - \eta_{0B} \mid X_{BA}) \\ &= E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_{BA}), \end{aligned} \quad (5.31)$$

where X_{BA} denote all X_B and X_A. Thus, $E(\varepsilon_{0A} - \varepsilon_{0B})$ is independent of D given X_B and X_A before and after the program. As a result:

$$E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_{BA}, D = 0) = E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_{BA}, D = 1) \quad (5.32)$$

$$\Leftrightarrow E(\varepsilon_{0A} \mid X_{BA}, D = 0) - E(\varepsilon_{0B} \mid X_{BA}, D = 0) = E(\varepsilon_{0A} \mid X_{BA}, D = 1) - E(\varepsilon_{0B} \mid X_{BA}, D = 1)$$

$$\Leftrightarrow E(Y_{0A} \mid X_{BA}, D = 0) - E(Y_{0B} \mid X_{BA}, D = 0) = E(Y_{0A} \mid X_{BA}, D = 1) - E(Y_{0B} \mid X_{BA}, D = 1) \quad (5.33)$$

Recall that $ATT_{(X)}$ is equal to:

$$ATT_{(X_B,X_A)} = E(Y_{1A} \mid X_{BA}, D=1) - E(Y_{0A} \mid X_{BA}, D=1). \quad (5.34)$$

Insert (5.33) into (5.34) to obtain:

$$\begin{aligned} ATT_{(X_B,X_A)} &= E(Y_{1A} \mid X_{BA}, D=1) - E(Y_{0A} \mid X_{BA}, D=1) - \left[E(Y_{0A} \mid X_{BA}, D=0) - E(Y_{0B} \mid X_{BA}, D=0)\right] \\ &\quad + \left[E(Y_{0A} \mid X_{BA}, D=1) - E(Y_{0B} \mid X_{BA}, D=1)\right] \\ &= \left[E(Y_{1A} \mid X_{BA}, D=1) - E(Y_{0A} \mid X_{BA}, D=0)\right] - \left[E(Y_{0B} \mid X_{BA}, D=1) - E(Y_{0B} \mid X_{BA}, D=0)\right] \end{aligned} \quad (5.35)$$

Similarly, we can identify the conditional average effect of non-treatment on the non-treated (ANTT):

$$\begin{aligned} ANTT_{(X_B,X_A)} &= E(Y_{1A} \mid X_{BA}, D=0) - E(Y_{0A} \mid X_{BA}, D=0) - \left[E(Y_{1A} \mid X_{BA}, D=0) - E(Y_{0B} \mid X_{BA}, D=0)\right] \\ &\quad + \left[E(Y_{1A} \mid X_{BA}, D=1) - E(Y_{0B} \mid X_{BA}, D=1)\right] \\ &= \left[E(Y_{1A} \mid X_{BA}, D=1) - E(Y_{0A} \mid X_{BA}, D=0)\right] - \left[E(Y_{0B} \mid X_{BA}, D=1) - E(Y_{0B} \mid X_{BA}, D=0)\right], \end{aligned} \quad (5.36)$$

which is the same as $ATT_{(X)}$. As a result, $ATE_{(X)}$ is identified, and it is equal to $ATT_{(X)}$.

The unconditional parameters are also identified due to (2.13) and (2.14).■

The method of matching in this context is similar to what is described in section (4.2). However, as (5.36) indicates, a participant is matched with a non-participant based on their conditioning variables before and after the program, X_B and X_A.

The above matching method requires panel data. If only independently pooled cross section data are available, the matching will be performed in a slightly different way. The identification assumption is revised as follows.

Assumption 5.8: The difference of the conditional expectation of outcomes before and after program is the same for the participant and non-participants, i.e.:

$$\left[E(Y_{0A} \mid X_A, D=1) - E(Y_{0B} \mid X_B, D=1)\right] = \left[E(Y_{0A} \mid X_A, D=0) - E(Y_{0B} \mid X_B, D=0)\right]$$

$$\left[E(Y_{1A} \mid X_A, D=1) - E(Y_{0B} \mid X_B, D=1)\right] = \left[E(Y_{1A} \mid X_A, D=0) - E(Y_{0B} \mid X_B, D=0)\right] \quad (A.5.8)$$

(A.5.8) is different from (A.5.6). For example, in condition (5.33) which results from assumption (A.5.6), all expectation terms include both X_B and X_A, while in the first equation of (A.5.8) the expectation terms include either X_B or X_A.

