Alternative Approaches to Evaluation in Empirical Microeconomics

Richard Blundell* and Monica Costa Dias
Institute for Fiscal Studies
August 2007

Abstract
This paper reviews a range of the most popular applied empirical evaluation methods: social experiments, natural experiments, matching methods, instrumental variables, discontinuity design and control functions. It discusses the identification of both the traditionally used average parameters and the more demanding distributional parameters. In each case, the necessary assumptions and the data requirements are considered. The adequacy of each approach is discussed drawing on the empirical evidence from the education and labour market policy evaluation literature.

Keywords: Evaluation methods, matching, instrumental variables, social experiments, natural experiments, discontinuity design, control functions.
JEL Classification: J21, J64, C33.

Acknowledgements: This research is part of the programme of work at the ESRC Centre for the Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies. We would like to thank the ESRC for financial support. The usual disclaimer applies.

Contents

1 Introduction 1

*Address: University College London and Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE. r.blundell@ucl.ac.uk, http://www ifs org.uk.
2 Which Treatment Parameter? 6
2.1 Average Treatment Effects ........................................... 6
2.2 The selection problem and the assignment rule ....................... 9
2.3 An example: returns to education ...................................... 10
2.3.1 Homogeneous treatment effects ...................................... 12
2.3.2 Heterogeneous treatment effects ...................................... 13

3 Social Experiments 14
3.1 Random assignment ...................................................... 14
3.2 Recovering the average return to education ............................ 16

4 Natural Experiments 16
4.1 The difference-in-differences (DID) estimator ............................ 16
4.2 Weaknesses of DID ...................................................... 20
4.2.1 Selection on idiosyncratic temporary shocks: the Ashenfelter’s dip ... 20
4.2.2 Differential macro trends ............................................ 20
4.2.3 Compositional changes over time .................................... 22
4.3 Non-linear models ....................................................... 22
4.4 Using DID to estimate returns to education ............................ 25
4.4.1 Monte-Carlo results ................................................... 30

5 Matching Methods 34
5.1 The matching estimator (M) .............................................. 34
5.2 Propensity score matching ................................................ 38
5.2.1 The popular alternative to matching: the linear regression model ... 41
5.3 Weaknesses of matching .................................................. 43
5.4 Using matching to estimate the returns to education .................... 44
5.4.1 Monte-Carlo results ................................................... 44
5.5 Combining matching and DID (MDID) ................................... 47
6 Instrumental Variables

6.1 The instrumental variables (IV) estimator ........................................ 49
6.2 Weaknesses of IV ............................................................................. 51
6.3 The LATE parameter ....................................................................... 52
   6.3.1 The LATE assumptions ............................................................... 55
   6.3.2 What does LATE measure? .......................................................... 56
6.4 The Marginal Treatment Effect .......................................................... 57
6.5 Using IV to estimate the returns to education ..................................... 59

7 Discontinuity Design

7.1 The discontinuity design estimator (DD) .......................................... 63
7.2 The link between discontinuity design and IV .................................... 65
7.3 Weaknesses of discontinuity design .................................................. 66
7.4 Using discontinuity design to estimate the returns to education .......... 67

8 Control Function Methods

8.1 The Control Function Estimator (CF) ............................................ 69
8.2 Weaknesses of the control function method .................................... 71
8.3 The link between the control function and the instrumental variables approach .... 71
8.4 Using the control function approach to estimate the returns to education .... 74

9 Summary ......................................................................................... 76
1 Introduction

The aim of this paper is to examine alternative evaluation methods in microeconomic policy analysis and to lay out the assumptions on which they rest within a common framework. The focus is on application to the evaluation of policy interventions associated with welfare programs, training programs, wage subsidy programs and tax-credit programs. At the heart of this kind of policy evaluation is a missing data problem. An individual may either be in a programme or may not, but no one individual can be in both states simultaneously. Indeed, there would be no evaluation problem of the type discussed here if we could observe the counterfactual outcome for those in the programme had they not participated. Constructing this counterfactual in a convincing way is a key ingredient of any serious evaluation method.

Which of the many available evaluation methods is appropriate for the policy or program under analysis? We will conclude that no single method is ‘best’ for all program evaluations. The choice will depend on three broad concerns: the nature of the question to be answered; the type and quality of data available; and the mechanism by which individuals are allocated to the program or receive the policy. This latter mechanism is typically labeled the ‘assignment rule’ and will be an important component in the analysis we present. In a perfectly designed social experiment assignment is random. In a structural microeconomic model assignment is assumed to obey rules from an appropriate economic theory. Alternative methods exploit different assumptions concerning assignment and differ according to the type of assumption made. Unless there is a convincing case for the reliability of the assignment mechanism being used, the results of the evaluation are unlikely to convince the thoughtful skeptic. Just as an experiment needs to be carefully designed a structural economic model needs to be carefully argued.

In this review we consider six distinct, but related, approaches: (i) social experiments methods, (ii) natural experiments, (iii) discontinuity design methods, (iv) matching methods, (v) instrumental methods and (vi) control function methods. The first of these approaches is closest to the ‘theory’ free method of a clinical trial, relying on the availability of a randomized assignment rule. The control function approach is closest to the structural econometric approach, directly modelling the
assignment rule in order to fully control for selection in observational data.\(^1\) The other methods can be thought of lying somewhere in between often attempting to mimic the randomized assignment of the experimental setting but doing so with non-experimental data. Natural experiments exploit randomisation to programs created through some naturally occurring event external to the researcher. Discontinuity design methods exploit ‘natural’ discontinuities in the rules used to assign individuals to receive a treatment. Matching attempts to reproduce the treatment group among the non-treated, this way re-establishing the experimental conditions in a non-experimental setting, but relies on observable variables to account for selection. Instrumental variables is a step closer to the structural method, relying on exclusion restrictions to achieve identification. Exactly what parameters of interest, if any, can be recovered by each method will typically relate to the specific environment in which the policy or programme is being conducted.

In many ways the social experiment method is the most convincing method of evaluation since it directly constructs a control (or comparison) group which is a randomized subset of the eligible population. The advantages of experimental data are discussed in papers by Bassi (1983,1984) and Hausman and Wise (1985) and were based on earlier statistical experimental developments (see Cockrane and Rubin (1973) and Fisher (1951), for example). Although a properly designed social experiment can overcome the missing data problem, in economic experiments it is frequently difficult to ensure that the experimental conditions have been met and not all parameters of interest are recovered from the experimental comparison. Since programs are typically voluntary those individuals ‘randomized in’ may decide not to participate in the treatment. The measured program impact will therefore recover an ‘intention to treat’ parameter, rather than the actual treatment effect. Further, unlike in many clinical trials, it is not possible to offer the control group a placebo in economic policy evaluations. Consequently individuals who enter a program and then are ‘randomized out’ may suffer a ‘disappointment’ effect and alter their behavior. Nonetheless, well designed experiments have much to offer in enhancing our knowledge of the possible impact of policy reforms. Indeed, a comparison of results from non-experimental data can help assess appropriate methods where experimental data is

\(^1\)The examination of fully specified structural evaluation models is beyond the scope of this review but for many important ex-ante policy evaluations they are the dominant approach, see Blundell and MaCurdy (1999) for some examples in the evaluation of tax and welfare policy proposals.
not available. For example, the important studies by LaLonde (1986), Heckman, Ichimura and Todd (1998) and Heckman, Smith and Clements (1997) use experimental data to assess the reliability of comparison groups used in the evaluation of training programmes. An example of a well conducted social experiment is the Canadian Self Sufficiency Project (SSP) which was designed to measure the earnings and employment responses of single mothers on welfare to a time-limited earned income tax credit programme. This study produced invaluable evidence on the effectiveness of financial incentives in inducing welfare recipients into work (see Card and Robbins, 1998). We draw on the results of this, and other experimental studies, below.

The natural experiment approach attempts to find a naturally occurring comparison group that can mimic the properties of the control group in the properly designed experiment. This method is also often labelled “difference-in-differences” since it is usually implemented by comparing the difference in average behaviour before and after the reform for the eligible group with the before and after contrast for the comparison group. The evaluation of the ‘New Deal for the Young Unemployed’ in the UK is a good example of a research design suited to this approach. It was an initiative to provide work incentives to unemployed individuals aged 18 to 24. The program is mandatory and was rolled out in selected pilot areas prior to the national roll out. The Blundell, Costa Dias, Meghir and Van Reenen (2004) study investigates the impact of this programme by using similar 18-24 year olds in non-pilot areas as a comparison group. The difference-in-differences approach can be a powerful tool in measuring the average effect of the treatment on the treated. It does this by removing unobservable individual effects and common macro effects. This approach relies on the two critically important identifying assumptions of (i) common time effects across groups, and (ii) no systematic composition changes within each group.

The discontinuity design method exploits situations where the probability of enrollment into treatment changes discontinuously with some continuous variable. For example, where eligibility to an educational scholarship depends on parental income falling below some cut-off or achieving a specific test score. It turns out to be convenient to discuss this approach in the context of the instrumental variable estimator since the parameter identified by discontinuity design is a local average treatment effect similar to the IV case but is not necessarily the same parameter. We contrast the IV and
discontinuity design approaches.

The matching method has a long history in non-experimental evaluation (see Heckman, Ichimura and Todd (1997), Rosenbaum and Rubin (1985) and Rubin (1979)). The aim of matching is simple. It is to line-up comparison individuals according to sufficient observable factors that any comparison individual with the same value of these factors will have no systematic differences in their reaction to the policy reform. Multiple regression is a simple linear example of matching. For this ‘selection on observables’ approach a clear understanding of the determinants of assignment rule on which the matching is based is essential. The measurement of returns to education, where scores from prior ability tests are available in birth cohort studies, is a good example. As Lalonde (1986) and Heckman, Ichimura and Todd (1998) demonstrate, experimental data can help in evaluating the choice of matching variables. As we document below, matching methods have been extensively refined and their properties examined in the recent evaluation literature and they are now a valuable part of the evaluation toolbox.

The instrumental variable method is the standard econometric approach to endogeneity. It relies on finding a variable excluded from the outcome equation but which is also a determinant of the assignment rule. In the simple linear model, the IV estimator identifies the treatment effect removed of all the biases which emanate from a non-randomized control. However, in heterogeneous treatment effect models, in which the impact of the programme can differ in unobservable ways across participants, the IV estimator will only identify the average treatment effect under strong assumptions and ones that are unlikely to hold in practise. Work by Angrist and Imbens (1994) and Heckman and Vytlacil (1999) has provided an ingenious interpretation of the IV estimator in terms of local treatment effect parameters. We provide a review of these developments.

Finally, the control function method directly analysis the choice problem facing individuals deciding on programme participation. It is, therefore, closest to the a structural microeconomic analysis. The control function approach specifies the joint distribution of the assignment rule and treatment. It uses the specification of the assignment rule together with an excluded ‘instrument’ to derive a control function which when included in the outcome equation fully controls for selection. This approach relates directly to the selectivity estimator of Heckman (1979).
As already noted, structural microeconometric simulation models are perfectly suited for ex-ante policy simulation. Blundell and McCurdy (1999) provide a comprehensive survey and a discussion of the relationship between the structural choice approach and the evaluation approaches presented here. A fully specified structural model can be used to simulate the parameter being estimated by any of the nonexperimental estimators above. Naturally, such a structural model would depend on a more comprehensive set of prior assumptions and will be less robust to the structural assumptions. We provide a running example of a structural model of schooling choices within which to evaluate each of the non-experimental methods. In our concluding section we draw out the relationship between the evaluation treatment effect parameters and those estimated in structural models. Results from evaluation approaches described above can be usefully adapted to assess the validity of a structural evaluation model.

Throughout this paper we illustrate the evaluation approaches within our simple stochastic model of education enrollment. The aim is to measure the returns to education. Individuals differ with respect to educational attainment, which is partly determined by a subsidy policy and partly determined by other factors. This ‘workhorse’ model of education and earnings is used to generate a simulated dataset and examine the performance of different estimators under different conditions. The specification of the education model is described in full detail in the appendix.

The rest of paper is organized as follows. In the next section we lay out the different definitions of treatment parameters and ask: what are we trying to measure in program evaluation? In this section we also develop an education evaluation example which we carry through the discussion of each alternative approach. Sections 3 to 8 are the main focus of this paper and present a detailed comparison of the six alternative methods of evaluation we examine here. In each case we use a common framework for analysis and apply each method to the education evaluation model. The order in which we discuss the various approaches follows the sequence described above with one exception; we choose to discuss discontinuity design after instrumental variables in order to relate the approaches together. Indeed an organising principle we use throughout this review is to relate the assumptions underlying each approach to each other, so that the pros and cons of each can be

\footnote{In the labor market area, from which we draw heavily in this review, the groundbreaking papers were those by Ashenfelter (1978), Ashenfelter and Card (1985) and Heckman and Robb (1985, 1986).}
assessed in common environment. Finally, in section 9 we provide a short summary.

2 Which Treatment Parameter?

2.1 Average Treatment Effects

Are individual responses to a policy homogeneous or do responses differ across individuals? If the responses differ, do they differ in a systematic way? The distinction between homogenous and heterogeneous treatment responses is central to understand what parameters alternative evaluation methods measure. In the homogeneous linear model, common in elementary econometrics, there is only one impact of the program and it is one that would be common to participants and nonparticipants alike. In the heterogeneous model, the treated and non-treated may benefit differently from program participation. In this case, the treatment on the treated parameter will differ from the treatment on the untreated parameter or the average treatment effect. Indeed, we can define a whole distribution of the treatment effects. A common theme in this review will be to examine the aspects of this distribution that can be recovered by the different approaches.

To simplify the discussion we consider a model of potential outcomes. In what follows we use upper case to denote vectors of random variables and lower case to denote random variables. We reserve Greek letters to denote the unknown parameters of the model.

Suppose we wish to measure the impact of treatment on an outcome, $y$. For the moment, we abstract from other covariates that may impact on $y$. Such covariates will be included later on. Denote by $d$ the treatment indicator: a dummy variable assuming the value 1 if the individual has been treated and 0 otherwise. The potential outcomes for individual $i$ at any time $t$ are denoted by $y^1_{it}$ and $y^0_{it}$ for the treated and non-treated scenarios, respectively. They are specified as

\[
\begin{align*}
  y^1_{it} &= \beta + \alpha_i + u_{it} \\
  y^0_{it} &= \beta + u_{it}
\end{align*}
\]

where $\beta$ is the intercept parameter, $\alpha_i$ is the effect of treatment on individual $i$ and $u$ is the unobservable component of $y$. The observable outcome is then

\[
y_{it} = d_{it} y^1_{it} + (1 - d_{it}) y^0_{it}.
\]
so that

\[ y_{it} = \beta + \alpha_i d_{it} + u_{it}. \]  

(3)

Notice that this is a very general model as, for now, we have not yet imposed any functional form or distributional assumptions on the components of the outcome. Different estimators use different sets of restrictions that we will then be discussed.

Selection into treatment determines the treatment status, \( d \). We assume this assignment occurs at a fixed moment in time, say \( k \), and depends on the information available at that time. This information is summarised by the observable variables, \( Z_k \), and unobservable, \( v_k \). Assignment to treatment is then assumed to be made on the basis of

\[ d_{it} = \begin{cases} 
1 & \text{if } d^*_{ik} > 0 \text{ and } t > k, \\
0 & \text{otherwise}
\end{cases} \]

(4)

where \( d^* \) is an index function

\[ d^*_{ik} = Z_{ik} \gamma + v_{ik} \]

(5)

in which \( \gamma \) is the vector of coefficients.

In this general specification, we have allowed for a heterogeneous impact of treatment, with \( \alpha \) varying freely across individuals.\(^3\) Estimation methods typically identify some average impact of treatment over some sub-population. The three most commonly used parameters are: the population average treatment effect (ATE), which would be the outcome if individuals were assigned at random to treatment, the average effect on individuals that were assigned to treatment (ATT) and the average effect on non-participants (ATNT). If it is the impact of the program on individuals of a certain type as if they were randomly assigned to treatment that is of interest, then ATE is the parameter to recover. On the other hand, the appropriate parameter to identify the impact of the program on individuals of a certain type that were assigned to treatment is the ATT.

Using the model specification above, we can express the average parameters at time \( t > k \) as

\(^3\)See, for example, Carneiro, Hansen and Heckman, 2001 and 2003, for a discussion of the distribution of treatment effects.
follows

\[
\alpha^{ATE} = E(\alpha_i) \\
\alpha^{ATT} = E(\alpha_i | d_{it} = 1) = E(\alpha_i | v_{ik} > -Z_{ik}\gamma) \\
\alpha^{ATNT} = E(\alpha_i | d_{it} = 0) = E(\alpha_i | v_{ik} < -Z_{ik}\gamma).
\]

Problems with the identification of these average parameters and increase interest on the distribution of treatment effects has led to the study of alternative parameters in the recent literature (Imbens and Angrist, 1994, Heckman and Vytlacil, 1999). Two particularly important parameters are the local average treatment effect (LATE) and the marginal treatment effect (MTE). To introduce them we need to assume that \(d^*\) is a non-trivial function of \(Z\). This means that at least one element in the set of parameters \(\gamma\) is non-zero and ensures that \(Z\) indeed affects the odds of participation such that variation in \(Z\) will affect the composition of the treatment group.

As is typical in evaluation studies, we also assume that \(v\) is a continuous random variable in \(\mathbb{R}\). Now suppose there exist two distinct values of \(Z\), say \(Z'\) and \(Z''\), such that \(Z'\gamma < Z''\gamma\). If \(v\) and \(Z\) are independent, meaning that the distribution of \(v\) does not change as \(Z\) changes, then some individuals will participate if drawing \(Z_{ik} = Z''\) but will not participate if drawing \(Z_{ik} = Z'\). The average impact of treatment on individuals that change their participation status when \(Z\) changes from \(Z'\) to \(Z''\) is the LATE parameter,

\[
\alpha^{LATE}(Z', Z'') = E(\alpha_i | v_{ik} > -Z''\gamma \text{ and } v_{ik} < -Z'\gamma)
\]

If now we set \(Z'\) and \(Z''\) arbitrarily close to each other, such that in the limit they are equal, \(Z' = Z'' = Z\), we obtain the MTE parameter,

\[
\alpha^{MTE}(Z) = E(\alpha_i | v_{ik} = -Z\gamma)
\]

This parameter measures the impact of participation among individuals that are indifferent about participation when drawing \(Z_{ik} = Z\). We will see later on that, under certain conditions, this parameter can be used to construct all other average parameters.

A distinctive feature of the LATE and MTE parameters under heterogeneous treatment effects is their dependence on the specific values of \(Z\) used regardless of whether \(Z\) and the treatment effect,
α, are related or not. This is a feature of the selection process, which blends participation costs and expected gains to determine the participation status. One would expect that the higher the participation cost (implying a lower, more negative \( Z_\gamma \)) the more concentrated will participants be towards the “high expected gains” (high \( \alpha \)) group and the higher the expected gains of agents that change participation status only at such high levels of cost.

All these parameters will be identical under homogeneous treatment effects. Under heterogeneous treatment effects, however, a non-random process of selection into treatment may lead to differences between them. However, whether the impact of treatment is homogeneous or heterogeneous, selection may affect our ability to identify the treatment effect parameter of interest.

2.2 The selection problem and the assignment rule

In non-experimental settings, assignment to treatment is most likely not random. Collecting all the unobserved heterogeneity terms together we can rewrite the outcome equation (2) as

\[
y_{it} = \beta + \alpha d_{it} + (u_{it} + d_{it} (\alpha_i - \bar{\alpha})) \\
= \beta + \alpha d_{it} + e_{it}.
\]

where \( \bar{\alpha} \) is the ATE parameter. Non-random selection occurs if the unobservable term \( e \) in (11) is correlated with \( d \). This implies that \( e \) is either correlated with the regressors determining assignment, \( Z \), or correlated with the unobservable component in the selection or assignment equation, \( v \). Consequently there are two types of non-random selection: selection on the observables and selection on the unobservables. When selection arises from a relationship between \( u \) and \( d \) we say there is selection on the non-treated outcomes as individuals with different untreated outcomes are differently likely to become treated. If, on the other hand, selection arises due to a relationship between \( \alpha \) and \( d \) we say there is selection on the (expected) gains, whereby individuals expecting to gain more from treatment are more likely to participate.

The result of selection is that the relationship between \( y \) and \( d \) is not directly observable from the data since participants and non-participants are not comparable. We will see later on that different estimators use different assumptions about the form of assignment and the nature of the impact.
to identify the treatment parameter of interest. Here we just illustrate the importance of some assumptions in determining the form and importance of selection by contrasting the homogeneous and heterogeneous treatment effect scenarios.

