Alternative Approaches to Evaluation in Empirical Microeconomics

Richard Blundell* and Monica Costa Dias

Institute for Fiscal Studies

April 2006

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.

*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.
## Contents

1 Introduction ........................................... 1

2 Which Treatment Parameter? .............................. 6
   2.1 Average Treatment Effects .................................. 6
   2.2 The selection problem and the assignment rule .............. 10
   2.3 An example: returns to education .......................... 11
       2.3.1 Homogeneous treatment effects ....................... 13
       2.3.2 Heterogeneous treatment effects ...................... 14

3 Social Experiments ...................................... 15
   3.1 Random assignment ......................................... 15
   3.2 Recovering the average return to education .................. 17

4 Natural Experiments .................................... 17
   4.1 The difference-in-differences (DID) estimator ................ 17
   4.2 Weaknesses of DID .......................................... 21
       4.2.1 Selection on idiosyncratic temporary shocks: the Ashenfelter’s dip 21
       4.2.2 Differential macro trends ............................. 22
       4.2.3 Compositional changes over time ...................... 23
   4.3 Non-linear models .......................................... 24
   4.4 Using DID to estimate returns to education .................. 27
       4.4.1 Monte-Carlo results .................................... 32

5 Matching Methods ...................................... 37
   5.1 The matching estimator (M) ................................. 37
   5.2 Propensity score matching .................................. 40
   5.3 Weaknesses of matching .................................... 41
   5.4 Using matching to estimate the returns to education .......... 42
5.4.1 Monte-Carlo results ........................................... 43
5.5 Combining matching and DID (MDID) ...................... 45

6 Instrumental Variables ......................................... 47
6.1 The instrumental variables (IV) estimator .................. 47
6.2 Weaknesses of IV .............................................. 49
6.3 The LATE parameter ............................................. 50
  6.3.1 The LATE assumptions ........................................ 52
  6.3.2 What does LATE measure? ............................. 53
6.4 Using IV to estimate the returns to education .............. 55

7 Discontinuity Design .............................................. 59
7.1 The discontinuity design estimator (DD) .................... 59
7.2 The link between discontinuity design and IV .............. 62
7.3 Drawbacks to discontinuity design ....................... 64
7.4 Using discontinuity design to estimate the returns to education 65

8 Control Function Methods ........................................ 67
8.1 The Control Function Estimator (CF) ......................... 67
8.2 The link between the control function and the instrumental variables approach .............................................. 70
8.3 A non-parametric CF approach: the local IV method ........ 71
8.4 Using the control function approach to estimate the returns to education ........................................ 72

9 Summary .......................................................... 74
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 some rules from an appropriate economic theory of behavior. 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

\(^1\)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.
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 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 analyses 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 educa-
tion 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 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 responses differ, do they differ in a systematic way? The distinction between homogenous and heterogeneous treatment responses is central to understanding what parameters alternative evaluation methods measure. In the homogeneous linear model, common in elementary econometrics, there is only one impact of the programme 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 programme

---

2In the labor market area, from which we draw heavily in this review, the ground breaking papers were those by Ashenfelter (1978), Ashenfelter and Card (1985) and Heckman and Robb (1985, 1986).
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

$$
\begin{align*}
y^1_{it} &= \beta + \alpha_i + u_{it} \quad \text{if } d_{it} = 1 \\
y^0_{it} &= \beta + u_{it} \quad \text{if } d_{it} = 0
\end{align*}
$$

(1)

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

$$
y_{it} = d_{it} y^1_{it} + (1 - d_{it}) y^0_{it}.
$$

(2)

so that

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

(3)

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 programme 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 programme 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 follows
\[
\alpha^{ATE} = E(\alpha_i)
\] (6)
\[
\alpha^{ATT} = E(\alpha_i | d_{it} = 1) = E(\alpha_i | v_{ik} > -Z_{ik}\gamma)
\] (7)
\[
\alpha^{ATNT} = E(\alpha_i | d_{it} = 0) = E(\alpha_i | v_{ik} < -Z_{ik}\gamma).
\] (8)

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.

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

\[
\alpha^{LATE}(Z', Z'') = E(\alpha_i | v_{ik} > -Z''\gamma \text{ and } v_{ik} < -Z'\gamma)
\] (9)

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,

\[
\alpha^{MTE}(Z) = E(\alpha_i | v_{ik} = -Z\gamma)
\] (10)

This parameter measures the impact of participation among the agents 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, \( \alpha \), 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 (the lower \( Z \)) 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, which is constant across the population in this case. The OLS estimator will then identify

$$E \left[ \hat{\alpha}^{OLS} \right] = \alpha + E [u_{it} | d_{it} = 1] - E [u_{it} | d_{it} = 0]$$

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

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

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 an education assignment model to study the behaviour of each of the non-experimental methods. In particular, we are interested in measuring the returns to education within a school enrollment 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 and will be used to generate a simulated dataset.

We consider individuals indexed by $i$ facing lifetime earnings $y$ that depend, among other things, on the level of education achieved. Individuals are heterogeneous at birth with respect to ability, $\theta$. Their lives are modelled in two periods, $t = 1, 2$. We assume
there are two levels of education, low and high. 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, \( z \), which we interpret as family background and the unobservable (to the researcher) \( v \),

\[
c_i = \delta_0 + \delta_1 z_i + v_i
\]  

where \( \delta_0 \) and \( \delta_1 \) are some parameters.