There is no argument for whether assumption (A.5.8) is stronger than (5.33) or vice versa.

Then, under this assumption (A.5.8), the $ATT_{(X)}$ is equal to:

$$
\begin{aligned}
ATT_{(X_A)} = {} & E(Y_{1A} \mid X_A, D=1) - E(Y_{0A} \mid X_A, D=1) \\
& - \left[E(Y_{0A} \mid X_A, D=0) - E(Y_{0B} \mid X_B, D=0)\right] + \left[E(Y_{0A} \mid X_A, D=1) - E(Y_{0B} \mid X_B, D=1)\right] \\
= {} & \left[E(Y_{1A} \mid X_A, D=1) - E(Y_{0A} \mid X_A, D=0)\right] - \left[E(Y_{0B} \mid X_B, D=1) - E(Y_{0B} \mid X_B, D=0)\right].
\end{aligned}
\tag{5.37}
$$

In implementation, firstly participants are matched to non-participants based on X_B to estimate the difference in their outcome before the program. Secondly, after the program participants are matched to non-participants again but based on X_A to estimate the difference in their outcome. Then, the estimate of the program impact $ATT_{(X)}$ is equal to the difference in the estimates before and after the program. That is why this method is also called double-matching.

Note that the term $\left[E(Y_{0A} \mid X, D=1) - E(Y_{0A} \mid X, D=0)\right]$ in (A.5.8) is set equal to zero in conditional independence assumption (A.4.1). This is bias when the conditional expectation of outcome of non-participants is used to predict the conditional expectation of outcome of participants if they had not participated in the program. Matching method using single cross-section data assumes this bias equal zero once conditional on X. Thus, the matching method is more robust than the matching method in sense that it allows this bias to differ from zero. However it requires that this bias be time-invariant.

Similarly, under the second condition of (A.5.8), $ANTT_{(X)}$ is identified. It is the same as $ATT_{(X)}$. As a result $ATE_{(X)}$ is also the same as $ATT_{(X)}$.

5.3.3. Advantages and disadvantages of the panel data methods

The main advantage of the panel data methods is that it allows for the selection of the program based on some unobservable variables. However, the methods have two disadvantages. The first is the requirement of the data set. Panel data that are collected before and after the program are not popular as single cross-section data. The second is that the methods require two assumptions to identify $ATE_{(X)}$ and $ATT_{(X)}$. The assumptions require that unobservable variables that affect the program selection are unchanged over time and the program statuses. These assumptions might be violated if the time period between two panel data sets is long enough so that the unobservable variables of subjects are altered. In addition, the unobservable variables can be changed as the subject participate in the program.

6. Simulation results

6.1. Simulation design

This section presents simulation results of measuring the ATT using the methods which have been discussed. The simulations have three main objectives. Firstly, they compare different estimators of program impacts in terms of mean-squared-error (MSE) for some simulation designs. Secondly, it examines biases in the "selection on observable" methods, i.e., regression and matching when the assumption on the conditional independence no longer holds. Thirdly, the simulations will investigate the role of instrumental variables in the estimation of the program impacts using the "selection on unobservable" methods, i.e., instrumental variables, sample selection.

Suppose that a program, denoted by a binary variable D, is assigned to some people in a population. Before program implementation, people have an observed outcome which is a function of covariates X and error terms ε:

$$Y_B = 5 + X_1 + X_{2B} + \varepsilon_B \,. \tag{6.1}$$

After the program, corresponding to the states of program and no-program, there are 2 potential outcomes, which are also expressed as functions of observed and unobserved variables:

$$Y_{0A} = 10 + X_1 + X_{2A} + \varepsilon_{0A}\,, \tag{6.2}$$

$$Y_{1A} = 15 + X_1 + 1.5X_{2A} + \varepsilon_{10}\,, \tag{6.3}$$

In (6.1), (6.2) and (6.3), X_1, X_{2B} and X_{2A} each follow a normal distribution $N(\mu,\sigma) = N(10,5)$, and each error term follows a normal distribution $N(\mu,\sigma) = N(0,5)$.