Under homogeneous treatment effects, selection bias occurs only if \( d \) is correlated with \( u \) since the outcome equation is reduced to

\[
y_{it} = \beta + \alpha d_{it} + u_{it}
\]

where \( \alpha \) is the impact of treatment on any individual. \( \alpha \) is constant across the population in this case. The OLS estimator will then identify

\[
E[\hat{\alpha}_{OLS}] = \alpha + E[u_{it}|d_{it} = 1] - E[u_{it}|d_{it} = 0]
\]

which is in general different from \( \alpha \) if \( d \) and \( u \) are related.

The selection process is expected to be more severe in the presence of heterogeneous treatment effects. The correlation between \( e \) and \( d \) may now arise through \( u \) (selection on non-treated outcomes) or through the idiosyncratic gains from treatment, \( \alpha_i - \overline{\alpha} \) (selection on gains). The parameter identified by the OLS estimator will now be

\[
E[\hat{\alpha}_{OLS}] = \overline{\alpha} + E[\alpha_i - \overline{\alpha}|d_{it} = 1] + E[u_{it}|d_{it} = 1] - E[u_{it}|d_{it} = 0]
\]

Note that the first term, \( \overline{\alpha} + E[\alpha_i - \overline{\alpha}|d_{it} = 1] \), is the ATT. Thus, even if \( d \) and \( u \) are not related, as long as \( E[d_{it} (\alpha_i - \overline{\alpha})] \neq 0 \), OLS will not recover the ATE. \( E[d_{it} (\alpha_i - \overline{\alpha})] \neq 0 \) implies that the idiosyncratic gains to treatment, \( \alpha_i \), are used in the participation decision itself.

### 2.3 An example: returns to education

Throughout this review we will use a dynamic model of educational choice and returns to education to illustrate the empirical use of each of the non-experimental methods. The model is solved and simulated under alternative conditions. The simulated data is then used to discuss the identification approach of each method and its ability to identify informative parameters. In the simulation exercise, the goal will be to measure the returns to education.

In the model, individuals differ with respect to educational attainment, which is determined by a number of observable and unobservable factors. Later on we will introduce an education subsidy
and explore its use in the context of natural experiments and instrumental variables. At this stage, however, we will only discuss the role of selection and heterogeneous effects in the evaluation problem. The model is described in full detail in the appendix.

We consider individuals indexed by $i$ facing lifetime earnings $y$ that depend, among other things, on education achievement. Individuals are heterogeneous at birth with respect to ability, $\theta$. Their lives are modeled in two periods, $t = 1, 2$. We assume there are two levels of education, low and high. The educational attainment is represented by the dummy variable $d$ where $d = 1$ for high education and $d = 0$ for low education. In period $t = 1$ the individual decides about investing in high education based on associated costs and expected gains from participation. The utility cost of education, $c$, depends on the observable characteristic, $b$, which we interpret as family background and the unobservable (to the researcher) $v$,

$$c_i = \delta_0 + \delta_1 b_i + v_i$$  \hspace{1cm} (12)

where $\delta_0$ and $\delta_1$ are some parameters.

Period $t = 2$ represents the individual’s working life. Lifetime earnings are realized, depending on ability, $\theta$, educational attainment, $d$, and the unobservable $u$. We assume that $u$ is unobservable to the researcher and is (partly) unpredictable by the individual at the time of deciding about education ($t = 1$). The logarithm of lifetime earnings is modeled as follows

$$\ln y_i = \beta + \alpha_1 d_i + \alpha_2 \theta_i d_i + u_i$$ \hspace{1cm} (13)

where $\beta$ is the intercept parameter and $\alpha_1$ and $\alpha_2$ are the treatment effect parameters for the general and ability-specific components, respectively.

As is obvious from the above equation, the returns to high education are heterogeneous in this model for as long as $\alpha_2 \neq 0$, in which case such returns depend on ability. The individual-specific return is

$$\alpha_i = \alpha_1 + \alpha_2 \theta_i$$

We assume $\theta_i$ is known by individual $i$ but not observable by the analyst. The educational decision of individual $i$ will be based on the comparison of expected lifetime earnings in the two alternative
The assignment (or selection) rule will therefore be
\[
d_i = \begin{cases} 
1 & \text{if } E[y_i|d_i = 1, \theta_i, v_i] - E[y_i|d_i = 0, \theta_i, v_i] > \delta_0 + \delta_1 b_i + v_i \\
0 & \text{otherwise}
\end{cases}
\]
so that investment in education occurs whenever the expected return exceeds the cost.

In this simple model, the education decision can be expressed by a threshold rule. Let \( \tilde{v} \) be the point at which individuals are indifferent between investing and not investing in education. It depends on the set of information available to the individual at the point of deciding, namely \((\theta, b)\). Then
\[
\tilde{v}(\theta_i, b_i) = E[y_i|d_i = 1, \theta_i, \tilde{v}(\theta_i, b_i)] - E[y_i|d_i = 0, \theta_i, \tilde{v}(\theta_i, b_i)] - \delta_0 - \delta_1 b_i.
\]
If tastes for education and work are positively related, \( v \) measures distaste for education and \( u \) measures unobserved productivity levels that are positively related with taste for work, then we expected \( v \) and \( u \) to be negatively correlated. This then means that, holding everything else constant, the higher \( v \) the higher the cost of education and the smaller the expected return from the investment. As \( v \) increases it will reach a point where the cost is high enough and the return is low enough for the individual to give up education. Thus, an individual \( i \) will follow the decision process,
\[
d_i = \begin{cases} 
1 & \text{if } v_i < \tilde{v}(\theta_i, b_i) \\
0 & \text{otherwise}
\end{cases}
\]
and this implies that educated individuals are disproportionately from the low-cost/high-return group.

2.3.1 Homogeneous treatment effects

Homogeneous treatment effects occur if the returns are constant across the population, that is either \( \alpha_2 = 0 \) or \( \theta_i = \theta \) over the whole population. In this case, the outcome equation (13) reduces to,
\[
\ln y_i = \beta + \alpha_1 d_i + u_i
\]
\[ \alpha^{ATE} = \alpha^{ATT} = \alpha^{ATNT} = \alpha_1 \text{ while } \alpha_1 \text{ also equals } \alpha^{LATE} \text{ and } \alpha^{MTE} \text{ for any choice of } b. \]

In this case, the selection mechanism simplifies to \( \tilde{v}(b_i) \). If, in addition, \( v \) and \( u \) are mean independent, the selection process will be exclusively based on the cost of education. In this case, OLS will identify the true treatment effect.

### 2.3.2 Heterogeneous treatment effects

Under heterogeneous treatment effects, education returns vary and selection into education will generally depend on expected gains. This causes differences in average treatment parameters. The ATE and ATT will now be,

\[
\begin{align*}
\alpha^{ATE} &= \alpha_1 + \alpha_2 E[\theta_i] \\
\alpha^{ATT} &= \alpha_1 + \alpha_2 E[\theta_i | v_i < \tilde{v}(\theta_i, b_i)]
\end{align*}
\]

If \( \alpha_2 \) is positive, meaning that high ability individuals will have higher returns from education, then the threshold rule \( \tilde{v} \) will be increasing in \( \theta \), meaning that higher ability individuals are also more likely to invest in education. This will then imply that the average ability among educated individuals is higher than the average ability in the population, \( E[\theta_i | v_i < \tilde{v}(\theta_i, b_i)] > E[\theta_i] \). But then, \( \alpha^{ATT} > \alpha^{ATE} \).

Assuming \( \theta \) is not observable by the analyst, the outcome equation (13) can be re-written as,

\[
\ln y_i = \beta + (\alpha_1 + \alpha_2 \overline{\theta}) d_i + (u_i + \alpha_2 d_i (\theta_i - \overline{\theta})).
\]

and OLS identifies

\[
E \left[ (\alpha_1 + \alpha_2 \overline{\theta})^{OLS} \right] = (\alpha_1 + \alpha_2 \overline{\theta}) + \alpha_2 E[\theta_i - \overline{\theta}|d_i = 1] + E[u_i|d_i = 1] - E[u_i|d_i = 0]
\]

\[
= \alpha_1 + \alpha_2 E[\theta_i|d_i = 1] + E[u_i|d_i = 1] - E[u_i|d_i = 0]
\]

This is the ATT if \( u \) and \( v \) are mean independent, while the ATE will not be identified by OLS (and is actually much harder to identify, as will become clear from the discussion of the empirical methodologies below).
3 Social Experiments

3.1 Random assignment

Suppose that an evaluation is proposed in which it is possible to run a social experiment that randomly chooses individuals from a group to be administered the treatment. If carefully implemented, random assignment provides the correct counterfactual, ruling out bias from self-selection. In the education model, a social experiment would randomly select potential students to be given some education while excluding the remaining individuals from the educational system. In this case, assignment to treatment would be random, and thus independent from the outcome or the treatment effect.

By implementing this sort of randomization, one ensures that the treated and the non-treated groups are equal in all aspects apart from the treatment status. In terms of the heterogeneous treatment effects model (3) we consider in this paper, randomisation corresponds to two key assumptions:

\[ \text{R1: } E[u_i|d_i = 1] = E[u_i|d_i = 0] = E[u_i] \]

\[ \text{R2: } E[\alpha_i|d_i = 1] = E[\alpha_i|d_i = 0] = E[\alpha_i] . \]

These randomisation ‘assumptions’ are required for recovering the average treatment effect (ATE).

Experiments are frequently impossible to implement. In many cases, such as in the education case, it is not conceivable that a government would agree to exclude/expose individuals from/to a given treatment at random. But even when possible, experimental information is frequently affected by two strong limitations. First, by excluding the selection behavior, experiments overlook intention to treat. However, the selection mechanism is expected to be strongly determined by the returns to treatment. In such case, the experimental results will not be generalizable to a economy-wide implementation of the treatment.

Second, a number of contaminating factors may interfere with quality of the information, affecting the experimental results. One possible problem concerns dropping-out behavior. For simplicity, suppose a proportion \( p \) of the eligible population used in the experiment prefer not to be treated and when drawn into the treatment group decide not to comply with treatment. Non-compliance might or not be observable, and this will determine the identifiable parameter.
Take the research design of a medical trial for a drug. The experimental group is split into treatments, who receive the drug, and controls, who receive a placebo. Without knowing whether they are treatments or controls, experimental participants will decide whether to take the medicine. A proportion $p$ of both groups will not take it. Suppose compliance is unrelated with the treatment effect, $\alpha$. If compliance is not observed, the identifiable treatment effect parameter is,

$$\tilde{\alpha} = (1 - p)E(\alpha)$$

which is a fraction of the ATE. If, on the other hand, compliance is observable, the ATE can be identified from the comparison of treatment and control compliers.

Unfortunately, non-compliance will unevenly affect treatments and controls in most economic experiments. Dropouts among the treated may correspond to agents that would not choose to be treated themselves if given the option; dropouts among the controls may be driven by many reasons, related or not to their own treatment preferences. As a consequence, the composition of the treatment and control groups conditional on (non)compliance will be different. It is also frequently the case that outcomes are not observable for the drop-outs. Taken together, these two conditions call for the use of non-experimental methods.

Another possible problem results from the complexity of contemporaneous welfare systems in developed countries and the availability of similar alternative treatments accessible to experimental controls. The experiment itself may affect experimental controls as, for instance, officers may try to “compensate” excluded agents with detailed information about other available treatments. This is another form of non-compliance, whereby controls obtain the treatment administered to experimental treatments.

Despite all the potential problems experiments might have, non-experimental data always requires special care. Except under very unlikely circumstances, the randomization conditions (R1) and (R2) do not hold. In what follows, we discuss a number of alternative identification hypothesis used with non-experimental data to recover the randomization hypothesis.
3.2 Recovering the average return to education

In the education example described in section 2.3, suppose we randomly select potential students to be given some specific education intervention while excluding the remaining students. In this case, assignment to treatment would be totally random, and thus independent from the outcome or the treatment effect. By implementing this sort of randomization, one ensures that the treated and the non-treated groups are in all equal apart from the treatment status. The randomization hypothesis (R1) and (R2) would be,

- \( E[u|d = 1] = E[u|d = 0] = E[u] \) and
- \( E[\theta|d = 1] = E[\theta|d = 0] = E[\theta] \).

These conditions are enough to identify the average returns to education in the experimental population using OLS,

\[
E\left( \alpha_1 + \alpha_2 \theta \right)^{OLS} = \alpha_1 + \alpha_2 \theta
\]

which is the ATE.\(^4\)

4 Natural Experiments

4.1 The difference-in-differences (DID) estimator

The natural experiment method makes use of naturally occurring phenomena that can be argued to induce some form of randomization across individuals in the eligibility or the assignment to treatment. Typically this method is implemented using a before and after comparison across groups. This is then formally equivalent to a difference-in-differences approach which uses some naturally occurring event to create a ‘policy’ shift for one group and not another. This may refer to a change of law in one jurisdiction but not another, it may refer to some natural disaster which changes a policy of interest

\(^4\)Notice that, given the dichotomous nature of the treatment we are considering, the OLS estimator in an experimental setting where the composition of the treatment and control groups is the same is given by the difference of means between the treated and control outcomes after treatment.
in one area but not another, or it may refer to the eligibility of a certain group to a change of policy for which a similar group is ineligible.

The difference between the two groups before and after the policy change is contrasted - thereby creating a difference-in-differences (DID) estimator of the policy impact. The DID estimator can either be applied to situations where there is longitudinal data, where the same individuals are followed over time, or to repeated cross section data, where samples are drawn from the same population before and after the intervention being studied. We start by considering the evaluation problem when the natural experiment involves longitudinal data. The DID estimator uses a decomposition of the error term to rewrite the outcome equation (2) as follows

\[ y_{it} = \beta + \alpha_i d_{it} + u_{it} \]  
where \( u_{it} \equiv \phi_i + \psi_t + \epsilon_{it} \)

where \( u \) is decomposed into three terms: an unobservable fixed effect, \( \phi \), an aggregate macro shock, \( \psi \), and an idiosyncratic transitory shock, \( \epsilon \).

Assume we observe individuals in two periods, before and after the policy change, designated by \( t = t_0 \) and \( t = t_1 \), respectively. For simplicity of notation, denote by \( d_i \) (without the time subscript) the treatment group, which is identified by the treatment status at \( t = t_1 \) (\( d_i \) is 1 among individuals that become treated at \( t = t_1 \) and is 0 among individuals that remain non-treated at \( t = t_1 \)). The main assumption underlying DID states that selection into treatment is independent of the temporary individual-specific effect, \( \epsilon_{it} \), so that

\[ E(u_{it} \mid d_i, t) = E(\phi_i \mid d_i) + \psi_t \]

for all \( i \) and \( t = t_0, t_1 \). That is, DID is based on the assumption that the randomization hypothesis (R1) holds in first differences

\[ E[u_{it_1} - u_{it_0} \mid d_i = 1] = E[u_{it_1} - u_{it_0} \mid d_i = 0] = E[u_{it_1} - u_{it_0}] . \]

This assumption does not rule out selection on the unobservables but restricts its source by ruling out the possibility of selection based on the transitory individual-specific effects \( \epsilon_{it} \). Also, it does not impose any conditions about selection on idiosyncratic gains from treatment that would mimic
the randomization hypothesis (R2). As a consequence, and as will be seen, it will only identify ATT under heterogeneous treatment effects in most cases.

Under the DID assumption we can write,

\[
E \left[ y_{d}^{t} \right] = \begin{cases} 
\beta + E [\alpha_{i} | d_{i} = 1] + E [\phi_{i} | d_{i} = 1] + \psi_{t_{1}} & \text{if } d = 1 \text{ and } t = t_{1} \\
\beta + E [\phi_{i} | d_{i} = d] + \psi_{t} & \text{otherwise}
\end{cases} 
\] (16)

where \( y_{d}^{t} \) is the average outcome over group \( d_{i} \) at time \( t \). It is now clear that we can eliminate both \( \beta \) and the remaining error components by sequential differences

\[
\hat{\alpha}^{DID} = [\bar{y}_{t_{1}}^{1} - \bar{y}_{t_{0}}^{1}] - [\bar{y}_{t_{1}}^{0} - \bar{y}_{t_{0}}^{0}] 
\] (17)

DID measures the excess outcome change for the treated as compared to the non-treated, this way identifying the ATT,

\[
E \left[ \hat{\alpha}^{DID} \right] = E [\alpha_{i} | d_{i} = 1] = \alpha^{ATT}.
\]

Notice that, under the DID assumption, the DID estimator is just the first differences estimator commonly applied to panel data when the presence of fixed effects is suspected. This means that an alternative way of obtaining \( \hat{\alpha}^{DID} \) is to take the first differences of (15) to obtain

\[
y_{it_{1}} - y_{it_{0}} = \alpha_{i}d_{it_{1}} + (\psi_{t_{1}} - \psi_{t_{0}}) + (\epsilon_{it_{1}} - \epsilon_{it_{0}})
\]

which can be consistently estimated using OLS. Notice also that the DID assumption implies that the transitory shocks \( \epsilon_{it} \) are uncorrelated with the treatment variable. Therefore, the standard within groups panel data estimator is analytically identical to the DID estimator of the ATT under these assumptions (see Blundell and MaCurdy (1999)).

Examining (16) it follows that repeated cross-sectional data would be enough to identify ATT for as long as treatment and control groups can be separated before the policy change, in period \( t = t_{0} \). Such information is sufficient for the average fixed effect per group to cancel out in the before after differences.

As an example, the DID approach has been used to study the impact of the ‘New Deal for the Young Unemployed’, a UK initiative to provide work incentives to individuals aged 18 to 24 and claiming Job Seekers Allowance (UI) for 6 months. The program was first introduced in January
1998, following the election of a new government in Britain in the previous year. It combines initial job search assistance followed by various subsidized options including wage subsidies to employers, temporary government jobs and full time education and training. Prior to the New Deal, young people in the UK could, in principle, claim unemployment benefits indefinitely. Now, after 6 months of unemployment, young people enter the New Deal ‘Gateway’, which is the first period of job search assistance. The program is mandatory, including the subsidized options part, which at least introduces an interval in the claiming spell.

The Blundell, Costa Dias, Meghir and Van Reenen (2004) study investigates the impact of the program on employment in the first 18 months of the scheme. In particular it exploits an important design feature by which the program was rolled out in certain pilot areas prior to the national roll out. Since the program is targeted at a specific age group, a natural comparison group is formed of similar individuals with corresponding unemployment spells but who are slightly too old to be eligible. A before and after comparison can then be made using a regular DID estimator. This can be improved by a matching DID estimator as detailed in section 5.5. These estimators are all implemented in the study. The pilot area based design also means that matched individuals of the same age can be used as an alternative control group.

The evaluation approach therefore consists of exploring sources of differential eligibility and different assumptions about the relationship between the outcome and the participation decision to identify the effects of the New Deal. On the ‘differential eligibility’ side, two potential sources of identification are used. First, the program is age-specific implies that using slightly older people of similar unemployment duration is a natural comparison group. Second, the program was first piloted for 3 months (January to March 1998) in selected areas before being implemented nation-wide (the ‘National Roll Out’ beginning April 1998). The same age group in non-pilot areas is not only likely to satisfy the quasi-experimental conditions more closely but also allows for an analysis of the degree to which the DID comparisons within the treatment areas suffer from both general equilibrium or market level biases and serious substitution effects. Substitution occurs if participants take (some of) the jobs that non-participants would have got in the absence of treatment. Equilibrium wage effects may occur when the program is wide enough to affect the wage pressure of eligible and ineligible
individuals.

The study focuses on the change in transitions from the unemployed claimant count to jobs during the Gateway period. It finds that the outflow rate for men has risen by about 20% as a result of the New Deal program. Similar results show up from the use of within area comparisons using ineligible age groups as controls and also from the use of individuals who satisfy the eligibility criteria but reside in non-pilot areas. Such an outcome suggests that either wage and substitution effects are not very strong or they broadly cancel each other out. The results appear to be robust to pre-program selectivity, changes in job quality and different cyclical effects.