Period \( t = 2 \) represents the individual’s working life. Lifetime earnings are realised, 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 modelled as follows

\[
\ln y_i = \beta + \alpha_1 d_i + \alpha_2 \theta_i d_i + u_i
\]  

where \( d \) is a dummy variable for having completed higher education, \( \beta \) 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 education are heterogeneous in this model for as long as \( \alpha_2 \neq 0 \), in which case they depend on ability. In this particular example, 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 expected lifetime earnings in the two alternative scenarios

\[
E \left[ \ln y_i | d_i = 1, \theta_i, v_i \right] = \beta_0 + \alpha_1 + \alpha_2 \theta_i + E \left[ u_i | v_i \right]
\]

\[
E \left[ \ln y_i | d_i = 0, \theta_i, v_i \right] = \beta_0 + E \left[ u_i | v_i \right].
\]
The assignment (or selection) rule will therefore be
\[
d_i = \begin{cases} 
1 & \text{if } E[y_i|d_i = 1, \theta_i, v_i] - E[y_i|d_i = 0, \theta_i, v_i] > \delta_0 + \delta_1 z_i + v_i \\
0 & \text{otherwise}
\end{cases}
\]
so that investment in education occurs whenever the expected return exceeds the cost.

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 \((\theta, z)\). Then
\[
\tilde{v} (\theta_i, z_i) = E[y_i|d_i = 1, \theta_i, \tilde{v} (\theta_i, z_i)] - E[y_i|d_i = 0, \theta_i, \tilde{v} (\theta_i, z_i)] - \delta_0 - \delta_1 z_i
\]
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 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,
\[
d_i = \begin{cases} 
1 & \text{if } v_i < \tilde{v} (\theta_i, z_i) \\
0 & \text{otherwise}
\end{cases}
\]
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,
\[
\ln y_i = \beta + \alpha_1 d_i + u_i
\]
\( \alpha^{ATE} = \alpha^{ATT} = \alpha^{ATNT} = \alpha_1 \) while \( \alpha_1 \) also equals \( \alpha^{LATE} \) and \( \alpha^{MTE} \) for any choice of \( z \). In this case, the selection mechanism simplifies to \( \tilde{\nu}(z_i) \). If, in addition, \( v \) and \( u \) are mean independent, the selection process will be exclusively based on the cost of education, thus not affecting the ability of OLS identifying the true treatment effect.

### 2.3.2 Heterogeneous treatment effects

Under heterogeneous treatment effects, education returns vary and selection into education will generally take into account idiosyncratic gains which will then lead to differences in average treatment parameters. The ATE and ATT will now be,

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

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 \( E[\theta_i | v_i < \tilde{\nu}(\theta_i, z_i)] > E[\theta_i] \), meaning that average ability among educated agents is higher than average ability in the population. But then, \( \alpha^{ATT} > \alpha^{ATE} \).

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

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

and OLS identifies

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

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:

**R1:** \( E[u_i|d_i = 1] = E[u_i|d_i = 0] = E[u_i] \)

**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, $\alpha_i$. If compliance is not observed, the identifiable treatment effect parameter is,

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

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,

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

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

\(^4\)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.
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 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

\[ E[u_{it1} - u_{it0}|d_i = 1] = E[u_{it1} - u_{it0}|d_i = 0] = E[u_{it1} - u_{it0}] . \]

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 can only identify ATT under heterogeneous treatment effects.

Under the DID assumption we can write,

\[
E[y^d_t] = \begin{cases} 
\beta + E[\alpha_i|d_i = 1] + E[\phi_i|d_i = 1] + \psi_{t1} & \text{if } d = 1 \text{ and } t = t_1 \\
\beta + E[\phi_i|d_i = d] + \psi_t & \text{otherwise}
\end{cases}
\]

(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

\[
\hat{\alpha}^{DID} = \left[ y^d_{t1} - y^d_{t0} \right] - \left[ y^0_{t1} - y^0_{t0} \right]
\]

(17)

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

\[ E[\hat{\alpha}^{DID}] = E[\alpha_i|d_i = 1] = \alpha^{ATT} . \]

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

\[ y_{it1} - y_{it0} = \alpha_i d_{it1} + (\psi_{t1} - \psi_{t0}) + (\epsilon_{it1} - \epsilon_{it0}) \]

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 programme was rolled out in certain pilot areas prior to the national roll out. Since the programme 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 programme. 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

$$E(\hat{\alpha}^{DID}) = \alpha^{ATT} + E(\epsilon_{i1} - \epsilon_{i0} | d_{i1} = 1) - E(\epsilon_{i1} - \epsilon_{i0} | d_{i1} = 0)$$

To illustrate the conditions such inconsistency might arise, suppose a training pro-
gramme is being evaluated in which enrolment is more likely if a temporary dip in earnings occurs just before the programme 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 programme 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 labour 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.

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

$$E(u_{it} \mid d_i = d, t) = E(\phi_i \mid d_i = d) + k^d\psi_t$$

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

$$E(\alpha^{DID}) = \alpha^{ATT} + (k^1 - k^0) [\psi_{t1} - \psi_{t0}]$$

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, $(k^1 - k^0)[\psi_t - \psi_{t^0}]$. That is,

$$(k^1 - k^0) [\psi_t - \psi_{t^*}] = (k^1 - k^0) [\psi_{t_1} - \psi_{t_0}]$$

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,

$$\hat{\alpha} = \{ [\bar{y}_{t_1}^1 - \bar{y}_{t_0}^1] - [\bar{y}_{t_1}^0 - \bar{y}_{t_0}^0] \} - \{ [\bar{y}_{t^*}^1 - \bar{y}_{t^*}^0] - [\bar{y}_{t^*}^0 - \bar{y}_{t^*}^0] \}.$$  \(18\)

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 partic-
ular, 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.\textsuperscript{5} We denote by $F^{-1}$ the inverse probability rule. We simplify the model further by assuming

\textsuperscript{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.
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,

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

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.