The assignment of the program D is designed as follows.

$$W = gX_1 + X_{2A} + kZ + u\,, \tag{6.4}$$

$D = 1$ if $W < W^*$,

$D = 0$ otherwise ,

where variable Z follows a normal distribution $N(\mu,\sigma) = N(10,5)$, and error term u follow a normal distribution $N(\mu,\sigma) = N(0,5)$. g and k are coefficients that reflect the roles of X_1 and Z in the program selection equation.

There are two points that should be noted. Firstly, the outcomes are assumed depends on two observed variables X_1 and X_2, of which X_1 is time-invariant, i.e., the same variable X_1 is specified in the outcome equations before and after the program. Secondly, the program selection depends on three variables X_1, X_2 and Z, of which Z is uncorrelated with the potential outcomes. As a result, Z is a valid instrumental variable for D.

The simulation is designed as follows. Firstly, suppose that we are able to observe both variables X_1 and X_{2A}, then the methods of matching and regression will be used to estimate the program impact. Secondly, suppose that only X_{2A} is observable, and X_1 is omitted in program impact estimation. The methods that will be used in this case are instrumental variable, sample selection, and panel data. Z will be selected as an instrumental variable. However, the methods of regression and matching are still used so that we can examine the selection bias due to the omitted variable X_1. The value of g and k in (6.4) will be changed to investigate the sensitivity of different methods to the omitted variable and instrumental variable.

6.2. Simulation results

Table 1 and 2 present the simulation results of estimation of ATT of program D using different estimators. Table 1 examines sensitivity of bias of the estimators of matching and regression to the role of the omitted variable, X_1. Table 2 investigates how the estimators of instrumental variable and sample selection work as the correlation of Z and D is changed. In each table, there are three panels corresponding to the values of g and k. In Table 1, k is set equal to 1, and g is changed from 0.5, to 1 and 2. In Table, g is fixed at 1, while k is changed through 0.5, 1 and 2.

When both X_1 and X_2 are used, the methods of regression and matching are used. There are two matching schemes. Matching 1 means the propensity score matching with 1 nearest neighbor, and matching 2 means the propensity score matching with 3 nearest neighbors. When the variable X_1 is omitted, the methods of regression and matching are still used to examine the bias of these methods when the assumption on the conditional independence does not hold. Then, estimation results from 7 estimators that do not assume the conditional independence are presented. There is a instrumental variable estimator in which the instrument is the variable Z. Sample selection 1 is the sample selection method using maximum likelihood estimation without the variable Z. Sample selection 2 is the maximum likelihood estimator with the variable Z, and sample selection 3 is the two-stage estimator with the variable Z. There are three estimators using panel data. First-difference is the linear regression using the differenced data. Diff-in-diff 1 means the difference-in-difference with propensity score matching using 1 nearest neighbors. Diff-in-diff 2 means the difference-in-difference with propensity score matching using 3 nearest neighbors.

It is showed in Tables 1 and 2 that in terms of MSE, the regression methods perform best since the models are correctly specified. The matching and regression will have rather low MSE when both X_1 and X_2 are controlled. The sample selection estimator without an instrument has the largest MSE as compared with other methods. However, since it is a consistent estimator, its MES tends to decrease when the sample size increase from 500 to 5000. The instrumental variable estimator has the second largest MSE.

Table 1 shows the bias of the regression and matching estimators increases remarkably as the correlation between X_1 and D increases with parameter g rising from 0.5 to 2. It is implied that bias tends to be higher if more variables that affect both the program selection and outcome are omitted. In Table 2, the correlation between Z and D is increased by raising parameter k from 0.5 to 2. As a result, MSE of the instrumental variable estimator is reduced significantly, especially in the small sample.