4.2 Weaknesses of DID

4.2.1 Selection on idiosyncratic temporary shocks: the Ashenfelter’s dip

The DID procedure does not control for unobserved temporary individual-specific shocks that influence the participation decision. If \( \epsilon \) is not unrelated to \( d \), DID is inconsistent for the estimation of ATT and instead approximates the following parameter

\[
E (\hat{\alpha}^{DID}) = \alpha^{ATT} + E(\epsilon_{it_1} - \epsilon_{it_0} | d_{it_1} = 1) - E(\epsilon_{it_1} - \epsilon_{it_0} | d_{it_1} = 0)
\]

To illustrate the conditions such inconsistency might arise, suppose a training program is being evaluated in which enrolment is more likely if a temporary dip in earnings occurs just before the program takes place - the so-called Ashenfelter’s dip (see Ashenfelter, 1978, and Heckman and Smith, 1994). A faster earnings growth is expected among the treated, even without program participation. Thus, the DID estimator is likely to over-estimate the impact of treatment.

4.2.2 Differential macro trends

The identification of ATT using DID relies on the assumption that treated and controls experience the same macro shocks. If this is not the case, the DID approach will yield a biased and inconsistent estimate of ATT. Differential trends might arise in the evaluation of training programs if treated and controls operate in different labor markets. For example, unemployment in different age groups is often found to respond differently to cyclical fluctuations. In particular, unemployment among the
youngest is generally more volatile, responding more strongly to changes in macro conditions and thus exhibiting more pronounced rises and drops as the economy evolves.

Figure 1 illustrates what is meant by common trends. It refers to the ND study of Blundell et al (2004) and compares treated and controls over time with respect to the outflows from unemployment. The common trends assumption holds when the curves for treated and controls are parallel. In our example, the curves are nearly parallel over most of the period. The only important exception is at the beginning of the observable period. The graph suggests that both control groups considered in the study are valid.

The possibility of differential trends motivates the differential trend adjusted DID estimator. Suppose we suspect that the common trend assumption of DID does not hold but can assume that selection into treatment is independent of the temporary individual-specific effect, \( \epsilon_{it} \), under differential trends

\[
E(u_{it} | d_i = d, t) = E(\phi_i | d_i = d) + k^d \psi_t
\]

where \( k^d \) is a scalar allowing for differential macro effects across the two groups (\( d \) represents the group and is either 1 or 0).

The DID estimator now identifies

\[
E \left( \hat{\alpha}^{DID} \right) = \alpha^{ATT} + (k^1 - k^0) \left[ \psi_{t_1} - \psi_{t_0} \right]
\]

which does not recover the true ATT unless \( k^1 = k^0 \), in which case we are back to the standard DID assumption.

In the availability of data, one possible solution is to compare the trends of treated and controls historically, prior to the intervention. Historical, pre-reform data can help if there exists another time interval, say \( \tau_0 \) to \( \tau_1 \) (with \( \tau_0 < \tau_1 < k \)), over which a similar macro trend has occurred. In that case, by comparing the DID estimate of the impact of treatment contaminated with the bias from differential trend with the estimate of the differential trend over \( (\tau_0, \tau_1) \) one can separate the true impact of treatment from the differential trend.

More precisely, suppose one finds a pre-reform period, \( (\tau_0, \tau_1) \) for which the differential macro trend matches the bias term in the DID estimator, \( (k^1 - k^0)[\psi_{t_1} - \psi_{t_0}] \). That is,

\[
(k^1 - k^0) \left[ \psi_{t_1} - \psi_{t_0} \right] = (k^1 - k^0) \left[ \psi_{t_1} - \psi_{t_0} \right]
\]
This means that there is a point in history where the relative conditions of the two groups being compared, treatments and controls, evolves similarly to what they do in the pre-post reform period, \((t_0, t_1)\). Together with the absence of policy reforms that affect the outcome \(y\) during \((\tau_0, \tau_1)\), this condition allows one to identify the bias term \((k^1 - k^0)[\psi_{t_1} - \psi_{t_0}]\) by applying DID to that pre-reform period. The impact of treatment can now be isolated by comparing DID estimates for the two periods, \((t_0, t_1)\) and \((\tau_0, \tau_1)\). This is the differentially adjusted estimator proposed by Bell, Blundell and Van Reenen (1999), which will now consistently estimate ATT, \(\hat{\alpha} = \{[\overline{y}_{t_1} - \overline{y}_{t_0}] - [\overline{y}_{t_{1*}} - \overline{y}_{t_{0*}}]\} - \{[\overline{y}_{t_{1*}} - \overline{y}_{t_{1*}}] - [\overline{y}_{t_{0*}} - \overline{y}_{t_{0*}}]\}. \tag{18}\)

It is likely that the most recent cycle is the most appropriate, as earlier cycles may have systematically different effects across the target and comparison groups. The similarity of subsequent cycles, and thus the adequacy of differential adjusted DID, can be accessed in the presence of a long history of outcomes for the treatment and control groups.

### 4.2.3 Compositional changes over time

Although DID does not require longitudinal data to identify the true ATT parameter, it does require similar treatment and control groups to be followed over time. In particular, the composition of the groups with respect to the fixed effects term must remain unchanged to ensure before-after comparability. If before-after comparability does not hold, the DID will identify a parameter other than ATT. We will later see an example where this occurs within the simulation model.

### 4.3 Non-linear models

A restrictive feature of the DID method is the imposition of additive separability of the error term conditional on the observables. Recent studies have proposed ways of relaxing this assumption. In their analysis of the New Deal for the Young People, Blundell et al. noticed that linearity in the error term can be particularly unrealistic when the outcome of interest is a dummy variable. In such case, the DID method can conceivably predict probabilities outside the \([0, 1]\) range. Instead, the authors suggest using the popular index models and assuming linearity in the index. Unfortunately, even with a very simple non-linear specification, DID loses much of its simplicity.
To see how to estimate the impact of treatment in this case, suppose the outcome equation is now,

$$y_{it} = 1(\beta + \alpha_i d_{it} + u_{it} > 0)$$

(19)

where $1(A)$ is the indicator function, assuming the value 1 if $A$ is true and 0 otherwise. As before,

$$u_{it} = \phi_i + \psi_t + \epsilon_{it}$$

and the DID assumption holds,

$$E(u_{it} | d_i, t) = E(\phi_i | d_i) + \psi_t$$

where $d_i$ represents the treatment status in the after treatment period, $t_1$.

In what follows we assume $\epsilon$ follows a distribution $F$ where $F$ is invertible.\(^5\) We denote by $F^{-1}$ the inverse probability rule. We simplify the model further by assuming a common group effect instead of allowing for an individual-specific effect: it is assumed that $\phi_i = \phi_d$ for $d = 0, 1$ being the post-program treatment status of individual $i$.\(^6\)

Under these conditions and given a particular parametric assumption about the shape of $F$, say normal, one could think of mimicking the linear DID procedure by just running a probit regression of $y$ on $d$ and dummy variables for group and time (and possibly other exogenous regressors $x$) hoping this would identify some average of the treatment parameter $\alpha$. One would then average over the treated the impact on $y$ to recover the average treatment effect (the individual impact would depend on the point of the distribution where the individual is before treatment).

Unfortunately, this is generally not a valid approach. The problem is that the model contains still another error component which has not been restricted and that, under general conditions, will not fulfill the probit requirements. To see this, notice we can re-write model (19) as follows,

$$y_{it} = 1(\beta + \alpha_{ATE} d_{it} + \phi_d + \psi_t + \epsilon_{it} + d_{it} (\alpha_i - \alpha_{ATE}) > 0)$$

where $d_{it} (\alpha_i - \alpha_{ATE})$ is part of the error term. Standard estimation methods would require a distributional assumption for $(\alpha_i - \alpha_{ATE})$ and its independence from the treatment status.

\(^5\)More precisely, we are assuming the transitory shocks, $\epsilon$, are iid continuous random variables with a strictly increasing cumulative density function, $F$, which is assumed known.

\(^6\)This is generally required for non-linear discrete choice models (see Nickell, 1981).
Instead of imposing further restrictions in the model, we can follow by noticing that under the above stated assumptions,

\[ E(y_{0i} | d_i = d, t) = F(\beta + \phi_d + \psi_t) \]

where, as before, \((y^0, y^1)\) are the potential outcomes in the absence and in the presence of treatment, respectively. But then the index is recoverable given invertibility of the function \(F\),

\[ \beta + \phi_d + \psi_t = F^{-1}[E(y^0 | d_i = d, t)] \]

Using this result it is obvious that the trend can be identified by comparing non-treated before and after treatment since,

\[ \psi_{t1} - \psi_{t0} = F^{-1}[E(y^0_{1i} | d_i = 1, t_1)] - F^{-1}[E(y^0_{1i} | d_i = 0, t_0)] \]

and given the common trend assumption it is also true that, would we be able to observe the counterfactual of interest, \(E(y^0_{1i} | d_i = 1, t_1)\),

\[ \psi_{t1} - \psi_{t0} = F^{-1}[E(y^0_{1i} | d_i = 1, t_1)] - F^{-1}[E(y^0_{1i} | d_i = 1, t_0)] \]

But then, from (??) and (??) one notices that the unobserved counterfactual is,

\[
F^{-1}\left[ E\left( y_{0i}^0 | d_i = 1, t_1 \right) \right] = \\
F^{-1}\left[ E\left( y_{0i}^0 | d_i = 1, t_1 \right) \right] + \left\{ F^{-1}\left[ E\left( y_{0i}^0 | d_i = 0, t_1 \right) \right] - F^{-1}\left[ E\left( y_{0i}^0 | d_i = 0, t_0 \right) \right] \right\}
\]

Let the average parameter which measures the average impact of treatment in the inverse transformation of the expected outcomes be \(\overline{\alpha}\). Then\(^7\)

\[
\overline{\alpha} = \left\{ F^{-1}\left[ E\left( y_{1i}^1 | d_i = 1, t_1 \right) \right] - F^{-1}\left[ E\left( y_{0i}^0 | d_i = 1, t_1 \right) \right] \right\} \\
= \left\{ F^{-1}\left[ E\left( y_{1i}^1 | d_i = 1, t_1 \right) \right] - F^{-1}\left[ E\left( y_{0i}^0 | d_i = 1, t_0 \right) \right] \right\} - \\
\left\{ F^{-1}\left[ E\left( y_{1i}^1 | d_i = 0, t_1 \right) \right] - F^{-1}\left[ E\left( y_{0i}^0 | d_i = 0, t_0 \right) \right] \right\}
\]

\(^7\)Notice that \(\overline{\alpha}\) is not \(\alpha^{ATT}\) since \(F^{-1}\left[ E\left( y_{1i}^1 | d_i = 1, t_1 \right) \right]\) is generally different from the average index for this group and time period (which is \(\beta + \alpha^{ATT} + \phi_1 + \psi_{t1}\)) given the non-linearity of \(F^{-1}\) and the heterogenous nature of the treatment effect. To see why notice that,

\[ E\left[ y_{1i}^i | d_i = 1, t_1 \right] = \int_{\alpha} F(\beta + \alpha_i + \phi_1 + \psi_{t1}) dG(\alpha_i | i \in T) \]

where \(G(\alpha_i | i \in T)\) is the distribution of the treatment effect among he treated. Applying the inverse transformation
Rearranging, the missing counterfactual is
\[
E \left( y^0 | d = 1, t_1 \right) = F \left\{ F^{-1} \left[ E \left( y^1 | d = 1, t_1 \right) \right] - \overline{\alpha} \right\}
\]

Using this expression, the ATT can be estimated by replacing the expected values by their sample analogues,
\[
\widehat{ATT} = \overline{y}_{t_1} - F \left[ F^{-1} \left( \overline{y}_{t_1} \right) - \overline{\alpha} \right]
\]
where
\[
\overline{\alpha} = \left[ F^{-1} \left( \overline{y}_{t_1} \right) - F^{-1} \left( \overline{y}_{t_0} \right) \right] - \left[ F^{-1} \left( \overline{y}_{t_1} \right) - F^{-1} \left( \overline{y}_{t_0} \right) \right]
\]

Recently, Athey and Imbens (2006) have developed a general non-linear DID method specially suited for continuous outcomes: the “changes-in-changes” (CIC) estimator.\(^8\) The discussion of this method is outside the scope of this paper (we refer the interested reader to the original paper by Athey and Imbens, 2006).

### 4.4 Using DID to estimate returns to education

In general, DID is not suited to evaluate the returns to education. The problem is that longitudinal data including earnings of treated and controls before and after the treatment is rarely available since education occurs earlier in the life cycle than labor market participation. Alternatively, the comparison of different cohorts can only help identifying the treatment effect under special circumstances, when some exogenous change leads to differences in educational investments between cohorts.

To explore this latter alternative, we consider a small extension to our model with the introduction of an education subsidy with the following design. Eligibility to subsidized education depends on a test performance: in a earlier period in their life-cycle, which is denoted by \( t = 0 \), the student takes

\[
F^{-1} \left( E \left[ y^1_{it} | d_i = 1, t_1 \right] \right) = F^{-1} \left( \int_\alpha F(\beta + \alpha_i + \phi_i + \psi_{t_1}) dG(\alpha_i | i \in T) \right) 
\]

\[
\neq \int_\alpha F^{-1} \left( F(\beta + \alpha_i + \phi_i + \psi_{t_1}) \right) dG(\alpha_i | i \in T)
\]

However, it can be used to recover the ATT as exposed in the main text.

\(^8\) An extension to the discrete case is also considered by the authors.
a test. The test score, \( s \), depends on ability, \( \theta \), the level of parental input, previous schooling and individual effort assigned to the test preparation, \( e \), and an unpredictable (to the individual) and unobservable (to the researcher) component, \( w \)

\[
s_i = \gamma_0 + \gamma_1 \theta_i (1 + e_i) + w_i
\]

where \( \gamma_0 \) and \( \gamma_1 \) are some parameters. Effort \( e \) carries some utility cost, as described in Appendix A. The (stochastic) payoff to this effort is the possibility of accessing subsidized education.

To explore the information benefits of pilot studies, we assume there are two regions within the country where the subsidy policy is introduced. We denote region by \( x \) with possible values \( x = 0, 1 \). Earnings levels may differ across regions but we exclude the possibility of (differential) time trends. The new earnings equation for an individual \( i \) among the generation working at time \( t \) is,

\[
\ln y_{it} = \beta_0 + \beta_1 x_i + (\alpha_1 + \alpha_2 \theta_i) d_i + u_{it}
\]

where the distribution of \( u \) remains constant over time.

We first assume that the subsidy is launched at time \( k \) in region \( x = 1 \). From time \( k \) onwards, the cost of education in region \( x = 1 \) follows an adjusted form of equation (12),

\[
c_i = \delta_0 + \delta_1 b_i - 1(s_i > s)S + v_i
\]

where \( s \) is the threshold rule defining eligibility to the subsidy and \( S \) is the new subsidy.

The question now is: Can we explore this policy change in some regions to estimate the returns to education using DID? We start by noticing that enrollment into education is not solely determined by the subsidy. Some eligible individuals (individuals in region \( x = 1 \) making their education decisions after time \( t = k \)) will decide to enroll into education even if no subsidy is available, while other eligible individuals will opt out even in the presence of the subsidy. Some investment in education is also expected among the non-eligible or in the no-subsidy region \( (x = 0) \), although the cost of education for these is not altered by the policy change. Thus, there will be some educated individuals even when and where the subsidy is not available. As a result, the ATT will not be identified in general. Instead, the average impact of treatment on individuals that change their educational decisions in response to the subsidy may be identified.
To estimate the returns to education among individuals that change their education status in response to the subsidy, we further assume that the introduction of the educational subsidy does not lead anyone to give up education. Instead, it makes education more attractive for all eligibles and does not change the incentives to invest in education among non-eligibles.  

Define the treatment and control groups as those living in regions affected \((x = 1)\) and not affected \((x = 0)\) by the policy change. We designate these groups by \(T\) and \(C\), respectively. Now suppose we have data on educational attainment and earnings in treated and control areas for different cohorts of individuals, both before and after the policy change. We choose two cohorts, making educational decisions before and after the policy change. Let \(t = t_0\) and \(t = t_1\) represent the periods when earnings of unaffected and affected cohorts are observed, respectively. We then compare the \(T\) and \(C\) groups over time using DID.

Designate by \(\ln y_{T0}, \ln y_{T1}, \ln y_{C0}\) and \(\ln y_{C1}\) the average log earnings among \(T\) and \(C\) before and after the policy change. As before, \(d_{it}\) is a dummy variable indicating whether individual \(i\) in cohort \(t\) has acquired high education, and we define the probabilities

\[
p_{jt} = P(d_{it} = 1|i \text{ in } j)
\]

where \(i\) indexes individuals, \(j\) represents the region \((j = T, C)\) and \(t\) represents time \((t = t_0, t_1)\). Thus, \(p_{jt}\) is the odds of participation in region \(j\) at time \(t\). The stated assumption that education is at least as attractive in the presence of the subsidy implies that \(d_{it1} \geq d_{it0}\) for all \(i\) in \(T\) and, therefore, \(p_{T1} \geq p_{T0}\). In the control region we assume \(p_{C1} = p_{C0}\) for simplicity, meaning that no other factors differentially affect the education investments of cohorts \(t_0\) and \(t_1\).

Assuming the decomposition of the error term as in equation (15),

\[
\begin{align*}
    u_{it} &= \phi_i + \psi_t + \epsilon_{it} \\
    E [\ln y_{T1} - \ln y_{T0}] &= (\psi_{t_1} - \psi_{t_0}) + (p_{T1} - p_{T0}) E [\alpha_i|d_{it1} = 1, d_{it0} = 0, i \in T]
\end{align*}
\]

\(^9\)We discuss this type of monotonicity assumption in more detail later on, along with the LATE parameter.
meaning that only the impact on the movers is picked. Similarly,

\[ E[\ln y_{Ct1} - \ln y_{Ct0}] = (\psi_{t1} - \psi_{t0}) \]

since individuals in the control region do not alter their educational decisions. Thus, under the DID assumption we identify,

\[ E[\hat{\alpha}_{DID}] = (p_{Tt1} - p_{Tt0}) E[\alpha_i | d_{it1} = 1, d_{it0} = 0, i \in T] \]

(23)

showing that the average returns to education on the individuals moving into education in response to the subsidy can be identified by dividing the DID estimator by the proportion of movers in the treated region, \( p_{Tt1} - p_{Tt0} \). This will identify the LATE parameter: the impact of education on individuals changing their educational status in response to a policy change.

Not correcting for the proportion of movers in \( T \) implies that a different parameter is estimated: the average impact of introducing an education subsidy on earnings in the treated region. This is a mixture between a zero effect for those that do not move in response to the subsidy and the return to education for the movers.

Under homogeneous treatment effects, all average parameters are equal and thus ATE and ATT are also identified. However, under heterogeneous treatment effects only the impact on the movers can be identified and even this requires especial conditions. In this example we have ruled out movers in the control regions. Notice, however, that if other conditions differentially affect the educational decisions in non-treated regions before and after the policy intervention, there will be some movers among the controls. Whether the monotonicity assumption mentioned above holds for the control group or not depends on the circumstances that lead these individuals to move. For simplicity, we assume monotonicity holds in control areas such that \( d_{it1} \geq d_{it0} \) for \( i \in C \). The DID will identify

\[ E[\hat{\alpha}_{DID}] = (p_{Tt1} - p_{Tt0}) E[\alpha_i | d_{it1} = 1, d_{it0} = 0, i \in T] + (p_{Ct1} - p_{Ct0}) E[\alpha_i | d_{it1} = 1, d_{it0} = 0, i \in C] \]

In this case, the ability to single out the impact of treatment on some group of movers (movers in \( T \) net of movers in \( C \)) depends on two factors: (i) that movers in \( T \) in the absence of a policy change
would have the same returns to education as movers in \( C \), which typically requires that they are similar individuals; and (ii) that different proportions of individuals move in the T and C areas.

Now suppose that instead of a local policy, we are exploring the use of a global policy change, simultaneously introduced in the whole country. Instead of using treated and non-treated regions, one can think of using the eligibility rules as the source of randomization. Let us now define the treatment and control groups, \( T \) and \( C \) respectively, as composed by agents scoring above and below the eligibility threshold, \( s \). Again, we assume data is available on two cohorts, namely those affected and unaffected by the policy change.

The use of the eligibility rule instead of regional variation suffers, in this case, from one additional problem: the identification of the eligibility group before the introduction of the program. The affected generations will react to the new rules, adjusting their behavior even before their treatment status is revealed (which amounts to becoming eligible to the subsidy). In our model, future eligibility status can be influenced in anticipation by adjusting individual studying efforts in period 0. As a consequence, a change in the selection mechanism in response to the policy reform will affect the size and composition of the \( T \) and \( C \) groups over time as defined by the eligibility status. This means that \( T \) and \( C \) groups are not comparable over time and since we are confined to use repeated cross-sections to evaluate the impact of education, this would exclude the DID approach as a valid candidate method to the present evaluation exercise if only eligibility can be use as a source of randomization.