Instead of imposing further restrictions in the model, we can follow by noticing that under the above stated assumptions,

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

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 \( F \),

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

\(^6\)This is generally required for non-linear discrete choice models (see Nickell, 1981).
Using this result it is obvious that the trend can be identified by comparing non-treated before and after treatment since,

$$\psi_{t_1} - \psi_{t_0} = F^{-1} \left[ E \left( y_{it}^0 | d_i = 0, t_1 \right) \right] - F^{-1} \left[ E \left( y_{it}^0 | d_i = 0, t_0 \right) \right]$$  \hspace{1cm} (20)

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

$$\psi_{t_1} - \psi_{t_0} = F^{-1} \left[ E \left( y_{it}^0 | d_i = 1, t_1 \right) \right] - F^{-1} \left[ E \left( y_{it}^0 | d_i = 1, t_0 \right) \right]$$  \hspace{1cm} (21)

But then, from (20) and (21) one notices that the unobserved counterfactual is,

$$F^{-1} \left[ E \left( y_{it}^0 | d_i = 1, t_1 \right) \right] = \left\{ F^{-1} \left[ E \left( y_{it}^0 | d_i = 1, t_1 \right) \right] - F^{-1} \left[ E \left( y_{it}^0 | d_i = 1, t_0 \right) \right] \right\} - \left\{ F^{-1} \left[ E \left( y_{it}^0 | d_i = 0, t_1 \right) \right] - F^{-1} \left[ E \left( y_{it}^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 \( \bar{\alpha} \). Then \(^7\)

$$\bar{\alpha} = \left\{ F^{-1} \left[ E \left( y_{it}^1 | d_i = 1, t_1 \right) \right] - F^{-1} \left[ E \left( y_{it}^0 | d_i = 1, t_1 \right) \right] \right\}$$

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

$$

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

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

where \( G(\alpha_i | i \in T) \) is the distribution of the treatment effect among he treated. Applying the inverse transformation yields,

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

$$\neq \int_\alpha F^{-1} (F(\beta + \alpha_i + \phi_1 + \psi_{t_1}))dG(\alpha_i | i \in T)$$

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

$$E(y^0|d = 1, t_1) = F\{F^{-1}\left[E(y^1|d = 1, t_1)\right] - \pi}\}

Using this expression, the ATT can be estimated by replacing the expected values by their sample analogues,

$$\hat{ATT} = \gamma_{t_1} - F\left[F^{-1}(\gamma_{t_1}) - \bar{\pi}\right]\]

where

$$\bar{\pi} = \left[F^{-1}(\gamma_{t_1}) - F^{-1}(\gamma_{t_0})\right] - \left[F^{-1}(\gamma_{t_1}) - F^{-1}(\gamma_{t_0})\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 labour 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 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

\(^8\)An extension to the discrete case is also considered by the authors.
researcher) component, \( w \)

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

(22)

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 unchanged 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),

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

(23)

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 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 \(\ln y_{Tt_0}, \ln y_{Tt_1}, \ln y_{Ct_0}\), and \(\ln y_{Ct_1}\) the average log earnings among \(T\) and \(C\) before and after the policy change, respectively. 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 \(d_{it_1} \geq d_{it_0}\) for \(i\) in \(T\) and, therefore, \(p_{Tt_1} \geq p_{Tt_0}\). In the control region we assume \(p_{Ct_1} = p_{Ct_0}\) 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),

\[
u_{it} = \phi_i + \psi_t + \epsilon_{it}\]

---

9 We discuss this type of monotonicity assumption in more detail later on, along with the LATE parameter.
yields under the DID assumptions,
\[ E \left[ \ln y_{Tt1} - \ln y_{Tt0} \right] = (\psi_{t1} - \psi_{t0}) + (p_{Tt1} - p_{Tt0}) E \left[ \alpha_i | d_{it1} = 1, d_{it0} = 0, i \in T \right] \]
meaning that only the impact on the movers is picked. Similarly,
\[ E \left[ \ln y_{Ct1} - \ln y_{Ct0} \right] = (\psi_{t1} - \psi_{t0}) \]
since individuals in the control region do not alter their educational decisions. Thus, under the DID assumption we identify,
\[ E \left[ \hat{\alpha}_{DID} \right] = (p_{Tt1} - p_{Tt0}) E \left[ \alpha_i | d_{it1} = 1, d_{it0} = 0, i \in T \right] \tag{24} \]
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
\[ E \left[ \hat{\alpha}_{DID} \right] = (p_{Tt1} - p_{Tt0}) E \left[ \alpha_i | d_{it1} = 1, d_{it0} = 0, i \in T \right] + (p_{Ct1} - p_{Ct0}) E \left[ \alpha_i | d_{it1} = 1, d_{it0} = 0, i \in C \right] \]
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 whether 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.

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 on two cohorts, affected and unaffected by the policy change, is observed.

The use of the eligibility rule instead of region variation suffers, in this case, from an 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 used 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 if such characteristics are affected by 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. We then estimate the impact of education in the alternative cases using DIDs 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 in our simulated model, agents that are aware of the new subsidy respond to it by significantly increasing their effort in period $t = 0$ in order to become 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

---

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 policy change</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 2: Monte Carlo experiment - DID estimates 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>Comparison by</td>
<td></td>
</tr>
<tr>
<td></td>
<td>eligibility status</td>
<td>region</td>
</tr>
<tr>
<td>(1) uncorrected estimates</td>
<td>0.348</td>
<td>0.056</td>
</tr>
<tr>
<td>(2) bias</td>
<td>62.3%</td>
<td>6.5%</td>
</tr>
<tr>
<td>(3) corrected estimates</td>
<td>0.595</td>
<td>0.531</td>
</tr>
<tr>
<td>(4) bias</td>
<td>21.0%</td>
<td>9.8%</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 (24)). 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 ???
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). If the subsidy is not announced in advance for the first generation of potential treated, however, 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 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 more strongly affected by the cost of education, related to family background, and less 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 education decisions 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 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.