7. Conclusions

The main issue in impact evaluation is missing data. We cannot observe subjects at the same time in both statuses: participation in a program and non-participation in the program. Unless the program is randomized, the missing data is not random. Subjects are selected in the program based on their decisions and program administrators' decisions. Different methods in impact evaluation rely on different assumptions on the relation between the outcome process and the program selection process to construct the counterfactual so that the program impacts are identified. The paper discusses alternative methods in terms of identification assumptions and estimation strategies in contexts of the two potential outcome equations and program selection equations with allowance for heterogeneous program impacts. The main parameters of interest in impact evaluation that are examined in this paper are ATE and ATT. In addition to a randomization-based method in which participants are selected randomly, these methods are categorized into: (1) methods assuming "selection on observable", and (2) methods assuming "selection on unobservable". If impact of factors that can affect subjects is correlated with impact of a program of interest, we need to separate the program impact. "Selection on observable" methods are based on an assumption that we can observe all these correlated factors. In contrast, if

we are not able to observe all the correlated factors, we need to resort "selection on unobservable" methods.

Finally, measurement of program impact are often very complicated. In reality, the treatment variable D can be continuous instead of binary. Besides the program to be assessed, there might be many other programs that can affect the participants and non-participants of the program in question. Unless the program selection of others programs is uncorrelated with the selection program of the program to be assessed, the omission of other contemporaneous programs can lead to serious bias. Furthermore, subjects can participate in a program, e.g. training program or micro-credit program several times. Even if they are allowed to participate in a program one time, they can joint the program at different points of time. However, data on subjects' outcome are often collected at the same point of times. Ignorance of these issues can make the results from impact evaluation misinterpreted. All of these issues require further study to improve the literature on program impact evaluation.

Reference

Abadie, A. and G. W. Imbens (2002), "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," unpublished manuscript.

Buddelmeyer, H. and E. Skoufias (2004), "An Evaluation of the Performance of Regression Discontinuity Design in Progresa." *World Bank Policy Research Working Paper* No. 3386.

Cochran, W. G. (1968), "The effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies". Biometrics 24, 295-313.

Cochran, W. G. and S. Paul Chambers (1965), "The Planning of Observational Studies of Human Population (with discussion" *Journal of the Royal Statistical Society. Series A (General)*, Vol. 128, No. 2. (1965), pp. 234-266.

Dawid, A. P. (1979), "Conditional Independence in Statistical Theory", *J. R. Statist. Soc*., 41, No. 1: 1-31.

Dehejia R. H. and S. Wahba (1998), "Propensity Score Matching Methods for Non-Experimental Causal Studies", NBER Working Paper 6829, Cambridge, Mass.

Fan, J. (1992), "Local Linear Regression Smoothers and their Minimax Efficiencies". *The Annals of Statistics* 21, 196-216.

Greene W. H. (2003), *Econometric Analysis,* Prentice Hall Press, Firth Edition, 2003.

Hahn, J., (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

Hahn, J., P. Todd and W. van der Klaauw (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1): 201-09.

Heckman, J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System" *Econometrica*, 46, 931-959.

Heckman, J., H. Ichimura, J. A. Smith, and P. E. Todd (1998a), "Characterizing Selection Bias Using Experimental Dada," *Econometrica*, Vol. 66, 1017-1098.

Heckman, J., J. Smith, and N. Clements (1997a), "Making the Most Out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in Program Impacts", *Review of Economic Studies,* (1997) 64, 487 – 535.

Heckman, J., L. Lochner and C. Taber (1998a), "General Equilibrium Treatment Effects: A Study of Tuition Policy", *American Economic Review*. 88(2):381-386.

Heckman, J., R. Lalonde and J. Smith (1999), "The Economics and Econometrics of Active Labor Market Programs," *Handbook of Labor Economics, Volume 3*, Ashenfelter, A. and D. Card, eds., Amsterdam: Elsevier Science.

Heckman, J., N. Hohmann, M. Khoo and J. Smith (2000), "Substitution and Dropout Bias In Social Experiments: Evidence from an Influential Social Experiment", *The Quarterly Journal of Economics*, May 2000

Heckman, J. and E. J. Vytlacil (1999), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects", *Proceeding of National Academy of Science*, Vol.96, 4730-4734.

Heckman, J., H. Ichimura, and P. Todd (1997b), "Matching as an Econometric Evaluation Estimators: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, 64 (4), 605- 654.