This is the problem identified by Abbring and van den Berg (2004) when the dynamic nature of labor market decisions is acknowledged. Individuals may react in anticipation of treatment, trying to explore the policy rules. If the rules change, anticipatory behavior may also change, thus rendering individuals with similar characteristics incomparable when such characteristics are affected by the endogenous selection behavior that is not explicitly modeled. Reactions in anticipation to treatment are not observable and tend to change over time. Their occurrence may create a problem similar to the Ashenfelter dip described above as their potential impact on the outcome will be absorbed by the transitory unobservable component. Treated and controls with similar pre-treatment characteristics and outcomes will be inherently different as observables are endogenously affected by the individuals prospects about treatment.
In our example, individuals may react to the new subsidy by increasing effort in the test, raising test performance on average and increasing the odds of becoming eligible to subsidized education. Thus, the ability distribution of eligibles will be affected by the policy change, not only the educational choice.

4.4.1 Monte-Carlo results

To illustrate the ability of DID to estimate the impact of treatment, we ran a Monte Carlo simulation. We tried different assumptions, depending on: (i) Whether or not the policy is experimented in some parts of the country before being nationally implemented; (ii) Whether or not the post-intervention generation has information about the policy change and (iii) Whether or not the unobservables \( v \) and \( u \) are correlated.\(^\text{10}\) We then estimate the impact of education in the alternative cases using DID and both correcting and not correcting for the fact that not all treated actually take up education.

Table 1 reports the results for a sample size of 2,000 individuals and 200 monte-carlo replications based on the assumption that the error terms, \( v \) and \( u \), are uncorrelated.

Rows 1 to 4 in table 1 display some measures of eligibility and educational attainment among different groups of individuals and policy scenarios. These numbers show that when agents are aware of the new subsidy they respond by significantly increasing their effort in period \( t = 0 \) in the intent of becoming eligible to subsidized education. The log-linear functional form for wages adopted in this example implies that this is specially true in region 1, where wages are higher. As a consequence, an additional 12% of the population invests in education if the subsidy is available (rows 3-4, column 1), amounting to almost 100% of the eligibles (row 4, column 2). However, if the subsidy is not announced in advance the change in educational attainment is much more modest as fewer individuals become eligible to subsidized education. As a consequence, only an additional 3% of the population changes educational attainment in response to an unexpected subsidy as most remain ineligible (rows 3-4, column 4).

Rows 5 to 8 show the true parameters. Education increases wages for the average individual but the selection process dictates that agents investing in education gain more from the investment than

\(^\text{10}\)Non-zero correlation between \( v \) and \( u \) implies that some selection on non-treatment outcomes is expected.
Table 1: Monte Carlo experiment - description assuming $u$ and $v$ are independent

<table>
<thead>
<tr>
<th></th>
<th>Expected policy change</th>
<th>Unexpected policy change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All population</td>
<td>Eligibles after</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
</tr>
<tr>
<td>Eligibility and education take up</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) % eligibles before policy</td>
<td>0.061</td>
<td>0.219</td>
</tr>
<tr>
<td>(2) % eligibles after policy</td>
<td>0.276</td>
<td>1.000</td>
</tr>
<tr>
<td>(3) % educated before policy</td>
<td>0.223</td>
<td>0.525</td>
</tr>
<tr>
<td>(4) % educated after policy</td>
<td>0.344</td>
<td>0.962</td>
</tr>
<tr>
<td>True parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5) ATE</td>
<td>0.354</td>
<td>0.502</td>
</tr>
<tr>
<td>(6) ATT</td>
<td>0.471</td>
<td>0.505</td>
</tr>
<tr>
<td>(7) LATE</td>
<td>0.492</td>
<td>0.492</td>
</tr>
<tr>
<td>(8) Aggregate effect</td>
<td>0.059</td>
<td>0.215</td>
</tr>
</tbody>
</table>

Notes: Simulated data based on 200 Monte-Carlo replications of 2000 observations each. Results refer to independent error terms $u$ and $v$. No time trends were included in the simulated data.

Estimates in columns 1-3 (4 and 5) are based on the assumption that the post policy generation is fully (not) aware of the availability of the subsidy and eligibility conditions when deciding about effort level in period $t = 0$. Columns 1 and 4 present results for the whole population; columns 2 and 5 present results for the population of individuals eligible for the subsidy; column 3 presents results for individuals living in region 1.

The figures in rows 1-4 show subsidy eligibility and education take up before and after the policy change. The figures in rows 5-8 are the true treatment effects on earnings among different populations depending on the group being considered in the respective column and on the parameter being estimated. Row 5 displays the impact of education on a randomly selected individual from the respective column population. Row 6 displays the impact of education on a randomly selected educated individual from the respective column population. Row 7 displays the impact of education on a randomly selected individual from the group of agents changing educational attainment in response to the policy among the respective column population. Row 8 displays the impact of the subsidy (not education) on a random individual selected from the respective column population.
<table>
<thead>
<tr>
<th></th>
<th>Expected policy change</th>
<th>Unexpected policy change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Comparison by</td>
<td></td>
</tr>
<tr>
<td></td>
<td>eligibility status</td>
<td>region</td>
</tr>
<tr>
<td>(1)</td>
<td>uncorrected estimates</td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>bias</td>
<td></td>
</tr>
<tr>
<td>(3)</td>
<td>corrected estimates</td>
<td></td>
</tr>
<tr>
<td>(4)</td>
<td>bias</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) uncorrected estimates</td>
<td>0.348</td>
<td>0.056</td>
<td>0.241</td>
</tr>
<tr>
<td>(2) bias</td>
<td>62.3%</td>
<td>6.5%</td>
<td>2.1%</td>
</tr>
<tr>
<td>(3) corrected estimates</td>
<td>0.595</td>
<td>0.531</td>
<td>0.479</td>
</tr>
<tr>
<td>(4) bias</td>
<td>21.0%</td>
<td>9.8%</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

**Notes:** Simulated data based on 200 Monte-Carlo replications of 2000 observations each. Results refer to independent error terms $u$ and $v$. No time trends were included in the simulated data.

Estimates in columns 1-2 (3) are based on the assumption that the post policy generation is fully (not) aware of the availability of the subsidy and eligibility conditions when deciding about effort level in period $t = 0$. Estimates in columns 1 and 3 explore the eligibility rule based on test scores; estimates in column 2 use regional variation in the timing of policy implementation.

Uncorrected DID estimates in row 1 are standard DID estimates. Corrected DID estimates in row 3 are re-scaled estimates to account for the fact that education take-up occurs even in the absence of the subsidy (see equation (23)). Row 2 displays the relative bias of uncorrected estimates as compared to the Aggregate Effect in row 8 of table ??; Row 4 displays the relative bias of corrected estimates as compared to LATE in row 7 of table ??.
the average agent (rows 5-6). Row 7 shows that agents moving into education in response to the policy change benefit more from the investment than agents investing in education in the absence of the subsidy. This is because the selection process in the absence of the subsidy is strongly affected by the cost of education, related to family background, and not so much by the gains from education, related to ability. By linking eligibility to performance, which itself is partly determined by ability, the subsidy strengthens the process of selection on ability but only mildly in our example. The Aggregate Effect in row 8 is the impact of introducing an education subsidy on average wages. It combines a null effect for agents unaffected by the policy (those not changing their educational decision in response to the subsidy) and the effect of education on agents that invest only if the subsidy is available. It is, therefore, much lower than any of the other parameters as it measures the impact of the subsidy, not the impact of high education.

Table 2 displays the DID estimates and respective bias. In producing these estimates we explore two sources of differential eligibility: region and test score. In the case of region (column 2), we assume the policy is first implemented in region 1 before being rolled out nationally. We compare outcomes in region 1 (the treated group) and region 0 (the control group) to assess the impact of treatment. In the case of test score (columns 1 and 3), we explore the eligibility rule in terms of test score by comparing agents scoring above the threshold (treatment group) with those scoring below the threshold (control group) over both regions.

Rows 1 and 2 in table 2 show results (estimated effects and bias) for the standard DID method. As argued before, this method will identify the Aggregate Effect under optimal conditions. It requires treated and controls to be correctly identified before and after treatment. Such requirement is fulfilled when the comparison uses regional variation in the implementation of the program, resulting in unbiased estimates (column 2). The same is true when the eligibility is the source of variation being explored for as long as the post-treatment generation being used is not aware of the policy change while making pre-treatment decisions (column 3). However, significant bias results from the comparison of eligibles and ineligibles within the context of an announced policy (column 1). In this case, reactions in anticipation to program participation will affect eligibility and change the composition of the treated and control groups over time, rendering them incomparable.
Rows 3 and 4 show similar results for the corrected DID method. Bias is now measured with respect to the LATE parameter and again, region and eligibility for an unexpected policy change can be used to identify the correct parameter (columns 2 and 3) but the identification conditions are not met by the use of eligibility within the context of an expected policy change (column 1).

All estimates are uninformative about the returns to education for and average individual or for the educated. Instead, they use the change in policy to identify the impact of education on a particular group of individuals: those at some margin of participating that respond to the extra incentive by becoming educated.

The results obtained under the alternative assumption of (negatively) correlated residuals resemble the ones presented here and are available under request. Since this additional source of selection does not affect our ability to identify treated and controls before the policy change, unbiased estimator will remain unbiased under the alternative setup.

5 Matching Methods

5.1 The matching estimator (M)

The main purpose of matching is to reproduce the treatment group among the non-treated, this way re-establishing the experimental conditions in a non-experimental setting. Under some assumptions we will discuss below, the matching method constructs the correct sample counterpart for the missing information on the treated outcomes had they not been treated by pairing each participant with members of the non-treated group. The matching assumptions ensure that the only remaining relevant difference between the two groups is program participation.

Matching can be used with cross-sectional or longitudinal data. In its standard formulation, however, the longitudinal dimension is not explored except perhaps on the construction of the matching variables. We therefore exclude the time subscript from this discussion but will consider the appropriate choice of the matching variables in what follows.

As a starting point we have to include some observable regressors in the outcome equation. We do this in a very general way. The covariates $X$ explain part of the residual term $u$ in (1) and part
of the idiosyncratic gains from treatment:

\[ y_i^1 = \beta + u(X_i) + \alpha(X_i) + [(u_i - u(X_i)) + (\alpha_i - \alpha(X_i))] \]

\[ y_i^0 = \beta + u(X_i) + (u_i - u(X_i)) \]

where \( u(X) \) is the predictable part of \( y_i^0 \), \( (u_i - u(X_i)) \) is what is left over of the disturbance \( u \) after conditioning for \( X \), \( \alpha(X) \) is some average treatment effect over individuals with observable characteristics \( X \) and \( \alpha_i \) is the individual \( i \) specific effect, which differs from \( \alpha(X_i) \) by the unobservable heterogeneity term.

The solution advanced by matching to estimate the ATT is based on the assumption that the set of observables, \( X \), contain all the information about the potential outcome in the absence of treatment, \( y^0 \), that was available to the individual at the point of deciding about whether to become treated, \( d \). This means that the econometrician has all the relevant information, namely the information that simultaneously characterize the participation rule and the non-treated outcome. This is called the Conditional Independence Assumption (CIA) and can be formally stated as follows

\[ y_i^0 \perp d_i \mid X_i \]  

Since all the information that simultaneously characterize \( y^0 \) and \( d \) is in \( X \), conditioning on \( X \) makes the non-treated outcomes independent from the participation status. Thus, treated and non-treated sharing the same observable characteristics, \( X \), draw the non-treated outcome, \( y_i^0 \), from the same distribution.

Within model (24), the CIA can be restated in terms of the unobservable in the non-treated outcome equation,

\[ (u_i - u(X_i)) \perp d_i \mid X_i \]

meaning that the unobservable component of the non-treated outcomes is independent of participation into treatment or, which is the same, that there is no selection on the unobservable term \( u_i \) in (24). This assumption obviously implies a conditional version of the randomization hypothesis (R1),

\[ E[u_i|d_i, X_i] = E[u_i|X_i] \]

This weaker version of the CIA is sufficient to estimate the ATT on individuals with observable characteristics \( X \) using matching. Again, nothing like the randomization hypothesis (R2) is required.
to identify the ATT, which means that selection on the unobservable gains can be accommodated by matching.

The implication of (25) or (26) is that treated and non-treated individuals are comparable in respect to the non-treated outcome, \( y^0 \), conditional on \( X \). Thus, for each treated observation \( (y^1) \) we can look for a non-treated (set of) observation(s) \( (y^0) \) with the same \( X \)-realization and be certain that such \( y^0 \) is a good predictor of the unobserved counterfactual.

Thus, matching is explicitly a process of re-building an experimental data set. Its ability to do so, however, depends on the availability of the counterfactual. That is, we need to ensure that each treated observation can be reproduced among the non-treated. This is only possible if the observables \( X \) do not predict participation exactly, leaving some room for unobserved factors to influence the treatment status. This is the second matching assumption, required to ensure that the region of \( X \) represented among participants is also represented among non-participants. Formally, it can be stated as follows

\[
P(d_i = 1 \mid X_i) < 1
\]

Given assumptions (26) and (27), we can now define the matching estimator. Let \( S \) represent the subspace of the distribution of \( X \) that is both represented among the treated and the control groups. \( S \) is known as the common support of \( X \). Under (27), \( S \) is the whole domain of \( X \). The ATT over the common support \( S \) is

\[
\alpha^{\text{ATT}}(S) = E \left[ y^1 - y^0 \mid d = 1, X \in S \right] = \frac{\int_S E(y^1 - y^0 \mid X, d = 1) \, dF(X \mid d = 1)}{\int_S dF(X \mid d = 1)}
\]

meaning that \( \alpha^{\text{ATT}}(S) \) is the mean of \( \alpha^{\text{ATT}}(X) \) (as defined in (7)) for \( X \in S \) weighted by the relative importance of \( X \) among the treated.

The matching estimator is the empirical counterpart of \( \alpha^{\text{ATT}}(S) \). It is obtained by averaging over \( S \) the difference in outcomes among treated and non-treated with equal \( X \)-characteristics using the empirical weights of the distribution of \( X \) among the treated. Formally, the matching estimator of the ATT is

\[
\hat{\alpha}^M = \sum_{i \in T} \left\{ y_i - \sum_{j \in C} \omega_{ij} y_j \right\} \omega_i
\]

36
where \( T \) and \( C \) represent the treatment and comparison groups respectively, \( \pi_{ij} \) is the weight placed on comparison observation \( j \) for the treated individual \( i \) and \( \omega_i \) accounts for the re-weighting that reconstructs the outcome distribution for the treated sample.

Identification of ATE requires a strengthened version of assumption (26) because the correct counterfactual needs to be constructed for both the treated and the non-treated. This means that both \( (u_i - u(X_i)) \) and \( (\alpha_i - \alpha(X_i)) \) need to be (mean) independent from \( d \) conditional on \( X \). That is, selection on unobserved expected gains must also be excluded for matching to identify the correct ATE. In its weaker version, the CIA is now formally:

\[
\begin{align*}
E[u_i | d_i, X_i] &= E[u_i | X_i] \\
E[\alpha_i | d_i, X_i] &= E[\alpha_i | X_i]
\end{align*}
\] (29)

Estimation of ATE also requires a modification of the overlapping support assumption (27) to ensure that both the treated and the non-treated are represented within the alternative group. Formally,

\[
0 < P(d_i = 1 | X_i) < 1
\] (30)

Under (29) and (30), the ATE over the common support \( S \) is

\[
\alpha^{ATE}(S) = E[y^1 - y^0 | X \in S] = \frac{\int_S E(y^1 - y^0 | X) \, dF(X)}{\int_S dF(X)}
\]

where now the conditional mean effects are weighted using the distribution of the \( X \)'s over the whole population \( (F(X)) \).

The choice of the appropriate matching variables, \( X \), is a delicate issue. Too much information and the overlapping support assumption will not hold. Too little and the CIA will not hold. The wrong sort of information and neither of the two assumptions will hold. So what is the right balance?

The appropriate matching variables are those describing the information available to the individual at the moment of deciding about participation and simultaneously explaining the outcome of interest. Only this set of variables ensures the CIA holds. However, the same is not necessarily
true for the overlapping support assumption as it will not hold when participation is determined with certainty within some regions of the support of $X$. In this case matching will identify a different parameter, namely the average impact over the region of common support. Typically, but not necessarily, individuals gaining the most and the least from treatment will be excluded from the analysis.

However, it is rarely clear what sort of information is in the individual information set at the moment of decision. There are a few rules that can help the econometrician in the selection of the matching variables:

- Matching variables should be determined before the time of decision and not after as this could compromise the CIA by having matching variables affected by the treatment status;

- Variables that are not simultaneously in the participation decision rule and outcome equations should not be used in matching as they could only contribute to reduce the common support region. A special case where this is true is that of variables that could be used to instrument the treatment status in the outcome equation;

- A structural model will generally shed some light on what the correct set of matching variables should be. Although such model is likely to suggest some unobservable variables, it can also show possible alternatives to make up for the unavailable information under specific conditions. For example, in studies about the impact of training on labor market outcomes, previous labor market history could contain all the relevant information in the unobservables ability and job-readiness as it is partly determined by such factors.

### 5.2 Propensity score matching

A serious limitation to the implementation of matching is the dimensionality of the space of the matching variables, $X$. Even if all variables are discrete with a finite domain, the dimensionality of the combined space increases exponentially with the number of variables in $X$, making it virtually impossible to find a match for each observation within a finite (even if large) sample when more than a few variables are being controlled for.
A popular alternative is to match on a function of $X$. Usually, this is carried out on the probability of participation given the set of characteristics $X$. Let $P(X)$ be such probability, known as the “propensity score”. It is defined as

$$P(X) = P(d = 1 | X).$$

Its has been motivated by Rosenbaum and Rubin’s result on the balancing property of the propensity score (1983, 1984). It is shown that if the CIA is valid for $X$ it is also valid for $P(X)$.

$$y_i^0 \perp d_i | X_i \Rightarrow y_i^0 \perp d_i | P(X_i)$$

This means that if the probability of participation, $P(X)$, is known it can be used to replace $X$ in the matching procedure.\(^{11}\)

The propensity score allows matching to be performed on a single dimension rather then on the number of variables in $X$, thus simplifying the matching procedure significantly. However, in concrete applications $P(X)$ is not known and needs to be estimated. Whether the overall estimation process is indeed simplified and the computing time reduced depends on what is assumed about $P(X)$. The popular procedure amounts to employing a parametric specification for $P(X)$, usually in the form of a logit, probit or linear probability model. This solves the dimensionality problem but relies on parametric assumptions. Alternatively, a non-parametric propensity score keeps the full flexibility of the matching approach but does not solve the dimensionality problem.

When using propensity score matching, the comparison group for each treated individual is chosen with a pre-defined criteria (established in terms of a pre-defined metric) of proximity between the propensity scores for the each treated and the controls. Having defined the neighborhood for each treated observation, the next step is that of choosing the appropriate weights to associate the selected set of non-treated observations for each participant. Several possibilities are commonly used. We briefly refer the most commonly applied alternatives and refer the interested reader to ??? (we need a good reference here) for a comprehensive discussion of alternative matching procedures.

\(^{11}\)More recently, a study by Hahn (1998) shows that $P(X)$ is ancillary for the estimation of ATE. However, it is also shown that knowledge of $P(X)$ may improve the efficiency of the estimates of ATT, its value lying on the “dimension reduction” feature.

39
The *Nearest Neighbor Matching* assigns a weight 1 to the closest non-treated observation and 0 to all others. A widespread alternative is to use a certain number of the closest non-treated observations to match the treated, generally the 10 closest observations. This reduces the variability of the nearest neighbor estimator and is more reliable specially when the sample of treated individuals is small as each match may significantly affect the results.

*Kernel Matching* defines a neighborhood for each treated observation and constructs the counterfactual using all control observations within the neighborhood, not only the closest one. It assigns a positive weight to all observations within the neighborhood while the weight is zero otherwise. Different weighting schemes define different estimators. For example, uniform kernel attributes the same weight to each observation in the neighborhood while other forms of kernel make the weights dependent on the distance between the treated and the control being matched, where the weighting function is decreasing in distance. By using more observations per treated, Kernel matching reduces the variability of the estimator as compared to the Nearest Neighbor and produces less bias then Nearest Neighbor with many matches per treated. However it still introduces significant bias at the edges of the distribution of $P(X)$. When this is a problem, the *Local Linear Matching* will effectively deal with this sort of bias.\(^\text{12}\)

Not only kernel and local linear matching produce more precise estimates than nearest neighbor matching, it is also simpler to compute the precision for these estimators. The complexity of propensity score matching demands for bootstrapping to be used in computing the standard errors for the effect of treatment. The problem with the nearest neighbor technique is that bootstrapping is not guaranteed to deliver consistent estimates since choosing only 1 (or a fixed number of) match(es) per treated means that the quality of the match does not necessarily improve as the sample (of controls) gets bigger. The same is not true for kernel and local linear matching as with these estimator the sample of matched controls expands with the sample size (for a thoroughly discussion of bootstrapping see Horowitz, 2001).