Row 1 in table 2 shows results for the standard DID method. As argued before,
this method will under optimal conditions identify the Aggregate Effect in the context of
this example. 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. However, significant bias results from the comparison of
eligibles and ineligibles within the context of an announced policy. In this case, reactions
in anticipation to program participation to affect eligibility will change the composition
of the treated and control groups over time, rendering them incomparable.

Row 3 shows 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 but the identification conditions are
not met by the use of eligibility within the context of an expected policy change.

Both estimators are uninformative about the returns to education for and average in-
dividual 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
and that need that extra incentive to become educated.

The results obtained under the alternative assumption of (negatively) correlated resid-
uals resemble the ones presented here and are available under request. Since this addi-
tional 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 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:

\[
y_1^i = \beta + u(X_i) + \overline{\pi}(X_i) + [(u_i - u(X_i)) + (\alpha_i(X_i) - \overline{\pi}(X_i))]
\]
\[
y_0^i = \beta + u(X_i) + (u_i - u(X_i))
\]

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

The choice of the appropriate matching variables, $X$, is a delicate issue. In particular, the inclusion of post-treatment information on variables that can be affected by the treatment status must be avoided since such data account for part of the treatment effect. To the extent that the goal of evaluation methods is to control for selection, which happens
before treatment, the correct information to be used is that available to the individual at the time of deciding about participation. What remains unexplained is random with respect to treatment status and should, therefore, be included in the error term in this analysis.

The solution advanced by matching to estimate the impact of treatment is based on the following assumption,

**M1: (conditional independence assumption - CIA)** Conditional on the set of observables $X$, the non-treated outcomes are independent of the participation status

$$ y_i^0 \perp d_i \mid X_i $$

which is equivalent to the unobservable in the non-treated outcome equation being independent of the participation status conditional on $X$

$$ (u_i - u(X_i)) \perp d_i \mid X_i $$

This means that, conditional on $X$, treated and non-treated individuals are comparable in what respect to the outcome in the non-treatment case, $y^0$. Thus, there can be no selection on the unobservable term $u_i$ in (25). This assumption obviously implies a conditional version of the randomization hypothesis (R1),

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

which, under the usual hypothesis of exogeneity of $X$ yields $E[u_i]$. 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 (M1) is that 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 $y^0$ constitutes the correct 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. The second matching assumption is therefore
All treated individuals have a counterpart on the non-treated population and anyone constitutes a possible participant

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

Let \( S \) represent the common support of \( X \), that is, the subspace of the distribution of \( X \) that is both represented among the treated and the control groups. Under assumption (M2), \( S \) is the whole domain of \( X \). The matching estimator for the ATT is the empirical counterpart of

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

which is just the average difference in outcomes among treated and non-treated with the same \( X \)-characteristics over the part of the distribution of \( X \) represented in the two groups (\( S \)) and weighted by the distribution of \( X \) among the treated (\( dF(X \mid d = 1) \)).

In general, the form of the matching estimator is given by

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

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

The parameter identified by matching, \( \alpha^M \), may differ from the actual ATT if the common support is not the whole domain of \( X \) represented among the treated.

Identification of ATE requires a strengthened version of (M1)

**M1’**: (conditional independence assumption - CIA) Conditional on the set of observables \( X \), the two potential outcomes are independent of the participation status

\[ (y_i^0, y_i^1) \perp d_i \mid X_i \]
That is, on the top of (M1), ATE requires no selection on the unobservable idiosyncratic gain. Under (M1’), the matching estimator of ATE is the sample counterpart of

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

where, as before, \( S \) is the common support and the average is now weighted with the distribution of \( X \) over the whole population (\( F(X) \)).

5.2 Propensity score matching

A serious limitation to the implementation of matching is the dimension of the matching space as defined by \( X \). A more feasible alternative is to match on a function of \( X \). Usually, this is carried out on the propensity to participate given the set of characteristics \( X \): \( P(X_i) = P(d_i = 1 | X_i) \) the propensity score. Its use is usually motivated by Rosenbaum and Rubin’s result (1983, 1984), which shows that the CIA remains valid if controlling for \( P(X_i) \) instead of \( X_i \):

\[ y_i^0 \perp d_i | P(X_i) \]

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.

When using \( P(X) \), 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 issue is that of choosing the appropriate weights to associate the selected set of non-treated observations for each participant one. Several possibilities are commonly used. The Nearest Neighbor matching assigns a weight 1 to the closest non-treated observation and 0 to all others. 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 neighbor 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. As before the form of the matching estimator is given by $\hat{\alpha}^M$ in (26).

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 (M2) ensures that the missing counterfactual can be constructed from the population of non-treated. What (M2) 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 (M1) and (M2) 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 (M2). 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 than 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

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

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 $u$. However, 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

Suppose we now extend the outcome equation in the returns to education example (equation (13)) to include an observable regressor, $x$, which we interpret as region and can take two values, 0 and 1. We have

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

The impact of education on log earnings is now considered to be region-specific. The average impact of treatment conditional on region does not differ from the overall ATE since we are not considering sorting by region

$$\alpha^{ATE}(x) = \alpha_1 + \alpha_2 E(\theta|x) = \alpha_1 + \alpha_2 E(\theta).$$

The ATT, however, will depend region since selection into treatment is affected by region

$$\alpha^{ATT}(x) = \alpha_1 + \alpha_2 E(\theta|x, d = 1).$$
5.4.1 Monte-Carlo results

As before, 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 3 details the results obtained using log earnings under the existence of an education subsidy.\(^{11}\)

ATT estimates are presented in Panel A of table 3. Columns (1)-(3) and (4)-(6) display the estimates for the non-correlated and negatively correlated error terms, respectively. In each case, we present the true effect together with the matching estimate and the bias. Notice that true effects do vary 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 very 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), columns (1) and (4) for non-correlated and negatively correlated error terms, respectively.