Hirano K., G. W. Imbens and G. Ridder (2002), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Working Paper*

Imbens G. W. and J. D. Angrist (1994), "Identification and Estimation of Local Average Treatment Effect", *Econometrica,* Vol. 62, No. 2, 467-475.

Mahalanobis P. C. (1936), "On the Generalized Distance in Statistics" *Proc. Nat. Inst. Sci. Ind*., Vol. 12 (1936), 49-55.

Moffitt, R. (1991), "Program Evaluation with Nonexperimental Data." *Evaluation Review*. 15(3). 291-314.

Powell, J. (1994), "Estimation of Semiparametric Models", in: R. Engle and D. McFadden, eds., *Handbook of Econometrics*, vol. 4 (North-Holland, Amsterdam, Netherlands) 2443-2521.

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

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

Rosenbaum, P. and R. Rubin (1984), "Reducing Bias in Observation Studies Using Subclassification on the Propensity Score", *Journal of Statistical Association*, Vol. 79, No. 387, September 1984, 516-523.

Rosenbaum, P. and R. Rubin (1985a), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score", *American Statistician*, 39 (1), 33-38.

Rosenbaum, P. and R. Rubin (1985b), "The Bias due to Incomplete Matching", *Biometrics*, Vol. 41, No. 1, March 1984, 103-116.

Roy, A. (1951), "Some Thoughts on the Distribution of Earnings", *Oxford Economic Papers* 3:135-146.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies", *Journal of Educational Psychology* 66:688-701.

Rubin, D. (1977), "Assignment to a Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, 2 (1), 1-26.

Rubin, D. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization", *The Annals of Statistics* 6, 34-58.

Rubin, D. (1979), "Using Multivariate Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74: 318–328.

Rubin, D. (1980), "Bias Reduction Using Mahalanobis-Metric Matching." *Biometrics* 36 (2): 293–298.

Smith, J. and P. Todd. (2005), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics* 125 (1–2): 305–353.

Van der Klaauw, W. (2002), "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review* 43(4): 1249-87.

Wooldridge J. M. (2001), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, Massachusetts London, England

Table 1: Estimation of ATT using different estimators: selection bias due to omitting X_1

Measurement	N=500		N=1000		N=5000	
Model parameter: k = 1; g = 0.5						
Proportion with D=1		0.2129		0.2135		0.2142
ATT		8.1572		8.1586		8.1991
Observed outcome for non-participants		31.5736		31.5665		31.5623
Observed outcome for participants		32.7802		32.7708		32.8326
	MSE	**IM**	**MSE**	**IM**	**MSE**	**IM**
X_1 and X_2 are used						
Regression	0.3474	1.0002	0.1996	1.0064	0.0370	1.0016
Matching 1	0.7532	0.9930	0.4269	1.0019	0.0809	1.0006
Matching 2	0.5645	0.9826	0.3020	1.0001	0.0556	0.9991
X_2 is used, X_1 is omitted						
Regression	7.9867	0.6676	7.1929	0.6802	7.1568	0.6754
Matching 1	9.2574	0.6569	8.0328	0.6725	7.6301	0.6715
Matching 2	8.7000	0.6579	7.5912	0.6745	7.5403	0.6693
IV method	3.9504	0.9994	1.9301	1.0293	0.4083	1.0160
Sample selection 1	18.7727	0.7865	12.1640	0.8133	3.6363	0.8548
Sample selection 2	2.3279	0.9672	1.1600	0.9770	0.2504	0.9742
Sample selection 3	2.3757	0.9497	1.2278	0.9675	0.2909	0.9679
First-difference	0.7007	1.0054	0.3312	1.0074	0.0679	1.0014
Diff-in-Diff 1	1.4905	1.0063	0.6602	1.0092	0.1405	1.0015
Diff-in-Diff 2	0.8889	1.0045	0.4224	1.0097	0.0868	1.0020
Model parameter: k = 1; g = 1						
Proportion with D=1		0.2072		0.2067		0.2067
ATT		8.3740		8.3932		8.3476
Observed outcome for non-participants		31.7920		31.8077		31.8166
Observed outcome for participants		31.7359		31.7710		31.7665
	MSE	**IM**	**MSE**	**IM**	**MSE**	**IM**
X_1 and X_2 are used						
Regression	0.4372	1.0021	0.2215	1.0000	0.0417	1.0006
Matching 1	1.1020	0.9877	0.5206	0.9937	0.0908	0.9989
Matching 2	0.8258	0.9708	0.3764	0.9873	0.0747	0.9962
X_2 is used, X_1 is omitted						
Regression	23.0570	0.4333	22.5919	0.4369	22.0575	0.4381
Matching 1	24.5679	0.4253	23.7076	0.4293	23.5033	0.4229
Matching 2	23.9210	0.4247	23.3541	0.4291	22.8178	0.4295
IV method	5.2825	1.0044	2.8733	1.0222	0.4590	1.0144
Sample selection 1	31.1031	0.6974	24.3219	0.7005	5.1599	0.8797
Sample selection 2	2.9061	0.9597	1.3888	0.9662	0.2991	0.9710
Sample selection 3	3.7438	0.9201	2.0336	0.9318	0.5586	0.9413
First-difference	0.7591	1.0010	0.3792	1.0006	0.0710	1.0002
Diff-in-Diff 1	1.3328	0.9960	0.6737	1.0028	0.1306	0.9995
Diff-in-Diff 2	0.8872	0.9998	0.4482	1.0009	0.0901	1.0008