The general form of the matching estimator is not altered by the sort of weights one decides to apply. As before, it is given by $\hat{\alpha}^M$ in (28).

\(^{12}\)For a discussion of non-parametric estimators including Kernel and Local Linear Regression methods see ???. - another good reference needed here.
While propensity score matching is affected by the same problems as fully non-parametric matching in what concerns to choosing the right set of controlling variables, it also faces the additional problem of finding a sufficiently flexible specification for the propensity score to ensure that the distribution of observables is indeed the same among treated and matched controls. That is, one wants to ensure that ensuring that (26) holds implies that (25) also holds. The evaluation literature has proposed a few balancing tests to assess whether the specification for the propensity score is statistically sound. For example, Rosenbaum and Rubin (1985) propose a test based on the comparison of means for each covariate between treated and matched controls. If the difference in means is too large, the test rejects the hypothesis that the samples (of treated and matched controls) are balanced with respect to the covariates when they are balanced with respect to the propensity score.

5.2.1 The popular alternative to matching: the linear regression model

The linear regression model is probably the most used evaluation model. It is often seen as the main alternative to matching as estimation also relies on the exclusion of selection on the unpredictable component of the outcome. It amounts to impose a fully parametric structure to model (24) by assuming that \( u_i \) and \( \alpha_i \) are linear functions of \( X \):

\[
\begin{align*}
    u(X_i) &= X_i \eta \\
    \alpha(X_i) &= \xi_0 + X_i \xi_1
\end{align*}
\]

where \((\eta, \xi_0, \xi_1)\) are the unknown coefficients. The model can then be written as

\[
\begin{align*}
    y_0^i &= \beta^0 + X_i \gamma^0 + e_0^i \\
    y_1^i &= \beta^1 + X_i \gamma^1 + e_1^i
\end{align*}
\]

where

\[
\begin{align*}
    \beta_d &= \beta + d\xi_0 \\
    \gamma_d &= \eta + d\xi_1 \\
    e_d^i &= (u_i - X_i \eta) + d(\alpha_i - \xi_0 - X_i \xi_1)
\end{align*}
\]
and $d$ is the treatment indicator.

Estimation of the ATT requires knowledge of the model for the untreated outcomes, (31). Under the CIA and the assumption of exogeneity of the covariates $X - E(y^0 | X) = E(y^0)$ - it can be estimated using OLS. The common support assumption (27) is not required as the parametric specification can be used to extrapolate $y^0$ outside the observable range for $X$ when predicting the counterfactual for each treated observation.

The imposition of a linear specification is not as restrictive as it might first seem. In fact, by including all sorts of interactions between the variables and higher order polynomials in the (continuous) regressors, one will closely approximate any smooth function $y^0$. The main requirement is then to use a flexible enough functional form for $y^0$.

More restrictive is the relaxation of the common support assumption. In its absence, the model needs to be extrapolated over unobservable regions of the distribution of $X$, where only the true model can be guaranteed to perform well. Of course, one could always think of imposing the common support assumption within the parametric linear model and estimate the average effect of treatment within regions of $X$ simultaneously observed among treated and controls. However, while this is feasible it is rarely done in the context of parametric models given the simplicity of extrapolating to outside the observable interval. Most frequently, the researchers are unaware that a common support problem exists.

Another drawback of the parametric linear model is the requirement of exogeneity of $X$ in the equation for $y^0$. Again, this is most important if the function is to be extrapolated to outside the estimation range. The purpose of estimating the equation for $y^0$ is to predict the unobservable counterfactual for the treated. Whether or not estimation is consistent is of less importance, what matters is that the predictions are accurate. Ensuring that the right counterfactual is being predicted is more difficult outside the estimation domain and will surely not be possible without a consistent estimator of the non-treated outcome.
5.3 Weaknesses of matching

The main weaknesses of matching are data driven: its availability and our ability to select the right information. The common support assumption (27) ensures that the missing counterfactual can be constructed from the population of non-treated. What (27) does not ensure is that the same counterfactual exists in the sample. If some of the treated observations cannot be matched, the definition of the estimated parameter becomes unclear. It is the average impact over some subgroup of the treated, but such subgroup may be difficult to define. The relevance of such parameter depends, of course, on the ability to define the population it corresponds to.

Taken together, assumptions (25) - or (26) - and (27) show how demanding matching is with data: the right regressors $X$ must be observed to ensure that what is left unexplained from $y^0$ is unrelated with the participation decision; any more than the right regressors will only contribute to make finding the correct counterfactual harder or even impossible. In particular, variables in the decision rule (in $Z$) but not in $X$ should be excluded from the matching procedure as they only interfere with our ability to ensure (27). To achieve the appropriate balance between the quantity of information at use and the share of the support covered can be very difficult. In a recent paper, Heckman and Lozano (2004) show how important and, at the same time, how difficult it is to choose the appropriate set of variables for matching. Bias results if the conditioning set of variables is not the right and complete one. In particular, if the relevant information is not all controlled for, adding additional relevant information but not all that is required may increase, rather then reduce, bias. Thus, aiming at the best set of variables within the information available may not be a good policy to improve the matching results.

If, however, the right amount of information is used, matching deals well with potential bias. This is made clear by the following decomposition of the treatment effect

$$E(y^1 - y^0 \mid X, d = 1) = \{E(y^1 \mid X, d = 1) - E(y^0 \mid X, d = 0)\} - \{E(y^0 \mid X, d = 1) - E(y^0 \mid X, d = 0)\}$$

where the second term on the rhs is the bias conditional on $X$. Conditional on $X$, the only reason the true parameter, $\alpha_{ATT}(X)$, might not be identified is selection on the unobservable term $u$. How-
ever, integration over the common support $S$ creates two additional sources of bias: non-overlapping support of $X$ and misweighting over the common support. Through the process of choosing and re-weighting observations, matching corrects for the latter two sources of bias and selection on the unobservables is assumed to be zero by the CIA.

5.4 Using matching to estimate the returns to education

In this section we again assume that earning levels change with region and adopt the earnings specification (4.4) which we reproduce here:

$$\ln y_i = \beta_0 + \beta_1 x_i + (\alpha_1 + \alpha_2 \theta_i) d_i + u_i$$

where $x$ is region and can assume 2 values, 0 or 1. The impact of education on earnings is now region-specific given the non-linear form of the earnings equation. In what follows we exclude sorting by region. Thus, the distribution of ability does not change with region. Thus the ATE on log earnings will not depend on region but the same does not hold with respect to the the ATT due to the selection process.

5.4.1 Monte-Carlo results

We ran some monte-carlo experiments under different assumptions about the relationship between $d$ and $u$, the source of endogeneity in evaluation problems. We estimated both ATT and ATNT using different sets of conditioning variables. Table ?? details the results obtained using log earnings when and educational subsidy is available.\textsuperscript{13}

ATT estimates are presented in Panel A of table 3. Columns (1)-(3) display the results for uncorrelated unobservables in the cost of education and outcomes. Columns (4)-(6) display the results for negatively correlated unobservables in the cost of education and outcomes where the correlation is -0.5. In each case, we present the true effect together with the matching estimate and the bias measured as the relative difference of the estimate to the true effect.

\textsuperscript{13} Estimates in levels and under the no-subsidy scenario show the similar patterns to the ones presented here and are available under request.
<table>
<thead>
<tr>
<th>Panel A: Estimates of the ATT</th>
<th>corr(u, v) = 0</th>
<th>corr(u, v) &lt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>true effect</td>
<td>estimate</td>
<td>bias</td>
</tr>
<tr>
<td>(1) Simple difference</td>
<td>0.474</td>
<td>0.494</td>
</tr>
<tr>
<td>Matching using conditioning variables:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) x (region)</td>
<td>0.474</td>
<td>0.469</td>
</tr>
<tr>
<td>(3) b (family background)</td>
<td>0.474</td>
<td>0.494</td>
</tr>
<tr>
<td>(4) s (test score)</td>
<td>0.444</td>
<td>0.518</td>
</tr>
<tr>
<td>(5) θ (ability)</td>
<td>0.473</td>
<td>0.534</td>
</tr>
<tr>
<td>(6) (x, b, s, θ)</td>
<td>0.455</td>
<td>0.487</td>
</tr>
</tbody>
</table>

| Panel B: Estimates of the ATNT |
|-----------------------------|----------------|
| true effect | estimate | bias |
| (7) Simple difference | 0.296 | 0.494 | 0.672 | 0.315 | 0.991 | 2.144 |
| Matching estimates using conditioning variables | | | | | | |
| (8) x (region) | 0.296 | 0.470 | 0.590 | 0.315 | 0.970 | 2.080 |
| (9) b (family background) | 0.295 | 0.585 | 0.984 | 0.315 | 1.163 | 2.691 |
| (10) s (test score) | 0.294 | 0.332 | 0.131 | 0.315 | 1.153 | 2.656 |
| (11) θ (ability) | 0.296 | 0.317 | 0.068 | 0.318 | 1.027 | 2.231 |
| (12) (x, θ) | 0.296 | 0.290 | 0.020 | 0.317 | 1.004 | 2.168 |
| (13) (x, b, s, θ) | 0.296 | 0.254 | 0.142 | 0.319 | 1.263 | 2.964 |

Notes: Simulated data based on 200 Monte-Carlo replications of 2000 observations each. All estimates obtained under the assumption that the true specification of the outcomes equation is additively separable in logs and that subsidised education places are available for those with high test scores. Columns (1) to (3) present results obtained for independent error terms, u and v. Columns (4) to (6) present results obtained for (negatively) correlated error terms, u and v, with a correlation of -0.5. Bias estimates result from the comparison of the average estimate with the true effect in column (1) and are measured in relative terms. ATT stands for “average treatment on the treated”. ATNT stands for “average treatment on the non-treated”. Estimates in rows (2)-(6) and (8) to (12) are based on propensity score matching using Epanechnikov kernel weights. Different matching variables are used in each row. Estimates in rows (1) and (7) are based on simple differences.
Notice that true effects in columns (1) and (4) change with the set of conditioning variables due to changes in the overlapping support. In our example, this is never a serious problem mainly because the state space being considered is small. Nevertheless, the more conditioning variables are included to perform matching, the more the identifiable effect differs from the population one displayed in row (1).

We start by considering the case of independent error terms presented in columns (1) to (3). Rows (2)-(6) display the matching estimates under different sets of conditioning variables while row (1) displays the simple difference estimates. In this example, the correct estimator of ATT uses region alone as this is the only regressor that simultaneously affect the educational decision and the outcome in the non-educated status. The results in row 2, columns 1-3 show that matching identifies the ATT in this case.

However, matching on other characteristics will induce some bias. For instance, suppose we decide to match on the test score - row (4). Individuals with the same score may decide differently about education because they expect different gains from the investment. In our case, the gains from treatment are higher in the region where the wages for both educated and non-educated individuals are also higher. Thus, individuals in the high-returns region are more likely to participate, but they would also enjoy from higher payments if they had remained uneducated then individuals in the low-returns region. This means that the regional distribution of individuals with the same but different educational attainment is different, and thus the expected outcome in the no education scenario is also different. That is, a comparison conditional on test scores only is not valid.

Rows (7)-(13) show how much more difficult it can be estimating ATNT than ATT. The conditional independence assumption underlying the matching estimation of the ATNT requires the treated outcome \( y^1 \) to be independent of the treatment status, \( d \) conditional on the covariates \( Z \). This excludes selection on the non-treated outcome, \( y^0 \), and on the gains from treatment. \( y^0 \) depends on region \( x \) and the gains from treatment depend on ability \( \theta \) and region \( x \). Thus, the correct conditioning set is now \( (\theta, x) \) and the first 3 columns of row 12 show the results that confirm this.

However, ability is rarely available in empirical studies. Unfortunately, rows 8-10 and 13 show that matching on alternative (sets of) covariates creates sizeable bias. The problem here, as compared
to the identification of ATT, is that selection on the expected gains is quite strong and ability is the main determinant of such gains. The test score can be quite valuable in the case of unobservable ability as it is strongly affected by ability. In fact, row (10) shows a much lower bias than other rows using alternative matching variables.

Results for correlated error terms are displayed in columns 4-6 of the table. Large bias is displayed in every case as this is now a model of selection on the unobservables since \( v \) has information on the future level of productivity \( u \). The results in the table are for a correlation of -0.5. Figure 2 displays the bias in the estimation of the ATT and ATNT for different levels of correlation when the correct observable matching set is controlled for. The graph shows quite considerable bias even for relatively low levels of correlation, particularly for the ATNT but also for the ATT. When selection on the unobservables is suspected, other methods such as IV and control function are more adequate than matching. These will be discussed in what follows.

5.5 Combining matching and DID (MDID)

In the presence of longitudinal or repeated cross-section data, matching and DID can be combined to weaken the underlying assumptions of both methods. The CIA is quite strong if individuals are allowed to decide according to their outcome’s forecast as data is rarely rich enough to describe the relevant available information. However, the combination of matching with DID can accommodate unobserved determinants of the non-treated outcome affecting participation for as long as these are constant over time.

To discuss MDID, we start by decomposing the unobservable term \( u_i \) in (24) into a fixed effect (\( \phi \)), macro shock (\( \psi \)) and an idiosyncratic transitory shock (\( \epsilon \)). MDID can be applied when treated and non-treated are observed over time with at least one observation before and one after the treatment. For simplicity we consider two time periods, \((t_0, t_1)\), where \( t_0 < k < t_1 \) and \( k \) is the time of treatment. MDID compares the evolution of treated outcomes with that of non-treated over the observation period \((t_0, t_1)\) and assigns any difference to the impact of treatment. To do so, MDID assumes that had the treated remained non-treated and they would have experienced a change in outcomes equal to that observed among the actual non-treated.
More formally, the model can now be written as

\[ y_{it} = \beta + u(X_i) + (\phi_i + \psi_t + \epsilon_{it} - u(X_i)) + (\alpha_i - \alpha(X_i)) \]
\[ y_{i0} = \beta + u(X_i) + (\phi_i + \psi_t + \epsilon_{it} - u(X_i)) \]

where \( y_{it}^d \) is the outcome for individual \( i \) at time \( t \) when his/her treatment status at that time is \( d \) - it is \( y^0 \) when the individual belongs to the non-treated group or when the time is \( t_0 \), and is \( y^1 \) when the individual is in the treated group and the time is \( t_1 \). The MDID assumption states that, conditional on the observables \( X \), the evolution of the unobserved part of \( y^0 \) is independent of the treatment status. Thus,

\[(u_{it1} - u_{it0}) \perp d_{it1} \mid X_i \quad (34)\]

The main matching hypothesis is now stated in terms of the before-after evolution instead of levels. It means that controls evolve from a pre- to a post-program period in the same way treatments would have evolved had they not been treated.

Assumption (34) is not enough to ensure identifiability of ATT. Just as in the matching case, we also need to impose a common support hypothesis. This will be the same as (27) when longitudinal data is available. If we only dispose of repeated cross-section data, however, we will need to strengthen it to ensure that the treated group can be reproduced in all three control groups characterized by treatment status before and after the program. This version of the common support assumption states that all treated individuals have a counterpart on the non-treated population before and after the treatment,

\[ P(d_{it1} = 1 \mid X_i, t) < 1 \quad (35)\]

where \( P(d_{it1} = 1 \mid X_i, t) \) is the probability that an individual observed at time \( t \) with characteristics \( X_i \) would belong to the treatment group at time \( t_1 \).

The effect of the treatment on the treated can now be estimated over the common support of \( X \), call it \( S \). The following estimator is adequate to the use of propensity score matching with longitudinal data

\[ \tilde{\alpha}^{MDID,L} = \sum_{i \in T} \left\{ \frac{[y_{it1} - y_{it0}] - \sum_{j \in C} \omega_{ij} [y_{jt1} - y_{jt0}]}{\omega_i} \right\} \]
where the notation is similar to what has been used before. With repeated cross-section data, however, matching must be performed over the three control groups: treated and non-treated at \( t_0 \) and non-treated at \( t_1 \). In this case, the matching-DID estimator would be

\[
\tilde{\alpha}_{MDID, RCS} = \frac{1}{\sum \omega_i} \left\{ \left[ y_{it_1} - \sum_{j \in T_0} \omega_{ijt_0}^T y_{jt_0} \right] - \left[ \sum_{j \in C_1} \omega_{ijt_1}^C y_{jt_1} - \sum_{j \in C_0} \omega_{ijt_0}^C y_{jt_0} \right] \right\} \omega_i
\]

where \( T_0, T_1, C_0 \) and \( C_1 \) stand for the treatment and comparison groups before and after the program, respectively, and \( \omega_{ijt_0}^G \) represents the weight attributed to individual \( j \) in group \( G \in \{C, T\} \) and time \( t \in \{t_0, t_1\} \) when comparing with treated individual \( i \).

The implementation of the MDID estimator using propensity score matching requires the propensity score to be estimated using treated and controls. In the presence of longitudinal data, one sets the dependent variable \( d \) equal to 1 if the agent is treated and to 0 otherwise. The controls are then matched to the treated and the re-weighted sample is used to compute the ATT using DID. In the presence of repeated cross-section data, the dependent variable is set to 1 if the agent is treated and the period of observation is \( t_1 \) and is set to 0 otherwise. Each of the control groups (treated before treatment and non-treated before and after treatment) are then matched to the treated after treatment separately. The overlapping region of support is now composed of the treated to whom a counterfactual is found in each of the three control samples. The three sets of weights can then be used to estimate the ATT using DID.

6 Instrumental Variables

6.1 The instrumental variables (IV) estimator

In contrast to the matching method, Instrumental Variables deals directly with selection on the unobservables. For this exposition, therefore, all the following results are conditional on \( X \) and in many cases we will omit \( X \) for ease of notation. We also omit the index \( t \) since longitudinal or

\footnote{As with the DID estimator, our ability to correctly separate treated from non-treated at \( t_0 \) is determinant for the quality of the estimates.}
repeated cross-section data is not necessarily required to estimate the effect of treatment under the IV assumptions.

Consider the model (1)-(2) under the selection process (5)-(4). IV requires the existence of at least one regressor exclusive to the decision rule, that is, a variable \( w \) in \( Z \) but not in \( X \). As \( w \) changes, so does the probability of participation such that individuals with different realizations of \( w \) participate at different rates. However, such change in \( w \) has no impact on the potential outcomes as \( w \) is not in \( X \). Thus the difference in the mean outcomes of two groups differing only with respect to \( w \) is due exclusively to the consequent difference in the participation rates. When the treatment effect is homogeneous, so that \( \alpha_{ATE} = \alpha_{ATT} = \alpha_i = \alpha \), it can be identified from the such differences in mean outcomes and participation rates.

To see this more clearly, we formally establish the three assumptions enunciated above. The first two assumptions state that the instrument \( w \) explains the participation decision but has no impact on the potential outcomes. These can be stated as, respectively:

\[
P[d = 1|Z_{-w}, w] \neq P[d = 1|Z_{-w}] \tag{36}
\]

\[
E[u|w] = E[u] \tag{37}
\]

where \( Z_{-w} \) is the set of regressors in the selection rule excluding the instrument \( w \). The third assumption states that the treatment effect is homogeneous:

\[
\alpha_i = \alpha \text{ for all } i \tag{38}
\]

Under these conditions, the instrumental variable \( w \) is the source of exogenous variation used to approximate randomized trials. It provides some variation that is correlated with the participation decision but does not affect the potential outcomes from treatment directly.