We start by considering the case of independent error terms 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 case, the correct estimator of ATT uses region alone (row (2)). This is because region is the only variable explaining earnings for the non-educated. The inclusion of additional conditioning variables will both affect the identifiable parameter due to non-overlapping support problems and increase bias. This is showed in rows (3)-(6). As expected, worse bias occurs if the appropriate conditioning variables are not controlled for. Rows (3)-(5) show that excluding region \( (x) \) from the conditioning set yields the worse results.

Rows (7)-(13) show how much more difficult it can be estimating ATNT than ATT. Reproducing how the non-treated would fare in the treatment case requires controlling for

\(^{11}\)Estimates in levels and under the no-subsidy scenario show the similar patterns to the ones presented here and are available under request.
Table 3: Monte Carlo experiment - Matching estimates and bias in logs

<table>
<thead>
<tr>
<th></th>
<th>corr($u, v$) = 0</th>
<th></th>
<th>corr($u, v$) &lt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>true effect</td>
<td>estimate</td>
<td>bias</td>
</tr>
<tr>
<td>(1)</td>
<td>Simple difference</td>
<td>0.474</td>
<td>0.494</td>
</tr>
<tr>
<td></td>
<td>Matching using conditioning variables:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2)</td>
<td>$x$</td>
<td>0.474</td>
<td>0.469</td>
</tr>
<tr>
<td>(3)</td>
<td>$z$</td>
<td>0.474</td>
<td>0.494</td>
</tr>
<tr>
<td>(4)</td>
<td>$s$</td>
<td>0.444</td>
<td>0.518</td>
</tr>
<tr>
<td>(5)</td>
<td>$\theta$</td>
<td>0.473</td>
<td>0.534</td>
</tr>
<tr>
<td>(6)</td>
<td>($x, z, s, \theta$)</td>
<td>0.455</td>
<td>0.487</td>
</tr>
</tbody>
</table>

Panel A: Estimates of the ATT

(7) Simple difference 0.296 0.494 0.672 0.315 0.991 2.144

Matching estimates using conditioning variables:

(8) $x$ 0.296 0.470 0.590 0.315 0.970 2.080
(9) $z$ 0.295 0.585 0.984 0.315 1.163 2.691
(10) $s$ 0.294 0.332 0.131 0.315 1.153 2.656
(11) $\theta$ 0.296 0.317 0.068 0.318 1.027 2.231
(12) ($x, \theta$) 0.296 0.290 0.020 0.317 1.004 2.168
(13) ($x, z, s, \theta$) 0.296 0.254 0.142 0.319 1.263 2.964

Panel B: Estimates of the ATNT

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) 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 $v$. 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.
differences affecting non-treatment outcomes and gains from treatment. In our example, this means that ability ($\theta$) should be controlled for in addition to region. Row (12) displays the unbiased results obtained using the correct conditioning set. In the absence of information on ability, however, controlling for a proxy may help reduce the selection bias problem. As displayed in rows (10) and (13), this is true in our example when the test score is included in the conditioning set.

The good performance of matching under uncorrelated error terms cannot be reproduced if $u$ and $v$ are correlated. In this case, selection occurs partly on the unobservable term, $v$, given the information it embodies about future earnings and gains from treatment. Columns (2)-(6) and (8)-(13) show that, irrespectively of the conditioning set one may consider, matching will always incur in very large bias in this case.

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 forecast. However, the combination of matching with DID is able to accommodate unobserved determinant of the non-treated outcome impacting on participation for as long as it lies on a separable individual fixed component of the error term. Start by decomposing the unobservable term $u$ in (25) into a fixed effect ($\phi$), macro shock ($\psi$) and an idiosyncratic transitory shock ($\epsilon$)

$$
g_{it} = \beta + u(X_i) + \pi(X_i) + [(\phi_i + \psi_t + \epsilon_{it} - u(X_i)) + (\alpha_i(X_i) - \pi(X_i))]
$$

$$
g_{it}^0 = \beta + u(X_i) + (\phi_i + \psi_t + \epsilon_{it} - u(X_i))
$$

Under this specification, the following transformation of the CIA can be used achieve identification of ATT,

**MDID1:** Conditional on the set of observables $X$, the before-after difference in the unobservable $u$ is independent of the participation status

$$(u_{it_1} - u_{it_0}) \perp d_{it_1} \mid X_i$$
which, under specification (27) is the same as assuming

$$\epsilon_{it_1} - \epsilon_{it_0} \perp d_{it_1} \mid X_i$$

where \(t_0 < k < t_1\).

Since DID effectively controls for the other components of the outcomes under non-treatment, only the temporary individual-specific shock requires additional control. 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 (MDID1) 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 (M2) 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. Thus,

**MDID2**: All treated individuals have a counterpart on the non-treated population before and after the treatment and anyone constitutes a possible participant,

$$0 < P(d_{it_1} = 1 \mid X_i, t) < 1$$

where \(P(d_{it_1} = 1 \mid X_i, t)\) 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 \(X, S\). The following estimator is adequate to the use of propensity score matching with longitudinal data

$$\hat{\alpha}_{MDID,L} = \sum_{i \in T} \left\{ [y_{it_1} - y_{it_0}] - \sum_{j \in C} \omega_{ij} [y_{jt_1} - y_{jt_0}] \right\} \omega_i$$

46
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

\[
\tilde{\alpha}_{MDID,RCS} = \sum_{i \in T_1} \left\{ \left[ y_{it_1} - \sum_{j \in T_0} \omega_{ijt_0} y_{jt_0} \right] - \left[ \sum_{j \in C_1} \omega_{ijt_1} y_{jt_1} - \sum_{j \in C_0} \omega_{ijt_0} y_{jt_0} \right] \right\} \omega_i
\]

where \( T_0, T_1, C_0 \) and \( C_1 \) stand for the treatment and comparison groups before and after the program, respectively, and \( \omega^G_{ijt} \) represent the weights attributed to individual \( j \) in group \( G \) (where \( G = C \) or \( T \)) and time \( t \) when comparing with treated individual \( i \).