IM: mean ratio of the impact estimate over the true impact.
MSE: mean-squared-error.
n: number of observations
Number of replications: 500

Table 1: Continued

Measurement	N=500		N=1000		N=5000	
Model parameter: k = 1; g = 2						
Proportion with D=1		0.2051		0.2052		0.2052
ATT		8.8183		8.8080		8.8024
Observed outcome for non-participants		32.1727		32.1805		32.1778
Observed outcome for participants		31.4752		31.4967		31.4848
	MSE	**IM**	**MSE**	**IM**	**MSE**	**IM**
X_1 and X_2 are used						
Regression	0.4868	1.0013	0.2261	1.0030	0.0481	0.9983
Matching 1	2.4676	0.9504	1.3425	0.9700	0.2954	0.9860
Matching 2	1.8609	0.9308	0.9009	0.9570	0.2189	0.9811
X_2 is used, X_1 is omitted						
Regression	46.3182	0.2307	46.2808	0.2287	46.0117	0.2296
Matching 1	46.7752	0.2309	47.4468	0.2219	46.2971	0.2283
Matching 2	46.4337	0.2307	47.0295	0.2234	46.3040	0.2274
IV method	9.0515	1.0422	3.9274	1.0356	0.9103	1.0222
Sample selection 1	59.9310	0.4987	45.9964	0.6268	23.9502	0.7762
Sample selection 2	2.7771	0.9874	1.3365	1.0126	0.2582	0.9893
Sample selection 3	6.0542	0.9349	2.7274	0.9521	0.8743	0.9431
First-difference	0.6225	1.0056	0.3243	1.0019	0.0619	0.9992
Diff-in-Diff 1	1.1915	1.0067	0.6026	0.9974	0.1276	0.9986
Diff-in-Diff 2	0.7955	1.0057	0.3762	0.9986	0.0747	0.9990

Table 2: Estimation of ATT using different estimators: the role of the instrumental variable Z