Under assumptions (37) and (38) we can write

\[
E(y_i \mid w_i) = \alpha P(d_i = 1 \mid w_i) + E(u_i \mid w_i)
= \alpha P(d_i = 1 \mid w_i) + E(u_i)
\]

which when used with two different values for \( w \), say \( w^* \) and \( w^{**} \), yields

\[
E(y_i \mid w_i = w^*) - E(y_i \mid w_i = w^{**}) = \alpha [P(d_i = 1 \mid w_i = w^*) - P(d_i = 1 \mid w_i = w^{**})]
\]
thus identifying the treatment effect from the ratio

$$\alpha = \frac{E(y_i \mid w_i = w^*) - E(y_i \mid w_i = w^{**})}{P(d_i = 1 \mid w_i = w^*) - P(d_i = 1 \mid w_i = w^{**})}$$

(39)
as long as $P(d_i = 1 \mid w_i = w^*) \neq P(d_i = 1 \mid w_i = w^{**})$ (which is the substance of assumption (36)). This is the standard IV identification strategy. The IV estimator is typically obtained from

$$\hat{\alpha}_{IV} = \frac{\text{cov}(y, w)}{\text{cov}(d, w)}.$$

### 6.2 Weaknesses of IV

A key issue in the implementation of IV is the choice of the instrument. Very frequently, it is impossible to find a variable that satisfies assumption (37), in which case IV is of no practical use. In other cases, the instrument $w$ has insufficient variation, which means that the estimation must rely on two very close values of $w$. In such case, the denominator in (39) can be very small, leading to very imprecise estimates of the treatment effect.

Even if a proper instrument is available, the identification of the true ATE (or ATT) relies on the homogeneity assumption (38). If (38) does not hold, (37) is unlikely to hold as well since the unobservable component is now (see equation (11))

$$e_i = u_i + d_i (\alpha_i - \overline{\alpha})$$

which, under (36), implies some dependence between $e$ and $w$. The one exception occurs when there is no selection on the idiosyncratic gains, $\alpha_i - \overline{\alpha}_{ATE}$, which then implies that $e$ is mean independent of $w$,

$$E[e \mid w] = E[u \mid w] + P[d = 1 \mid w] E(\alpha_i - \overline{\alpha}_{ATE} \mid d = 1, w)$$

$$= E[u]$$

since $E(\alpha_i - \overline{\alpha}_{ATE} \mid d = 1, w) = 0$. In this case, IV still identifies ATE which is not different from ATT as individuals do not use information on their idiosyncratic gains to decide about participation.

However, in the more general case of heterogeneous effects with selection on idiosyncratic gains, IV will not identify ATE or ATT. If individuals are aware of their own idiosyncratic gains from
treatment, they are expected to make a more informed participation decision. The resulting selection process generates some correlation between $\alpha$ and $w$. To see why notice that, in this case, both $\alpha$ and $w$ affect the participation process. As a consequence, the distribution of $w$ among participants with high gains will be different from the distribution of $w$ among participants with low gains. For example, it may be that participants with values of $w$ that make participation more unlikely gain on average more from treatment than participants with values of $w$ that make participation more likely to occur. This results in some correlation between $\alpha$ and $w$ caused by the selection process.

Consider the education example discussed earlier. Suppose we use family background to instrument the level of education relying on the assumption that family background is uncorrelated with ability. In such case, family background will be uncorrelated with potential earnings under the two treatment scenarios, $y^0$ and $y^1$. However, in the data family background will be related with the idiosyncratic component of the returns to education, determined by ability, since individuals with a “good family background” (facing relatively low educational costs) are more likely to invest in education than individuals with “low family background” (facing high educational costs), and do so even if expecting relatively low returns.

Thus, if the homogeneity assumption (38) fails to hold, IV will not generally identify ATE or ATT. This happens because the average outcomes of any two groups differing only on the particular $w$-realizations are different for two reasons: (i) different participation rates and (ii) compositional differences in the treated/non-treated groups according to the unobservables. The later precludes identification of ATE or ATT. However, a different “local” average parameter can be identified under slightly modified hypothesis - the LATE to which we now turn.

### 6.3 The LATE parameter

The solution advanced by Imbens and Angrist (1994) is to identify the impact of treatment from local changes in the instrument $w$ when the effect is heterogeneous. The rationale is that, under certain conditions, a change in $w$ reproduces random assignment locally by inducing individuals to alter their participation status without affecting the potential outcomes, $(y^0, y^1)$. As with standard IV, the difference in average outcomes between two groups differing only in the realization of $w$ results
exclusively from the consequent difference in participation rates. Unlike standard IV, the identifiable effect will not correspond to the ATE or the ATT. Instead, it will depend on the particular values of \( w \) used to make the comparison and the identifiable effect is the average impact on individuals that change their participation status when faced with the change in \( w \) used to estimate the effect of treatment.

To discuss this parameter and the identification strategy more formally we denote by \( y_i(w) \) the outcome of individual \( i \) at a given point of the instrument \( w \). This outcome depends on \( w \) because the treatment status may change as \( w \) changes. By the standard IV assumptions, this is the only source of dependence between \( y \) and \( w \). Using this notation we can rewrite equation (2) by taking the instrument explicitly into account

\[
y_i(w) = d_i(w)y_i^1 + (1 - d_i(w))y_i^0
\]

where \( d_i(w) \) is the random variable representing the treatment status of individual \( i \) at a given point \( w \).

The LATE identification strategy requires stronger assumptions than the standard IV one to compensate for the relaxation of the homogeneity hypothesis. We start by assuming that the potential outcomes and the treatment effect as a function of \( w \) are independent of the instrument \( w \):

\[
\left( y_i^1, y_i^0, d_i(w) \right) \perp w_i \quad (40)
\]

This means that \( w \) is unrelated with the unobservable in the selection equation (5), \( v \), the unobservable in the outcome equation (1), \( u \), and the idiosyncratic gain from treatment, \( \alpha_i \).

Assumption (40) ensures that \( y_i(w) \) is independent of \( w \) and thus

\[
E[y_i(w) | w] = E[y_i(w)] \quad (41)
\]

since \( w \) contains no information about the potential outcomes conditional on \( d(w) \) and only affects \( d(w) \) through its observable part.

We now use this result to compare the observed outcomes at two distinct values of the instrument
$w$, say $w^{\ast}$ and $w^{**}$:

$$E[y_i(w)|w = w^{**}] - E[y_i(w)|w = w^{\ast}] = E[y_i(w^{**}) - y_i(w^{\ast})]$$

$$= E [(d_i(w^{**}) - d_i(w^{\ast})) (y^1_i - y^0_i)]$$

$$= P[d_i(w^{**}) > d_i(w^{\ast})] E [y^1_i - y^0_i|d_i(w^{**}) > d_i(w^{\ast})]$$

$$- P[d_i(w^{**}) < d_i(w^{\ast})] E [y^1_i - y^0_i|d_i(w^{**}) < d_i(w^{\ast})]$$

where the second equality is obtained by substituting in the expression for $y(w)$ and the third equality uses the fact that whenever $d_i(w^{\ast}) = d_i(w^{**})$ the expression in the expectations operator is nil.

The above equality means that under the independence assumption in (40), any change in the average outcome $y$ when $w$ changes is solely due to changes in the treatment status of a subset of the population. The last equality shows two treatment parameters that one may be willing to identify: the impact of treatment on the treated under $w^{**}$ but not treated under $w^{\ast}$ and the impact of treatment on the treated under $w^{\ast}$ but not treated under $w^{**}$. It is frequently true that one of the alternatives can be eliminated. For example, it may be the case that everyone that participates at $w = w^{\ast}$ also participates at $w = w^{**}$ but the reverse is not true. This is the monotonicity assumption, which states that the decision rule, $d(w)$ is a non-trivial monotonic function of $w$.

Suppose the monotonicity assumption holds. Let $d(w)$ be increasing in $w$ and $w^{**} > w^{\ast}$ so that $P[d_i(w^{**}) < d_i(w^{\ast})] = 0$. In such case

$$E[y_i(w)|w = w^{**}] - E[y_i(w)|w = w^{\ast}] = P[d_i(w^{**}) > d_i(w^{\ast})] E [y^1_i - y^0_i|d_i(w^{**}) > d_i(w^{\ast})].$$

and this equation rearranged yields the LATE parameter:

$$\alpha^{LATE} (w^{\ast}, w^{**}) = \frac{E [y_i(w)|w = w^{**}] - E[y_i(w)|w = w^{\ast}]}{P[d_i(w^{**}) > d_i(w^{\ast})]}$$

$$= \frac{E [y_i(w)|w = w^{**}] - E [y_i(w)|w = w^{\ast}]}{P[d_i = 1|w^{**}] - P[d_i = 1|w^{\ast}]}. $$

LATE measures the impact of treatment on individuals that move from non-treated to treated when $w$ changes from $w^{\ast}$ to $w^{**}$.
To illustrate the LATE approach, return to the education example and suppose $w$ is family background. Participation is assumed to become more likely as $w$ increases. To estimate the effect of education, consider a group of individuals that differ only in the family background dimension. Among those that enroll into education when the family background $w$ equals $w^{**}$ some would not do so if $w = w^*$ where $w^* < w^{**}$. LATE measures the impact of education on the “movers” by assigning any difference on the average outcomes of the two groups to the different enrollment rates caused by the difference in family background.\footnote{Abadie, Angrist and Imbens (1998) extend this approach to the evaluation of quantile treatment effects. The goal is to assess how different parts of the outcome’s distribution are affected by the policy. As with LATE, a local IV procedure is used, making the estimated impacts representative only for the sub-population of individuals changing their treatment status in response to the particular change in the instrument being considered.}

6.3.1 The LATE assumptions

The independence assumption in (40) is required to establish the result in (41). It states that $w$ is independent of the unobservable components of the outcomes and participation rules, namely $e_i^0 = u_i$ and $e_i^1 = u_i + d_i(\alpha_i - \pi)$ in equation (11) and $v_i$ in equation (5). This means that $w$ should not affect the observed outcome through any effect on the potential outcomes or any relation with the unobserved components of the model. While the former is easy to understand the later requires some explanation. Suppose $w$ is related with $v$ in the participation rule and $v$ is related with $u$ in the outcome equation. Then the potential outcome will generally be related with $w$.

To illustrate the conditions under which the independence assumption may not apply, consider our earlier education example. Take $w$ to be some measure of family background and assume it is related with the taste for education, included in $v$, and the taste for education then affects the taste for working, included in $u$. When comparing two groups with different realizations of $w$ one is implicitly comparing two populations with a different distribution of $v$ and, therefore, a different distribution of $u$. That is, if $w$ is not exogenous in the participation equation then the comparison of two groups with different realizations of $w$ is not valid since the populations being compared are different with respect to other variables affecting the potential outcomes.

The monotonicity assumption is required for interpretation purposes. Under monotonicity of $d$
with respect to \( w \), the LATE parameter measures the impact of treatment on individuals that move from non-treated to treated as \( w \) changes. If monotonicity does not hold, LATE measures the change in average outcome caused by a change in the instrument, which is due to individuals moving in and out of participation. However, it is not possible to separate the effect of treatment on individuals that move in from that on individuals that move out as a consequence of a change in \( w \) (see Heckman, 1997).

Notice that the LATE assumptions are local: they only need to hold locally, for the specific values of \( w \) used in the estimation process. As a consequence, the identified parameter is a local parameter, only respecting to the specific population defined by the instrument. This is further discussed in the next point.

6.3.2 What does LATE measure?

Although very similar to the IV estimator presented in (39), LATE is intrinsically different since it does not represent ATT or ATE. LATE depends on the particular values of \( w \) used to evaluate the treatment and on the particular instrument chosen. The group of “movers” is not in general representative of the whole treated or, even less, the whole population. Whether the parameter is of policy interest or not depends on the instrument and the specific values of the instrument used in the estimation (see, for example, the discussion in Heckman, Lalonde and Smith, 1999). When a discrete variable, namely a change in policy, is used to instrument participation, LATE will measure the effect of treatment on individuals changing their treatment status in response to the policy change. In this case, LATE focus on an important subpopulation and may provide an important measure of the impact of the policy. If, on the other hand, a continuous variable measuring some individual characteristic is used to instrument participation, LATE will generally be much less informative.

In our education example, notice that we discussed two alternative instruments to measure a local ATE. In the first case, in the context of DID, we used a change in policy to measure the impact of education on individuals moving into education. DID differs from the standard LATE estimator based on a change in policy only by allowing the aggregate conditions to vary over time (although it requires treated and controls to be similarly affected by the market conditions). In the second
case, we discussed the use of family background to instrument participation. Clearly, the former is much more informative for the policy maker than the later. The estimated parameter based on family background will depend on the specific values being compared, may not represent a specific population that can be easily targeted and is more likely to raise arguments about the validity of the instrument (just as the discussion in the previous section).

6.4 The Marginal Treatment Effect

The meaning of the parameters identified by local IV methods is clarified in Heckman and Vytlacil (1999, 2000) and Carneiro, Heckman and Vytlacil (2004). These authors consider the estimation of the impact of treatment over the whole distribution of a continuous instrument. To do so, they use infinitesimal changes in the instrument and measure the limit of LATE as the change in instrument becomes arbitrarily small. As the whole distribution of local treatment effects is identified, all more aggregate parameters can also be estimated by integration over the distribution of the instrument.

To see this more formally, we start by considering the selection equation in our general model (equation (4)),

\[
d = \begin{cases} 
1 & \text{if } v > -Z\gamma \\
0 & \text{otherwise}
\end{cases}
\]

For simplicity, we assume that the only variable in \( Z \) is the instrument \( w \). The propensity score as a function of \( w \) is,

\[
p(w) = P[d = 1|w] = P[v > -w\gamma] = 1 - F_v(-w\gamma)
\]

where \( F_v \) is the distribution function of the unobservable \( v \).

We can now define \( \Xi = 1 - F_v(v) \) and notice that \( \Xi \) follows a uniform \([0, 1]\) distribution. \( \Xi \) is attached to the actual unobservable drawn by the individual, \( v \), and defines the group of individuals that are indifferent about participation at a given \( w \), say \( \Xi \). At \( w = \Xi \), indifference between participation and non-participation occurs for individuals drawing \( \Xi = p(\Xi) \). The participation decision
can now be re-written as,

\[ d = \begin{cases} 
1 & \text{if } v < p(w) \\
0 & \text{otherwise}
\end{cases} \]

Take two possible values for the instrument, \( w^* \) and \( w^{**} \). The LATE parameter is,

\[
\alpha^{LATE}(w^*, w^{**}) = \frac{E[y_i(w)|w = w^{**}] - E[y_i(w)|w = w^*]}{p(w^{**}) - p(w^*)}
\]

which identifies the impact of treatment among individuals with unobservable \( v \) between \(-w^{**} \gamma \) and \(-w^* \gamma \) or, which is the same, \( \subseteq \) between \( p(w^*) \) and \( p(w^{**}) \).

This can also be seen by taking the limits when \( w^{**} \) becomes arbitrarily close to \( w^* \). Let \( \delta \) represent the difference \( w^{**} - w^* \). As \( \delta \) approaches 0, the propensity score \( P[d = 1|w^* + \delta] \) approaches \( P[d = 1|w^*] = p \). In the limit, the LATE parameter measures the impact of treatment on individuals that are indifferent between participating and not participating at the point \( p(w) = p \). These are the individuals that draw \( \subseteq = p \). The limit of LATE as \( \delta \) approaches zero is called the Marginal Treatment Effect (MTE). It is defined as,

\[
\alpha^{MTE}(p) = \frac{E[y_i^1 - y_i^0 | \subseteq = p]}{\frac{\partial E[y_i | \subseteq = p]}{\partial p}}
\]

If data is rich enough to explore changes in treatment status over the whole distribution of \( v \) then all the average parameters, namely ATE, ATT, ATNT and LATE, can be expressed as averages of MTE using different weights (see appendix B for details). However, data may not be rich enough to allow for the estimation of MTE over the whole distribution of \( v \), in which case LATE may be the best available option. An example of this can be found in the next section, when IV and local IV are used to estimate the returns to education.

The estimation of ATT using MTE with a continuous instrument \( w \) requires the space of \( V, [0, 1] \), to be finely discretized. Then the MTE can be estimated each point on the grid. Estimation may use some non-parametric regression procedure to identify the slope of \( y \) with respect to \( w \) at the point being studied. This is the MTE at that point \( p \). The ATT may then be obtained by integrating the
set of MTE over the space of \( p \) using the distribution of \( p \) among the treated to weight the local effects.

6.5 Using IV to estimate the returns to education

Under the IV conditions, the variables in the selection process but not in the outcome equation may be used to instrument educational investment when ability is not observed. When applied to the model of educational investment we have been using, this means the family background \((b)\) is a valid instrument while the test score \((s)\) is not since it is correlated with ability, which directly affects earnings.

Table 4 displays some estimates of the ATT using the standard IV. We present the figures for the two scenarios depending on the availability of an educational subsidy and consider both uncorrelated and negatively correlated error terms, \( u \) and \( v \). In all cases, estimates use the correct logarithmic specification of the outcomes.

We expect the estimates based on standard IV techniques to be biased as the homogeneity assumption (IV1) is not met. Given this, the estimator based on the instrument \( b \) does surprisingly well, with biases around 10% in most cases (rows (2) and (5)). On the contrary, and as expected, the invalid instrument \( s \) produces significantly biased estimates in most cases (rows (3) and (6)).

Similar local IV estimates are presented in table 5 and show an interesting pattern. Columns (1)-(2) show, as expected, that family background \( b \) is a valid instrument while the test-score \( s \) always induces significant bias. Under the assumption of uncorrelated disturbances in the cost of education \((v)\) and earnings \((u)\) equations, the local IV technique based on the instrument \( b \) performs well, not displaying significant bias.

However, the bias is considerably larger in the case of correlated unobservable terms when education is instrumented with \( b \) (columns (3) and (4)). To understand the source of bias in the case of correlated residuals, recall that the local IV technique estimates the ATT by integrating the MTE over \( V \) and then \( X \). The MTE at \( p \) measures the impact of treatment on agents that change treatment status at the point \( V = p \) where \( p \) is the propensity score and \( V \) is the transformation of the unobservable in the selection rule, \( V = 1 - F_v(v) \). If \( p \) is not observable in the whole unit interval, we
will be unable to recover the ATT. In particular, the identification of the ATT will be affected in the absence of observations in the interval for the propensity score $[0, \bar{p}]$ for some $\bar{p}$ significantly larger than zero. In this case, we know that agents experiencing $V < \bar{p}$ will always prefer to participate within the observable range for $p$. That is, we never observe these sort of agents at their indifference point between participation and non-participation. Unfortunately, these agents are not a random sample of the population: they prefer to participate even at low levels of $p$ which means they expect to earn more from participation than most of the population. For these reasons, the estimated effect will not be the ATT but the average treatment for agents indifferent between participation and non-participation at the values of $p$ in the observable interval, $[\bar{p} > 0, \bar{p}]$. Given the above discussion, we can expect this impact to be lower then the ATT.\footnote{Not observing the top of the distribution of $p$ does not affect the identification of ATT since agents with $p > \bar{p}$ will never participate for the range of $p$’s observable. They are always non-participants.}

The above discussion is close to what has been argued in the literature about the ability of IV to produce interpretable parameters (see Heckman, 1997, Heckman and Vytlacil, ??, HE). The local parameters produced by IV are dependent on the particular instrument used and its ability to induce a change of treatment status in each individual. Even the estimation based on the MTE, which explicitly runs over the whole observable distribution of $p$ and uses the correct weights to aggregate the local parameters, may produce estimates that are not global. Instead, such estimates may be dependent on the particular instrument being used and apply only to the subpopulation of individuals that would switch treatment status at observable values of the instrument. Whether or not the identified parameter is of interest depends on the specific policy/evaluation question.

The lack of support affects the results in columns (3) and (4). In the uncorrelated disturbance case, the range of observable $p$ starts very close to zero. We are, therefore, able to identify the impact of education even for the agents that show a strong preference towards education. If the disturbances are correlated, however, only values of $p$ above 0.06 are observable. As an expected result, the obtained estimates are always downward biased.