### 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 \( X \) (in many cases we will omit \( X \) for ease of notation). We also omit the index \( t \) since longitudinal or repeated cross-section data is not necessarily required to estimate the effect of treatment under the IV assumptions.

Consider the model (1)-(2) observed under the selection process (5)-(4). IV requires the existence of at least one regressor exclusive to the decision rule, that is, a variable \( z \) in \( Z \) but not in \( X \).

To begin our discussion we make the following assumptions:

**IV1:** The impact of treatment is homogeneous

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

---

\(^{12}\)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.
IV2: The unobservable component in the outcome equation (1), \( u \), is mean-independent of \( z \)

\[
E[u|z] = E[u]
\]

IV3: Conditional on the remaining regressors in \( Z \) (which we denote by \( Z_{-z} \)), the decision rule is a non-trivial (non-constant) function of \( z \)

\[
P[d = 1|Z_{-z}, z] \neq P[d = 1|Z_{-z}]
\]

Assumption (IV2) is the exclusion restriction, meaning that \( z \) has no impact on outcomes apart from through the treatment status, \( d \). The homogeneous treatment effects assumption (IV1) implies that selection occurs on the outcomes level only, \( \beta + u_i \), not on the gains from treatment, \( \alpha_i \).

The variable \( z \) is called the instrument: the source of exogenous variation used to approximate randomized trials. It provides variation that is correlated with the participation decision but does not affect the potential outcomes from treatment directly.

Under assumptions (IV1) and (IV2) we can write

\[
E(y_i | z_i) = \alpha P(d_i = 1 | z_i) + E(u_i | z_i)
\]

\[
= \alpha P(d_i = 1 | z_i) + E(u_i)
\]

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

\[
E(y_i | z_i = z^*) - E(y_i | z_i = z^{**}) = \alpha [P(d_i = 1 | z_i = z^*) - P(d_i = 1 | z_i = z^{**})]
\]

thus identifying the treatment effect from the ratio

\[
\alpha^{IV} = \frac{E(y_i | z_i = z^*) - E(y_i | z_i = z^{**})}{P(d_i = 1 | z_i = z^*) - P(d_i = 1 | z_i = z^{**})}
\]

(28)

as long as \( P(d_i = 1 | z_i = z^*) \neq P(d_i = 1 | z_i = z^{**}) \) (IV3). This is the standard IV estimator typically obtained from

\[
\alpha^{IV} = \frac{cov(y, z)}{cov(d, z)}
\]
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 (IV2), in which case IV is of no practical use. In other cases, the instrument $z$ has insufficient variation, which means that the estimation must rely on two very close values of $z$. In such case, the denominator in (28) 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 (IV1). If (IV1) does not hold, (IV2) is unlikely to hold as well since the unobservable component is now (see equation (11))

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

which, under (IV3), implies some dependence between $e$ and $z$. This will not be a problem if there is no selection on the idiosyncratic gains, $\alpha_i - \bar{\alpha}$, which then implies that $e$ is mean independent of $z$.

$$E[e|z] = E[u|z] + P[d = 1|z]E(\alpha_i - \bar{\alpha}|d = 1, z) = E[u]$$

since $E(\alpha_i - \bar{\alpha}|d = 1, z) = 0$. In this case, IV still identifies ATE which is not different from ATT given that 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 $\alpha_i$ and $z$. This is easily understood given that $z$ impacts on $d$, facilitating or inhibiting participation. For example, it may be that participants with values of $z$ that make participation more unlikely gain on average more from treatment than participants with values of $z$ that make participation more likely to occur.

Consider our education illustration. 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 (IV1) fails to hold, IV will not in general identify ATE or ATT. This happens because the average outcomes of any two groups differing on the particular $z$-realizations alone are different not only as a consequence of different participation rates but also because of compositional differences in the treated/non-treated groups according to the unobservables. 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 $z$ when (IV1) does not hold. The rationale is that some local changes in the instrument $z$ reproduce random assignment by inducing individuals to decide differently as they face different conditions unrelated to potential outcomes.

To discuss this parameter we define $y_i(z)$ as the outcome of individual $i$ at a given point of the instrument $z$. Thus, we can rewrite equation (2) by taking the instrument explicitly into account

$$y_i(z) = d_i(z)y^1_i + (1 - d_i(z))y^0_i$$

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

The use of IV when (IV1) does not hold requires a strengthened version of (IV2) and (IV3). We start by considering the following transformation of (IV2):
**IV2**: \((y^1_i, y^0_i, d_i(z))\) is jointly independent of \(z_i\).

which means that \(z\) 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, \(\alpha_i\).