Measurement	**N=500**		**N=1000**		**N=5000**	
Model parameter: k = 0.5; g = 1						
Proportion with D=1		0.2155		0.2142		0.2136
ATT		8.2014		8.1741		8.1794
Observed outcome for non-participants		32.0339		32.0550		32.0630
Observed outcome for participants		31.0088		31.0014		30.9982
	MSE	**IM**	**MSE**	**IM**	**MSE**	**IM**
X_1 and X_2 are used						
Regression	0.4550	1.0066	0.2123	1.0061	0.0486	1.0010
Matching 1	1.4092	0.9769	0.7674	0.9995	0.1592	0.9936
Matching 2	1.0157	0.9598	0.4956	0.9855	0.1204	0.9920
X_2 is used, X_1 is omitted						
Regression	29.0957	0.3477	28.0584	0.3549	28.4994	0.3479
Matching 1	31.7644	0.3288	29.5871	0.3425	30.7746	0.3257
Matching 2	31.0796	0.3281	29.3458	0.3418	30.3038	0.3286
IV method	17.4283	1.0412	8.9465	1.0840	1.8401	1.0503
Sample selection 1	23.0359	0.7337	18.3429	0.7260	2.7907	0.8628
Sample selection 2	6.9304	0.9019	3.3401	0.9056	0.8159	0.9266
Sample selection 3	8.3259	0.8340	5.0899	0.8366	2.1682	0.8491
First-difference	0.7676	1.0054	0.3513	1.0066	0.0915	1.0016
Diff-in-Diff 1	1.4104	1.0003	0.6972	1.0065	0.1690	1.0016
Diff-in-Diff 2	0.9183	1.0044	0.4595	1.0080	0.1100	1.0016
Model parameter: k = 1; g = 1						
Proportion with D=1		0.2072		0.2067		0.2067
ATT		8.3740		8.3932		8.3476
Observed outcome for non-participants		31.7920		31.8077		31.8166
Observed outcome for participants		31.7359		31.7710		31.7665
	MSE	**IM**	**MSE**	**IM**	**MSE**	**IM**
X_1 and X_2 are used						
Regression	0.4372	1.0021	0.2215	1.0000	0.0417	1.0006
Matching 1	1.1020	0.9877	0.5206	0.9937	0.0908	0.9989
Matching 2	0.8258	0.9708	0.3764	0.9873	0.0747	0.9962
X_2 is used, X_1 is omitted						
Regression	23.0570	0.4333	22.5919	0.4369	22.0575	0.4381
Matching 1	24.5679	0.4253	23.7076	0.4293	23.5033	0.4229
Matching 2	23.9210	0.4247	23.3541	0.4291	22.8178	0.4295
IV method	5.2825	1.0044	2.8733	1.0222	0.4590	1.0144
Sample selection 1	31.1031	0.6974	24.3219	0.7005	5.1599	0.8797
Sample selection 2	2.9061	0.9597	1.3888	0.9662	0.2991	0.9710
Sample selection 3	3.7438	0.9201	2.0336	0.9318	0.5586	0.9413
First-difference	0.7591	1.0010	0.3792	1.0006	0.0710	1.0002
Diff-in-Diff 1	1.3328	0.9960	0.6737	1.0028	0.1306	0.9995
Diff-in-Diff 2	0.8872	0.9998	0.4482	1.0009	0.0901	1.0008

Table 2: Continued

Measurement	N=500		N=1000		N=5000	
Model parameter: k = 2; g = 1						
Proportion with D=1		0.2003		0.1978		0.1986
ATT		8.7200		8.7468		8.7352
Observed outcome for non-participants		31.3448		31.3514		31.3568
Observed outcome for participants		33.6909		33.6948		33.7069
	MSE	**IM**	**MSE**	**IM**	**MSE**	**IM**
X_1 and X_2 are used						
Regression	0.3609	1.0055	0.1905	0.9996	0.0362	1.0003
Matching 1	0.7137	1.0023	0.3670	0.9970	0.0703	1.0003
Matching 2	0.5209	0.9940	0.2481	0.9942	0.0493	0.9993
X_2 is used, X_1 is omitted						
Regression	11.7778	0.6164	11.8345	0.6115	11.5968	0.6112
Matching 1	12.1054	0.6211	12.2814	0.6110	11.7931	0.6112
Matching 2	11.9005	0.6175	12.1397	0.6089	11.7044	0.6106
IV method	1.9908	1.0061	1.0126	1.0012	0.2219	1.0002
Sample selection 1	34.2011	0.7010	28.9788	0.7563	15.4823	0.8223
Sample selection 2	1.2352	1.0021	0.6164	0.9993	0.1249	0.9956
Sample selection 3	1.3569	0.9874	0.7139	0.9855	0.1660	0.9859
First-difference	0.5989	1.0066	0.2970	1.0011	0.0795	1.0019
Diff-in-Diff 1	1.0943	1.0021	0.6126	0.9964	0.1390	1.0016
Diff-in-Diff 2	0.7586	1.0049	0.3712	0.9992	0.0899	1.0012