A final remark concerns the use of an invalid instrument such as the test score, $s$. In both tables 4 and 5, the estimated bias is significantly reduced by the consideration of a positive subsidy. The reason for this lies on the agents response to the introduction of a educational subsidy. Contrary
Table 4: Monte Carlo experiment - IV estimates of ATT and respective bias in logs

<table>
<thead>
<tr>
<th></th>
<th>corr ((u, v) = 0)</th>
<th>corrr ((u, v) &lt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>bias</td>
</tr>
<tr>
<td>(1) True parameters</td>
<td>0.459</td>
<td>0.434</td>
</tr>
<tr>
<td>Panel A: No subsidy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV estimates using the instruments:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) (b) (family background)</td>
<td>0.421</td>
<td>0.083</td>
</tr>
<tr>
<td>(3) (s) (test score)</td>
<td>0.654</td>
<td>0.425</td>
</tr>
<tr>
<td>Panel B: Positive subsidy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4) True parameters</td>
<td>0.471</td>
<td>0.453</td>
</tr>
<tr>
<td>IV estimates using the instruments:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5) (b) (family background)</td>
<td>0.404</td>
<td>0.120</td>
</tr>
<tr>
<td>(6) (s) (test score)</td>
<td>0.537</td>
<td>0.170</td>
</tr>
</tbody>
</table>

Notes: Simulated data based on 200 Monte-Carlo replications using samples of 2000 observations each. All estimates obtained under the assumption that the true specification of the outcomes equation is additively separable in logs. Columns (1) and (2) present results obtained for independent error terms, \(u\) and \(v\). Columns (3) and (4) present results obtained for (negatively) correlated error terms, \(u\) and \(v\). ATT stands for “average treatment on the treated”.

to when no subsidy is available, many agents will now make a positive effort to score better on the test, and this effort is related to family background, \(b\). Thus, the relationship between test score and ability will be reduced while family background will now exhibit some relation with the test score. Although still an invalid instrument, \(s\) will now incorporate more exogenous variation that is related with participation, which helps in the identification of the true effect.
<table>
<thead>
<tr>
<th></th>
<th>corr ((u, v) = 0)</th>
<th>corr ((u, v) &lt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>bias</td>
</tr>
<tr>
<td>(1) True parameters</td>
<td>0.459</td>
<td>0.434</td>
</tr>
<tr>
<td>Panel A: No subsidy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV estimates using the instruments:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) (b) (family background)</td>
<td>0.491</td>
<td>0.070</td>
</tr>
<tr>
<td>(3) (s) (test score)</td>
<td>0.676</td>
<td>0.473</td>
</tr>
<tr>
<td>Panel B: Positive subsidy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IV estimates using the instruments:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5) (b) (family background)</td>
<td>0.484</td>
<td>0.028</td>
</tr>
<tr>
<td>(6) (s) (test score)</td>
<td>0.401</td>
<td>0.147</td>
</tr>
</tbody>
</table>

Notes: Simulated data based on 200 Monte-Carlo replications using samples of 2000 observations each. All estimates obtained under the assumption that the true specification of the outcomes equation is additively separable in logs. Estimation of the marginal treatment effect (MTE) over the support of the propensity score was based on a local quadratic regression using Epanechnikov kernel weights and a bandwidth of 0.4. Columns (1) and (2) present results obtained for independent error terms, \(u\) and \(v\). Columns (3) and (4) present results obtained for (negatively) correlated error terms, \(u\) and \(v\). ATT stands for “average treatment on the treated”.

62
7 Discontinuity Design

7.1 The discontinuity design estimator (DD)

Certain non-experimental policy designs provide sources of randomization that can be explored to estimate treatment effects under relatively weak assumptions. This is really the motivation for the natural experiment approach discussed earlier. However, a special case that has attracted considerable recent attention occurs when the probability of enrollment into treatment changes discontinuously with some continuous variable $w$. The variable $w$ is an observable instrument, typically used to determine eligibility. The discontinuity design (DD) approach uses the discontinuous dependence of $d$ on $w$ to identify the local average treatment effect even when the instrument does not satisfy the IV assumptions discussed before. Instead of some independence assumption, DD relies on a continuous relationship between the instrument and all the determinants of the outcome except for the participation variable. Any discontinuity in $y$ as a function of $w$ is, therefore, attributed to a discontinuous change in the participation rates as a function of $w$. As will be discussed, the parameter identified by DD is a local average treatment effect like the LATE parameter discussed under IV but is not necessarily the same parameter.17

Let $d(w)$ be a random variable indicating the treatment status. As above, the dependence of $d$ on $w$ means that the distribution of $d$ changes with $w$. The main source of identification used by DD is a discontinuity in $d(w)$ or its expected value at a given point in the distribution of $w$. Its most popular case, although empirically less frequent, is what is known by sharp design. It happens when $w$ fully determines participation on the basis of a threshold, $w^*$. The treated (non-treated) are individuals with values of $w$, say, above (below) the threshold. In this case, $d(w)$ is discontinuous at $w^*$, moving from being deterministically equal to 0 to being deterministically equal to 1. Thus, the identification condition with sharp design can be stated as follows,

$$d(w^*-\epsilon) \neq d(w^+\epsilon)$$

where, to simplify the notation, $d(w^*-\epsilon)$ ($d(w^+\epsilon)$) represents the limit of $d(w)$ as $w$ approaches $w^*$ from below (above). Both limits are assumed to exist. $d$ being a dummy variable, one of the limits

17For a detailed discussion of DD see Hahn, Todd and Van der Klaauw, 2001.
is 1 and the other is 0.

The fact that participation is a deterministic function of \( w \) means that individuals do not contribute to the decision process.\(^{18}\) The sharp design implies that the decision process is exogenously determined by \( w \) and all the selection is on the observables. Thus, the impact of treatment is arguably independent from the selection process, at least locally. Although selection occurs only on the observables, matching is not feasible given the absence of overlap between treated and controls once \( w \) is included in the set of covariates. Instead of the common support assumption used in matching, DD is based on the hypothesis that the outcome is a continuous function of \( w \) at \( w^* \). This way, any observed discontinuity in \( y \) at \( w^* \) results exclusively from the discontinuity in the participation rate. Under these conditions, the DD estimator for the sharp design case is:

\[
\alpha_{DD}(w^*) = E(y|w^*+) - E(y|w^*^-).
\]

where \( E(y|w^*+) \) (\( E(y|w^*^-) \)) is the limit of \( E(y|w) \) when \( w \) approaches \( w^* \) from above (below).

There are a few examples of economic studies that fall in the category of sharp design. They typically involve some exogenously imposed eligibility rule with a cut-off point. One example is the New Deal evaluation discussed above. Among other things, eligibility is based on age. Eligibles are those individuals that have not completed 25 years of age at completion of 6 months of unemployment. This rule is explored in Di Giorgi to estimate the impact of the New Deal on the oldest participants (see Di Giorgi??).

Possibly more common in economics is the fuzzy design. It occurs when the conditional probability of participation, \( P(d = 1|w) \), is discontinuous at \( w^* \). It happens when dimensions other than \( w \), particularly unobserved dimensions, also determine participation. To illustrate a possible fuzzy design consider our education example and suppose a subsidy is available for individuals scoring above a certain threshold \( w = w^* \) in a test required to enroll in university. Both subsidized and unsubsidized individuals will enroll. However, the threshold-rule is expected to create a discontinuity in the probability of enrollment at \( w^* \) since the cost of high education changes discontinuously at that point.

\(^{18}\) The possibility that individuals adjust \( w \) in response to the eligibility criteria in the intent of changing their participation status is ruled-out from the DD analysis.
Just as above, estimation under fuzzy design requires the outcome \( y \) to be a continuous function of \( w \) at \( w^* \) so that any observed discontinuity at that point can be linked to the discontinuous change in the participation rate. Also as above, the impact of treatment must be assumed locally independent of the participation decision \( d \). However, and unlike the sharp design case, this assumption is less credible here since unobserved factors are expected to contribute to determine participation along with \( w \), and such factors may be related with the potential gains from treatment when gains are heterogeneous (see van der Klaauw, 1997).

Under these assumptions, the DD estimator with fuzzy design is the sample counterpart of,

\[
\alpha^{DD}(w^*) = \frac{E(\frac{y}{w^*+}) - E(\frac{y}{w^*-})}{P(d=1|w^*+)} - P(d=1|w^*-)
\]

(42)

which identifies the local average treatment effect, \( E(\alpha|w^*) \). It measures the impact of treatment on individuals with characteristics \( w = w^* \). This is an average treatment effect at the local level since selection on idiosyncratic gains is excluded at the local level.

The local continuity and independence assumptions recover randomization under discontinuity in the odds of participation at the discontinuity point. The independence assumption is precisely a local version of (R2), meaning that ATE, is identifiable locally by discontinuity design. Note also that, under the independence assumption, ATE and ATT are locally equal. Assumption (R1) is not guaranteed to hold but instead the error term for the non-treated, \( u \), is required to be a continuous function of \( w \) at \( w^* \). Continuity ensures that it vanishes by differencing and taking the limits, thus ceasing to be a problem.

The DD estimator is simple to implement non-parametrically. It only requires running non-parametric regressions of \( y \) and \( d \) on \( w \) at each side of the discontinuity point. The expected value of \( y \) and \( d \) at \( w^* \) can then be computed at each side of the discontinuity and used to estimate the impact of treatment using (42).

### 7.2 The link between discontinuity design and IV

Interestingly, we have discussed the average treatment effect at a local level before, under IV. This was the LATE parameter or, when taking the limits using a continuous instrument, the MTE. To
understand the similarities and differences between DD and local IV we discuss what is identified by each method in the presence of a discontinuous jump in the participation rates as a function of an instrument, $w$.

Consider first the sharp design: on one side of the discontinuity point no one participates and at the other side everyone participates. In this case, gains from treatment are independent from $d$ since participation is exogenously determined. Thus both DD and local IV identify the ATE and this is the same as the ATT. Local IV assumes independence between potential outcomes and $w$ while DD makes the weaker assumption that potential outcomes are continuous function of $w$ at the cut-off point. For this reason, DD must compare the limit of the expected outcome as $w$ becomes arbitrarily close to $w^*$ while local IV may consider two discrete points close enough to $w^*$.

The fuzzy design is slightly more complex. DD is based on the assumption that gains from treatment are locally independent from the participation decision, thus identifying the local ATE. On its turn, local IV allows for selection in idiosyncratic gains, even locally. Instead, it imposes the monotonicity assumption to ensure switchers occur in only one direction when the instrument changes from one side of the threshold to the other. Thus, the identified parameter, LATE, is a local version of the ATT.

7.3 **Weaknesses of discontinuity design**

An obvious drawback of discontinuity design is its dependence on discontinuous changes in the odds of participation. This means that only the average parameter at a given point in the distribution of $w$ is identifiable. As in the binary instrument case of local IV, the discontinuity design is restricted to the discontinuity point which is dictated by the design of the policy. As discussed with respect to LATE with continuous instruments, the interpretation of the identified parameter can be a problem whenever the treatment effect, $\alpha$, changes with $w$.

To illustrate these issues, consider the context of our educational example. Suppose a subsidy is available for individuals willing to enroll in high education for as long as they score above a certain threshold $s$ in a given test. The introduction of such subsidy together with the eligibility rule creates a discontinuity in the odds of participation. On the other hand, the test score, $s$, and the returns
to education, $\alpha$, are expected to be (positively) correlated if both depend on, say, ability. DD excludes this relation by restricting the analysis to the neighborhood of $x$. But the local analysis only considers a specific subpopulation with a particular distribution of ability which is not that of the whole population or of the treated population. That is, the returns to education are estimated at a certain margin from where other more general parameters cannot be inferred.

Even if $\alpha$ is theoretically unrelated to $w$, the selection process may create a relation in the data that again precludes the estimation of population average effects. For example, suppose in the above example that from $x$ onwards, the subsidy is a monotonic increasing function of $s$. Thus, even if $\alpha$ and $s$ are independent, the fact that individuals with higher $s$ receive higher subsidies implies that they are also more willing to enrol into education, even if expecting lower returns. In such case, the local parameter identified by discontinuity design would over-estimate the ATT.

A final downside of DD which is also common with local IV relates to the implementation stage. By restricting analysis to the local level, the sample size may become too small to produce precise estimates of the treatment effect parameters.

7.4 Using discontinuity design to estimate the returns to education

Estimation using the discontinuity design method is only possible when the educational subsidy is available and the score eligibility rule is used for assignment.

Table 6 displays the monte carlo results using discontinuity design to estimate the impact of education at the eligibility threshold. We present estimates under the assumption of no correlation and negative correlation between the error components in the selection and outcome equations.

As expected, this method performs well, independently of whether $v$ and $u$ are correlated and of whether levels or logs are used. The robustness of DD is the expected outcome from the weak set assumptions on which it is based.

Despite the good performance of DD, whether the identified parameters are interesting depends on the characteristics of the problem and the policy question one wishes to address. To see why, notice that the DD measures the impact of education for the individuals at the threshold (in the present example, scoring about 4) that would invest in education would they be eligible to a subsidy
Table 6: Monte Carlo experiment - DD estimates of local ATE and bias

<table>
<thead>
<tr>
<th>Panel A: Estimates in levels</th>
<th>True parameter</th>
<th>global ATE</th>
<th>local ATE estimate</th>
<th>bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) corr ((u, v) = 0)</td>
<td></td>
<td>1.676</td>
<td>2.248</td>
<td>0.085</td>
</tr>
<tr>
<td>(2) corr ((u, v) &lt; 0)</td>
<td></td>
<td>1.673</td>
<td>2.297</td>
<td>0.040</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Panel B: Estimates in logs</th>
<th>True parameter</th>
<th>global ATE</th>
<th>local ATE estimate</th>
<th>bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3) corr ((u, v) = 0)</td>
<td></td>
<td>0.354</td>
<td>0.449</td>
<td>0.065</td>
</tr>
<tr>
<td>(4) corr ((u, v) &lt; 0)</td>
<td></td>
<td>0.354</td>
<td>0.465</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Notes: Simulated data based on 200 Monte-Carlo replications using samples of 2000 observations each. Estimation of the DD parameter at the eligibility point (score \(s = 4\)) was based on a local linear regression using Epanechnikov kernel weights and a bandwidth of 0.5. The true local parameters represent the impact of education for agents scoring between 3.99 and 4.01. Rows (1) and (3) present results obtained for independent error terms, \(u\) and \(v\). Rows (2) and (4) present results obtained for (negatively) correlated error terms, \(u\) and \(v\). Panel A uses the assumption that the true specification of the outcomes equation is additively separable in levels. Panel B uses the assumption that the true specification of the outcomes equation is additively separable in logs. ATE stands for “average treatment effect”.

but not otherwise. Like the LATE parameter, the DD is specially suited to address the questions related with potential extensions of the policy to a wider group by partially relaxing the eligibility rules.

The more uncertainty the individuals face about gains from treatment at the moment of deciding about participation, the closer the DD parameter (the local ATE) will be to the ATT. In our example, individuals make a more informed participation decision under correlated error terms, \(v\) and \(u\). In this case, and as noticed before, participants and non-participants will exhibit more pronounced differences in the returns to education. Therefore, the ATE and the ATT will also show larger differences. As expected, a local ATE as the one evaluated at \(s = 4\) by DD and displayed in column (2) in the table is much further away from the global ATT when the error terms are correlated (which equals 2.993) than when they are independent (which equals 2.356 and is very close to the local ATE of 2.248).
This is true in levels but not in logs, the reason being that the effect of treatment in log earnings does not depend on the shock $u$ as it is additively separable in logarithm.

8 Control Function Methods

8.1 The Control Function Estimator (CF)

When selection is on the unobservables, one attractive approach to the evaluation problem is to take the nature of the selection rule (4)-(5) explicitly into consideration in the estimation process (see Heckman, 1976). The control function method does exactly this, treating the endogeneity of $d$ as an omitted variable problem.\footnote{Below we examine the link between CF and IV methods in the binary treatment evaluation model considered here.}

Consider the outcome equation (11) together with the selection rule (4)-(5). In this section we omit the time subscript since only cross section data is generally required by the CF method. For simplicity of notation and exposition, we also drop the regressors $X$ in the outcome equation (considered under matching) as these can be easily introduced.

The CF approach is based on the assumption that, conditional on $v$, $u$ is independent of $d$ and $Z$. It can be formally stated as

$$u \perp d, Z \mid v$$

and it means that, would we be able to control for $v$, $d$ would become exogenous in the outcome’s equation. That is, assumption (43) allows for the variation in $d$ to be separated from that in $u$ by conditioning on $v$ (see, for example, Blundell and Powell, 2003).

Often it is only a conditional mean restriction that is required. After conditioning on other possible regressors in the outcome equation, $X$, (or, alternatively, if $d$ is additively separable from $X$) all is required is mean independence of $u$ from $d$ and $Z$ conditional on $v$:

$$E[u|v, d, Z] = E[u|v] = g(v)$$

where $g$ is a function of $v$, the control function.
Both (43) and (44) recover the randomization hypothesis (R1) *conditional* on the unobservable term $v$. As discussed before with respect to other non-experimental approaches, assumption (R2) is harder to reproduce and is not considered in most empirical studies. Thus, only the ATT will be identified in general.

The control function method is close to a fully structural approach in the sense that it *explicitly* incorporates the decision process in the estimation of the impact of the treatment. The problem is how to identify the unobservable term, $v$, in order to include it in the outcome equation. If $d$ is a *continuous* variable and the decision rule is invertible, then $d$ and $Z$ are sufficient to identify $v$. In such case, $v$ is a deterministic function of $(d, Z)$, making conditioning on $(v, d, Z)$ equivalent to conditioning on $(d, Z)$ alone, which is observable. However, if $d$ is discrete, and in particular if it is a dummy variable, then all that can be identified under typical assumptions is a threshold for $v$ as a function of $d$ and $Z$. This is made clear from the selection rule (4)-(5), where all that can be inferred when the parameters $\gamma$ are known is whether $v$ is above or below $-Z\gamma$ depending on whether $d = 1$ or $d = 0$.

Applications of the control function approach typically make a parametric assumption on the joint distribution of the error terms, $u$ and $v$. The most commonly encountered assumption imposes joint normality, which together with a dichotomous decision variable yields the Heckit estimator. Suppose $u$ and $v$ follow a bivariate normal distribution. The control function assumption (44) under the joint normality assumption becomes

$$
\begin{align*}
E[u|d = 1, Z] &= \rho \lambda_1(Z\gamma) \\
E[u|d = 0, Z] &= \rho \lambda_0(Z\gamma)
\end{align*}
$$

(45)

where $\rho = \sigma_u \text{corr}(u, v)$, $\sigma_u$ is the standard error of $u$, and the control functions are (adopting the standardization $\sigma_v = 1$ where $\sigma_v$ is the standard error of $v$)

$$
\lambda_1(Z\gamma) = \frac{\phi(Z\gamma)}{\Phi(Z\gamma)} \quad \text{and} \quad \lambda_0(Z\gamma) = \frac{-\phi(Z\gamma)}{1 - \Phi(Z\gamma)}
$$

Thus, joint normality implies that the conditional expectation of $u$ conditional on $d$ and $Z$ is a known function of the threshold, $Z\gamma$, that determines the assignment propensity: $P(d_i = 1|Z_i) = P(v_i > -Z_i\gamma|Z_i)$. 

70
Estimates of the control functions specified above can be obtained from a first stage regression of $d$ on $Z$. By including the estimated control functions in the outcome equation (11) we obtain

$$y_i = \beta + d_i (\alpha + E [\alpha_i - \alpha | d_i = 1]) + \left[ \rho d_i \frac{\phi (Z_i \gamma)}{\Phi (Z_i \gamma)} + \rho (1 - d_i) \frac{-\phi (Z_i \gamma)}{1 - \Phi (Z_i \gamma)} \right] + \delta_i$$

where $\delta$ is what remains of the error term in the outcome equation and is mean independent from $d$:

$$\delta_i = u_i + d_i \left( [\alpha_i - \alpha] - E [\alpha_i - \alpha | d_i = 1] - \hat{E} [u_i | Z, d_i = 1] \right) - (1 - d_i) \hat{E} [u_i | Z, d_i = 0]$$

It is clear from the regression equation (46) that all that is identified when the impact of treatment is heterogeneous is $\alpha^{ATT} = \alpha + E [\alpha_i - \alpha | d_i = 1]$.

### 8.2 Weaknesses of the control function method

The relative robustness of CF comes from the structure it imposes on the selection process. This makes this approach particularly informative for the policy maker by allowing for selection on the unobservables and supporting the extrapolation of results to alternative policy scenarios. However, this same feature has been strongly criticized being overly restrictive and possibly at odds with the data, in which case the estimated effects would be inconsistent and the predictions meaningless.

A number of semi-parametric CF estimators have been proposed that deal (at least partially) with this problem. Moreover, in its non-parametric setup, the CF approach has been shown to be equivalent to the LATE approach. While such advances deal with the main criticism to CF, they also reduce the usefulness of the CF approach to inform about possible changes in the policy scenario.

In what follows, we discuss precisely the less restrictive semi-parametric CF method.

### 8.3 The link between the control function and the instrumental variables approach

There are two key assumptions underlying specification (46): (i) the parametric assumption for the joint distribution of unobservables and (ii) the linear index assumption on the selection rule. These assumptions can be relaxed and the LATE approach can be shown to be an application of the non- (or semi-) parametric version of CF (see Vytlacil, 2001). To see this, we first compare the two methods and then briefly discuss the equivalence result of Vytlacil.
Consider the non-parametric selection model under additive separability of the error term. It can be written as:

\[ d_i = 1(v_i > -h_Z(Z_i)) \]  \hfill (47)

where \( h_Z \) is some function of \( Z \). This is a more general version of the selection model as specified in (4)-(5). Also consider the outcome’s model as specified in (1)-(2).