Under (IV2') we can write

\[
E[y_i(z)|z] = P[d_i(z) = 1|z] E[y^1_i|d_i(z) = 1, z] + (1 - P[d_i(z) = 1|z]) E[y^0_i|d_i(z) = 0, z]
\]

\[
= P[d_i(z) = 1] E[y^1_i|d_i(z) = 1] + (1 - P[d_i(z) = 1]) E[y^0_i|d_i(z) = 0]
\]

\[
= E[y_i(z)]
\]

(29)
since, conditional on \(d\), \(z\) contains no extra information about the potential outcomes, \(y^0\) and \(y^1\). We can now use this result together with two possible values of \(z\), say \(z^*\) and \(z^{**}\), to write

\[
E[y_i(z)|z = z^{**}] - E[y_i(z)|z = z^*] = E[y_i(z^{**}) - y_i(z^*)]
\]

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

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

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

where the second equality is obtained by substituting in the expression for \(y(z)\) and the third equality uses the fact that whenever \(d_i(z^*) = d_i(z^{**})\) the expression in the expectations operator is nil. This expression means that, under (IV2'), any change in the average outcome \(y\) when \(z\) changes is solely due to changes in the treatment status of a subset of the population. However, the identification of the impact of treatment on individuals that change their participation decision still requires another assumption:

**IV3':** The decision rule is a non-trivial monotonic function of \(z\).

Under (IV3'), one of the terms in the last equality vanishes. Without loss of generality, suppose \(d\) is increasing in \(z\) and \(z^{**} > z^*\). Then \(P[d_i(z^{**}) < d_i(z^*)] = 0\) and

\[
E[y_i(z)|z = z^{**}] - E[y_i(z)|z = z^*] = P[d_i(z^{**}) > d_i(z^*)] E[y^1_i - y^0_i|d_i(z^{**}) > d_i(z^*)].
\]
This equation can be rearranged to yield the LATE parameter

\[
\alpha_{LATE}(z^*, z^{**}) = E\left[y_i^1 - y_i^0 | d_i(z^{**}) > d_i(z^*)\right] \\
= \frac{E\left[y_i(z) | z = z^{**}\right] - E\left[y_i(z) | z = z^*\right]}{P[d_i(z^{**}) > d_i(z^*)]} \\
= \frac{E\left[y_i(z) | z = z^{**}\right] - E\left[y_i(z) | z = z^*\right]}{P[d_i = 1 | z^{**}] - P[d_i = 1 | z^*]}
\]

(30)

which measures the impact of treatment on individuals that move from non-treated to treated when \(z\) changes from \(z^*\) to \(z^{**}\).

To illustrate the LATE approach, return to our education example and suppose \(z\) is family background. Participation is assumed to become more likely as \(z\) 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 \(z\) equals \(z^{**}\) some would not do so if \(z = z^*\). LATE measures the impact of education on the “movers” by attributing any difference in the average outcomes between the two groups defined by different family backgrounds to the different enrollment rates.\(^{13}\)

6.3.1 The LATE assumptions

Assumption (IV2’) is required to establish the result in equation (29). Even if \((y^0, y^1)\) are not directly related with \(z\), some relation may arise if \(d(z)\) is not independent of \(z_i\). \(z_i\) is related with \(d(z)\) if it is related to the unobservable in the selection rule, \(v\). If, furthermore, \(v\) is related to the unobservable in the outcome equation, \(u\), then the potential outcomes will in general be correlated with \(z\). In the education example, take \(z\) to be family background and assume it has an impact on the taste for education,

\(^{13}\)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 as a consequence of the particular change in the instrument considered.
included in \( v \). Thus, a change in \( z \) affects \( v \). If the taste for education is related with the taste for working, which is included in \( u \), a change in \( z \) may affect \( u \), thus altering the potential outcomes even among those that do not change treatment status. In such case, the population average outcome responds to a change in \( z \) not only through individuals altering their treatment status but also through changes in potential outcomes for the whole population, irrespective of their treatment status in the two \( z \)-scenarios.

The monotonicity assumption in (IV3’) is required for interpretation purposes. Under monotonicity of \( d \) with respect to \( z \), the LATE parameter measures the impact of treatment on individuals that move from non-treated to treated as \( z \) changes. If monotonicity does not hold, LATE measures the change in average outcome caused by a change in the instrument, which, under (IV2’), 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 \( z \) (see Heckman, 1997).

### 6.3.2 What does LATE measure?

Although very similar to the IV estimator presented in (28), LATE is intrinsically different since it does not represent ATT or ATE. LATE depends on the particular values of \( z \) 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. For instance, individuals benefiting the most from participation are more unlikely to be observed among the movers. The LATE parameter answers a different question, of how much individuals at the margin of participating benefit from participation given a change in policy. That is, it measures the effect of treatment on the sub-group of treated at the margin of participating for a given value of \( z \).

The meaning of the parameters identified by local IV methods is clarified in Heckman and Vytlacil (1999, 2000) and Carneiro, Heckman and Vytlacil (2004). Start by
considering the selection equation in our general model (equation (4)),

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

and notice that the propensity score as a function of \(Z\) is,

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

We can now define \(V = 1 - F_v(v)\) and notice that \(V\) follows a uniform \([0, 1]\) distribution. \(V\) 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 \(Z = z\). At \(Z = z\), indifference between participation and non-participation occurs for individuals drawing \(V = p(z)\). The participation decision can now be re-written as,

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

and the LATE parameter is,

\[
\alpha_{LATE}(z^*, z^{**}) = \frac{E[y_i(z) | z = z^{**}] - E[y_i(z) | z = z^*]}{p(z^{**}) - p(z^*)} = E[y_i^1 - y_i^0 | p(z^*) < V < p(z^{**})]
\]

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

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

$$\alpha^{MTE}(p) = E[y_i - y_i^0 | V = p]$$

$$= \frac{\partial E[y | V = p]}{\partial p}$$

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).