The CF method uses the additive separability of the unobservable term in the selection rule to acquire all the required information about the unobservable \( v \) from the observation of \( d \) and \( Z \). Estimation is based on the following regression function:

\[
E(y_i|Z_i, d_i) = \beta + d_i \left[ \alpha + E(\alpha_i - \overline{\alpha}|d_i = 1) \right] + d_i \Lambda_1(Z_i) + (1 - d_i) \Lambda_0(Z_i) \]  \hfill (48)

where \( \Lambda_1(Z_i) = E(u_i|d_i = 1, Z_i) \) and \( \Lambda_0(Z_i) = E(u_i|d_i = 0, Z_i) \). On the top of additive separability of \( v \) in the selection model, the following assumptions are required to establish (48):

- \( d \) is a non-trivial function of \( Z \) (which, for a continuous random variable \( v \) defined over the real line is equivalent to say that \( h_Z \) is a non-trivial function of \( Z \)).
- \( Z \) is independent of \((y^0, y^1, v)\).

The first assumption ensures that the probability of participation does change with \( p \). The second assumption ensures that \( Z \) does not affect the outcomes other than through its impact on participation and is exogenous on the participation equation, thus not affecting the distribution of \( v \).

The regression model (48) explicitly controls for the part of the error term \( u \) correlated with the participation status \( d \), thus eliminating the endogeneity problem.

On its turn, the LATE approach is based on the following regression model:

\[
E(y_i|Z_i) = \beta + p(d_i = 1|Z_i) \left[ \overline{\alpha} + E(\alpha_i - \overline{\alpha}|d_i = 1) \right] \]  \hfill (49)

To establish this relationship, we need the LATE assumptions discussed before. These are:

- \( d(Z) \) is a non-trivial function of \( Z \), where \( d(Z) \) is a random variable representing the treatment status and its dependence on \( Z \) means that the distribution of \( d \) depends on \( Z \).
\[ d(Z') \geq d(Z'') \ (\text{or} \ d(Z') \leq d(Z'')) \] for all individuals.

- \( Z \) is independent of \((y^0, y^1, d(Z))\).

The first assumption ensures that the probability of participation changes with \( Z \). The second assumption is the monotonicity assumption. It is required to ensure that the difference in the average outcomes evaluated at two distinct realizations of \( Z \), say \( Z' \) and \( Z'' \), is solely due to the move into treatment of some individuals (or out of treatment, depending on the direction of the monotonic relation) and not to a complete change in the composition of the treatment group due to some individuals moving in and some others moving out of treatment. The third assumption is the exclusion restriction together with the exogeneity of \( Z \) in the selection rule. The later is required to ensure that \( Z \) does not affect the outcome other than through \( d \), even if only indirectly through some relation with \( v \) that may itself be related with \( u \).

Notice that the LATE approach does not impose the selection model (47) to identify the treatment effect. It does not require any functional form distributional assumptions, instead relying on the general form for the decision process,

\[
d(Z) = 1(h_{Z,v}(Z_i, v_i) > 0)
\] (50)

Instead of the decision structure, the LATE approach uses the monotonicity assumption. As for the rest of the CF and LATE assumptions, they are equivalent. To see this, notice that the first assumption in each set is exactly the same. The independence assumptions are also equivalent: if \( v \) is independent of \( Z \) so will \( d(Z) \); if it is not then for general forms of \( h_{Z,v} \) in (50) neither will \( d(Z) \).

Now compare the two remaining assumptions: the selection rule under CF and the monotonicity assumption under LATE. The additive separability of the unobservable term in the selection rule obviously implies the monotonicity assumption of LATE since the decision process is based on a threshold rule: \( h_{Z}(Z') \) is either greater or smaller than \( h_{Z}(Z'') \) and so everyone that participates under the lowest one will also participate under the highest one.

The reverse implication, however, is not necessarily true. However, when taken together the LATE assumptions are equivalent to the CF assumptions. Vytlacil (2001) shows that under the LATE assumptions it is always possible to construct a selection model \( \tilde{d}(Z) \) of the type (47) satisfying
the CF assumptions and such that \( \tilde{d}(Z) = d(Z) \) almost everywhere. This means that under the LATE assumptions stated above, we can always find a selection model that rationalizes the data at hand. This equivalence result shows that the LATE approach can be seen as an application of a non-parametric version of the CF method.

Also notice that the local IV method of Heckman and Vytlacil (1999 and 2001) discussed earlier withdraws the monotonicity assumption of LATE and is instead based on the additive separability of the selection rule, as in (47). Thus, it is explicitly an application of CF approach.

8.4 Using the control function approach to estimate the returns to education

Table 7 displays the estimates for the ATT using the fully parametric CF approach. For the non-parametric CF estimates, we re-direct the reader to section 6.5, where local IV is discussed. The specification used in the estimation assumes that the outcome depends linearly on education and region. For the educational decisions we used a probit specification where the underlying variable depends linearly on the covariates listed in column (1).

Rows (2)-(3) and (9)-(10) of table (7) show the estimates of the ATT when only the correct observables in the selection process are included. In general, estimates exhibit large bias, not identifying the correct parameter. There are two reasons for this. The first is the problem of specification. The decision rule in (14) is a non-linear function of the observables \( b \) and \( x \). The second relates to the underlying assumption of joint normality of the disturbances in the selection rule and outcome’s equations. Although \((u, v)\) are jointly normally distributed, these are not the only unobservable components of the two equations. The permanent (unobservable) ability also affects the selection rule through its effect on the returns to education. The specification in levels suffers from the additional problem that \( \exp(u) \), not \( u \), is part of the disturbance.

We explore some possible solutions to these problems: (i) including the test score \( s \) to proxy ability; (ii) including ability itself; and (iii) allowing for a more flexible specification of the selection rule by including higher order terms for the continuous explanatory variables as well as interaction terms. Results are presented in rows (4)-(7) and (11)-(14).

Results for the specification in levels (panel A) improve significantly with the inclusion of the new
Table 7: Monte Carlo experiment - CF estimates of ATT and respective bias using the Heckit estimator. Simulations the assumption of negatively correlated residuals, $(u, v)$

<table>
<thead>
<tr>
<th></th>
<th>no subsidy</th>
<th>positive subsidy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>bias</td>
</tr>
<tr>
<td>(1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>$b$</td>
<td>1.280</td>
</tr>
<tr>
<td>(3)</td>
<td>$(b, x)$</td>
<td>1.485</td>
</tr>
<tr>
<td>(4)</td>
<td>$(b, s, x)$</td>
<td>2.070</td>
</tr>
<tr>
<td>(5)</td>
<td>$(b, s, x)+$interactions</td>
<td>2.205</td>
</tr>
<tr>
<td>(6)</td>
<td>$(b, \theta, x)$</td>
<td>3.299</td>
</tr>
<tr>
<td>(7)</td>
<td>$(b, \theta, x)+$interactions</td>
<td>3.353</td>
</tr>
</tbody>
</table>

Panel A: Estimates in levels

<table>
<thead>
<tr>
<th></th>
<th>no subsidy</th>
<th>positive subsidy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>bias</td>
</tr>
<tr>
<td>(8)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>True parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(9)</td>
<td>$b$</td>
<td>0.359</td>
</tr>
<tr>
<td>(10)</td>
<td>$(b, x)$</td>
<td>0.366</td>
</tr>
<tr>
<td>(11)</td>
<td>$(b, s, x)$</td>
<td>0.555</td>
</tr>
<tr>
<td>(12)</td>
<td>$(b, s, x)+$interactions</td>
<td>0.562</td>
</tr>
<tr>
<td>(13)</td>
<td>$(b, \theta, x)$</td>
<td>0.426</td>
</tr>
</tbody>
</table>
variables. Test score $s$ seems to help in the identification of ATT when ability is not observed. If available, the inclusion of ability drives the bias towards zero. Allowing for a more general specification of the selection rule does not improve results. Results for the specification in logs are not as bright. The inclusion of $s$ does not help and even ability may only improve the results slightly. The remaining problem in this case is the specification of the selection rule.

9 Summary

This paper has presented an overview of alternative empirical methods for the evaluation of policy interventions at the microeconomic level. The choice of appropriate evaluation method has been shown to depend on three central considerations: the policy parameter to be measured, the data available and the mechanism by which individuals are allocated to the program or receive the policy. Through studying a combination of the econometric underpinnings and the actual implementation of each method we hope to have convinced the reader that no method dominates. Indeed the requirements placed on the design of any evaluation to fully justify the use of any of the standard evaluation methods are typically difficult to satisfy.

One key to the appropriate choice of method has been shown to be a clear understanding of the ‘assignment rule’. That is, the mechanism by which assignment of individuals are allocated to the policy or program. In a sense this is a precursor to the choice of appropriate evaluation method. At one end of the spectrum, in a perfectly designed social experiment, assignment is random and at the other end of the spectrum, in a structural microeconomic model, assignment is assumed to obey some plausible model of economic allocation. Perfect designs and fully plausible allocation theories are rare. We have shown how alternative methods exploit different assumptions concerning assignment and differ according to the type of assumption made. Unless there is a convincing case for the reliability of the assignment mechanism being used, the results of the evaluation are unlikely to convince the thoughtful skeptic. Just as an experiment needs to be carefully designed a structural economic model needs to be convincingly argued.

But we have also seen that knowledge of the assignment mechanism alone is not enough. Each method will have a set of possible evaluation parameters it can recover. That is, even if the arguments
behind the assumed assignment rule is convincing, any particular method will typically only permit a limited set of policy questions to be answered. For example, we have seen that ex-ante evaluations that seek to measure the impact of policy proposals place inherently more stringent demands on the research design than ex-post measurements of existing policies. Similarly, measuring distributional impacts rather than simple average impacts typically also rests on stronger assumptions. Even where the randomisation assumption of an experimental evaluation is satisfied and is fully adopted in implementation, the experiment can only recover a limited set of parameters. In the end any reasonable evaluation study is likely to adopt a number of approaches, some being more robust but recovering less while others answering more complex questions at the cost of more fragile assumptions.
Appendix A: the educational model

Consider an economy of heterogeneous individuals indexed by $i$ facing lifetime earnings $y$ that depend on the highest level of education achieved. We distinguish between two levels of education, low and high. There is a (utility) cost linked to the acquisition of the highest level of education, denoted by $c$. The cost of education depends on the individual’s characteristics and on eligibility to an education subsidy when such subsidy is available. Eligibility is based on a test score, which itself depends on the individual’s characteristics and on the endogenously selected level of effort.

The prototypical individual in this model lives for three periods and solves a dynamic model of educational decisions. At birth, the individual is characterized by three variables: ability ($\theta$), family background ($z$) and region ($x$). These characteristics are assumed to remain unaltered throughout the individual’s life.

In period $t = 0$, the individual decides about the level of effort in school. This is the endogenous component of the test score, which will then impact on the cost of education faced by the individual. The test score conditional on effort is given by,

$$s_i = \gamma_0 + \gamma_1 \theta_i e_i + w_i$$  \hfill (51)

where $e$ is effort, $(\gamma_0, \gamma_1)$ are the parameters and $w$ is the unpredictable component. The test score is revealed in period $t = 1$ after the effort choice has been made. Conditional on the test score, the individual faces a cost of education that depends on family background as follows,

$$c_i = \delta_0 + \delta_1 z_i + 1(s_i > \underline{s})S + v_i$$  \hfill (52)

where $c_i$ is the cost of education faced by individual $i$, $\underline{s}$ is the threshold on the test score to determine eligibility, $S$ is the educational subsidy available to eligible individuals, the function $1(A)$ is the characteristic function, assuming the value 1 if $A$ is true and 0 otherwise, and $v$ is the unpredictable part of the cost of education.

Conditional on how effort affects the test-score (equation (51)) and its impact on the cost of education (equation (52)), the individual chooses effort in period $t = 0$ to maximise lifetime utility,

$$V_{0i}(\theta_i, z_i, x_i) = \max e_i \{ -\lambda e_i + \beta E_s,v[V_1i(\theta_i, z_i, x_i, s_i, v_i)] \}$$  \hfill (53)

where $V_{0i}$ represents the discounted value of present and future utility for individual $i$ in period $t$, $\beta$ is the discount factor and the index in the expectations operator lists the random variables at the moment of selecting effort, with respect to which the expected value is computed. From the above equation the optimal level of effort is a function of $\theta$, $z$ and $x$: $e^*(\theta, z, x)$.

Education is decided in period $t = 1$. The test score, $s$, and the unpredictable part of the cost of education, $v$, are revealed at the start of this period, before educational investment is decided. So,
the cost of education is known by the time of deciding about the investment. What is not known
certainty at this stage is the return to education as it depends on an unpredictable component
as viewed from period 1. Only in period 2 is this uncertainty resolved, when the individual observes
lifetime earnings which are specified as,

$$\ln y_i = \beta_0 + \beta_1 x_i + (\alpha_1 + \alpha_2 \theta_i) d_i + u_i$$  \hspace{1cm} (54)

where $y$ is earnings, $d$ is a dummy variable representing the education decision, the $\beta$’s and $\alpha$s are
the parameters of the earnings function and $u$ is the unpredictable component of earnings. Thus, the
individual’s problem at time 1 can be specified as

$$V_{1i}(\theta_i, z_i, x_i, s_i, v_i) = \max_d \{-c_i d_i + \beta E_u[y_i(\theta_i, d_i, x_i, u_i)|v_i]\}$$  \hspace{1cm} (55)

where we allow for $v$ and $u$ to be related and thus conditioning the expected value on $v$.

Under the model specification in equation (55), the education decision follows a reservation rule
defined in the cost of education,

$$d_i = \begin{cases} 
1 & \text{if } E_u(y_i|d_i = 1, v_i) - E_u(y_i|d_i = 0, v_i) > c_i \\
0 & \text{otherwise} 
\end{cases}$$  \hspace{1cm} (56)

Finally, in period 2 the individual works and collects lifetime earnings. There is no decision to be
made at this stage.

**Average parameters**

Suppose we aim at identifying the impact of education on the logarithm of earnings. The impact on
individual $i$ is simply given by,

$$\alpha_i = \alpha_1 + \alpha_2 \theta_i$$

We can use this expression to specify the ATE as

$$\alpha^{ATE} = \alpha_1 + \alpha_2 E(\theta_i)$$

$$= \alpha_1 + \alpha_2 \int_{\Theta} \theta f^{\theta}(\theta) \ d\theta$$

where $f^{\theta}$ is the probability density function of $\theta$ and $\Theta$ is the domain of $\theta$.

In a similar way, the ATT is just

$$\alpha^{ATT} = \alpha_1 + \alpha_2 E(\theta_i|d_i = 1)$$
However, it is now more difficult to derive the exact expression for $E(\theta|d_i = 1)$ as it depends on the endogenous individuals’ choices. To do this, we will assume that $v$ and $u$ are not positively correlated, thus $\text{corr}(v, u) \leq 0$. In particular, we take $u$ to be a linear random function of $v$,

$$u_i = \rho v_i + v_i$$

where $\rho \leq 0$ is the slope parameter and $v$ is a iid shock. In this case, the reservation policy described in equation (56) in terms of the cost of education $c$ can now be expressed in terms of the unobservable component, $v$. We denote it by $\tilde{v}$ and note that it is a function of the variables known at time 2 that impact either on the cost of education or on the expected future earnings. Thus, $\tilde{v} (\theta, z, x, s)$ but since $s = \gamma_0 + \gamma_1 \theta c (\theta, z, x) + u^s$ it is equivalent to write it as $\tilde{v} (\theta, z, x, u^s)$. The reservation policy $\tilde{v}$ fully characterizes the educational decision: whenever the individual draws a shock $v > \tilde{v}$ the decision will be not to participate while the opposite happens when $v < \tilde{v}$. Thus, the decision rule (56) can be re-written as,

$$d_i = \begin{cases} 
1 & \text{if } v_i < \tilde{v} (\theta, z_i, x_i, u^s_i) \\
0 & \text{otherwise}
\end{cases}$$

Conditional on the set of variables $(\theta, z, x, u^s)$, the size of the population investing in education will be given by,

$$F^{v} (\tilde{v} (\theta, z, x, u^s)) = \int_{-\infty}^{\tilde{v}(\theta, z, x, u^s)} f^v (v) \, dv$$

which is just the cumulative density function of $v$ at the reservation point, $\tilde{v} (\theta, z, x, u^s)$. If we integrate this over the whole population of $(\theta, z, x, u^s)$ we will obtain the size of the educated population. The expected value of $\theta$ conditional on enrolling in education will be,

$$E(\theta|d = 1) = \int_{\Theta} \int_{Z} \int_{X} \int_{-\infty}^{+\infty} \theta f^{v^*} (\tilde{v} (\theta, z, x, u^s)) f^{\theta, z, x, u^s} (\theta, z, x, u^s) \, du^s \, dx \, dz \, d\theta$$

where $\Theta$, $Z$, and $X$ stand for the domains of $\theta$, $z$ and $x$, respectively, and $f^{\theta, z, x, u^s}$ is the joint density function of $\theta$, $z$, $x$ and $u^s$.

**Parameters used in the simulations**

- Discount parameter: $\beta = 1$
- Utility cost of effort to prepare test: $\lambda = 0.9$
- Test score equation
\begin{align*}
\gamma_0 & : 1.0 \\
\gamma_1 & : 2.5 \\
w & : N(0,1)
\end{align*}

- Cost of education
  \begin{align*}
  \delta_0 & : 3.0 \\
  \delta_1 & : -1.2 \\
s & : 4.0 \\
S & : 2.5 \\
v & : N(0,1)
\end{align*}

- Earnings equation (54)
  \begin{align*}
  \beta_0 & : 0.70 \\
\beta_1 & : 0.30 \\
\alpha_1 & : 0.01 \\
\alpha_2 & : 0.70 \\
u_i & = \rho v_i + \xi_i \text{ where} \\
\rho & : -0.5 \text{ in the correlated case or} \\
& \quad 0 \text{ in the non-correlated case} \\
\xi & : N(0, \sigma^2 = 0.75) \text{ in the correlated case or} \\
& \quad N(0,1) \text{ in the non-correlated case}
\end{align*}

- The state variables, \( \theta, z \) and \( x \) are drawn from the following distributions,
  \begin{align*}
  \theta & : N(0.5, \sigma = 0.25) \text{ truncated at 0 and 1} \\
z & : N(0,1) \text{ truncated at } -2 \text{ and 2} \\
x & : \text{Bernoulli } p = 0.4
\end{align*}
Appendix B: Average treatment parameters

All the average parameters can be expressed as averages of the MTE using different weights. Consider the ATT. Participants at any point \( p \) of the distribution of \( V \) are those that draw \( V < p \). Thus,

\[
\alpha_{ATT}(p) = \int_0^p \alpha_{MTE}(V) f_V(V | V < p) \, dV
\]

\[
= \frac{1}{p} \int_0^p \alpha_{MTE}(V) \, dV
\]

where the second equality results from the fact that \( V \) is uniformly distributed. Integrating over all the support of \( p \) yields the overall ATT,

\[
\alpha_{ATT} = \int_0^1 \alpha_{ATT}(p) f_p(p | d = 1) \, dp
\]

\[
= \int_0^1 \int_0^p \alpha_{MTE}(V) \left( \frac{f_p(p | d = 1)}{p} \right) \, dV \, dp
\]

Similarly, ATE, ATNT and LATE are,

\[
\alpha_{ATE} = \int_0^1 \int_0^1 \alpha_{MTE}(V) f_p(p) \, dV \, dp
\]

\[
\alpha_{ATNT} = \int_0^1 \int_p^1 \alpha_{MTE}(V) \left( \frac{f_p(p | d = 0)}{1 - p} \right) \, dV \, dp
\]

\[
\alpha_{LATE}(p^*, p^{**}) = \frac{1}{p^{**} - p^*} \int_{p^*}^{p^{**}} \alpha_{MTE}(V) \, dV
\]
References


Figure 1: Outflows from the unemployment between months 6 and 10 conditional on having completed 6 months in unemployment. Comparing the treated group 19-24s in pilot areas with the two alternative control groups (19-24s in non-pilot areas and 25-30 in pilot areas).

Note: Figure from Blundell et al (2004).
Figure 2: Bias in matching estimates of the ATT and the ATNT by level of correlation between u and v.