### 6.4 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 ($z$) 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 $z$ 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 $z$ 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 $z$ performs well, not displaying significant bias.
However, the bias is considerably larger in the case of correlated unobservable terms when education is instrumented with $z$ (columns (3) and (4)). To understand the source of bias in the case of correlated residuals, recall that the local IV technique estimates ATT by integrating the estimated 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 $[0, \underline{p}]$ for some $\underline{p}$ significantly larger than zero. In this case, we know that agents experiencing $V < \underline{p}$ 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, $[\underline{p} > 0, \overline{p}]$. 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 $p > \overline{p}$ will never participate for the range of $p$'s observable. They are always non-participants.}

This sort of problem 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
Table 4: Monte Carlo experiment - IV estimates of ATT and respective bias in logs

\[
\begin{array}{cccc}
\text{corr} (u, v) = 0 & \text{corr} (u, v) < 0 \\
\text{estimate} & \text{bias} & \text{estimate} & \text{bias} \\
(1) & (2) & (3) & (4) \\
\hline
\end{array}
\]

\[
\begin{array}{cccc}
\text{Panel A: No subsidy} \\
(1) & \text{True parameters} & 0.459 & 0.434 \\
\text{IV estimates using the instruments:} & \\
(2) & z & 0.421 & 0.083 & 0.392 & 0.097 \\
(3) & s & 0.654 & 0.425 & 0.652 & 0.502 \\
\hline
\text{Panel B: Positive subsidy} \\
(4) & \text{True parameters} & 0.471 & 0.453 \\
\text{IV estimates using the instruments:} & \\
(5) & z & 0.404 & 0.120 & 0.417 & 0.039 \\
(6) & s & 0.537 & 0.170 & 0.583 & 0.343 \\
\end{array}
\]

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”. \( z \) represents family background and \( s \) represents the test score.

of a educational subsidy. Contrary 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, \( z \). 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 5: Monte Carlo experiment - Local IV estimates of ATT and respective bias in logs

<table>
<thead>
<tr>
<th></th>
<th>corr ((u, v) = 0)</th>
<th></th>
<th>corr ((u, v) &lt; 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td>bias</td>
<td>estimate</td>
</tr>
<tr>
<td>Panel A: No subsidy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) True parameters</td>
<td>0.459</td>
<td>0.434</td>
<td></td>
</tr>
<tr>
<td>IV estimates using the instruments:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) (z)</td>
<td>0.491</td>
<td>0.070</td>
<td>0.381</td>
</tr>
<tr>
<td>(3) (s)</td>
<td>0.676</td>
<td>0.473</td>
<td>0.731</td>
</tr>
<tr>
<td>Panel B: Positive subsidy</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4) True parameters</td>
<td>0.471</td>
<td>0.453</td>
<td></td>
</tr>
<tr>
<td>IV estimates using the instruments:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5) (z)</td>
<td>0.484</td>
<td>0.028</td>
<td>0.384</td>
</tr>
<tr>
<td>(6) (s)</td>
<td>0.401</td>
<td>0.147</td>
<td>0.382</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”. \(z\) represents family background and \(s\) represents the test score.
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 $z$.\textsuperscript{15} The variable $z$ is an observable instrument, typically used to determine eligibility. The discontinuity design approach uses the discontinuous dependence of $d$ on $z$ to identify the local average treatment effect even when the instrument does not satisfy the IV assumptions discussed before. As will be discussed, the parameter identified by discontinuity design is a local average treatment effect like the LATE parameter discussed under IV but is not necessarily the same parameter.

The discontinuity design approach applies when $d$, or more generally $E(d|z) = P(d = 1|z)$, is a discontinuous function of $z$ at a certain point $z = z^\ast$. Suppose, therefore, that the following condition holds

$$\lim_{z \to z^\ast^-} P(d = 1|z) \neq \lim_{z \to z^\ast^+} P(d = 1|z)$$

(31)

where both limits exist. The superscripts $-$ and $+$ mean that the limit is taken as $z$ approaches $z^\ast$ from below and above, respectively. Equation (31) represents the requirement for $P(d = 1|z)$ to have a discontinuity at $z^\ast$.

If $d$ varies deterministically with $z$, condition (31) would simplify to

$$\lim_{z \to z^\ast^-} d(z) \neq \lim_{z \to z^\ast^+} d(z)$$

where, again, both limits exist. Therefore, $d$ being a dummy variable, one of the limits is 1 and the other is 0. This is typically known as the sharp design, where participation

\textsuperscript{15}For a detailed discussion of the discontinuity design method see Hahn, Todd and Van der Klaauw, 2001.
is deterministically resolved by the instrument $z$. It occurs when individuals are given no choice of whether to participate. To exemplify one such rule consider the case of compulsory assignment to special education for children scoring above (below) a certain threshold in a test but not otherwise. Possibly more common is a discontinuity in the conditional probability of participation, $P(d = 1|z)$. This is known as the *fuzzy* design. It happens when dimensions other than $z$, particularly unobserved dimensions, also determine participation. For instance, suppose a subsidy is available for individuals scoring above a certain threshold $z = z^*$ in a test required to enroll in university. Both subsidized and unsubsidized individuals will enroll but the threshold-rule is expected to create a discontinuity in the probability of enrollment at $z^*$ since the cost of high education changes discontinuously at that point.

Instead of requiring the independence assumptions used in IV, the discontinuity design is based on continuity assumptions. We will discuss this approach in the context of the outcome model used so far, again excluding the time dimension that is not explored by this method. Thus, consider the simplified version of equation (11)

$$y_i = \beta + \alpha_i d_i + u_i$$

$$= \beta_i + \alpha_i d_i$$

$$= \beta + \alpha d_i + e_i$$

where

$$e_i = (\beta_i - \beta) + d_i (\alpha_i - \alpha)$$

together with the following assumptions:

**DD1**: $E(\beta_i|z)$ as a function of $z$ is continuous at $z = z^*$.

**DD2**: $E(\alpha_i|z)$ as a function of $z$ is continuous at $z = z^*$.

**DD3**: The participation decision, $d$, is independent from the participation gain, $\alpha$, in the neighborhood of $z^*$.  

60
In the special homogeneous treatment effects case, assumptions (DD2) and (DD3) are always true, thus only (DD1) being required to identify \( \alpha \). In the general heterogeneous case, conditions (DD1) and (DD2) ensure that the expected values of the potential outcomes

\[
E(y_i^1|z) = E(\beta_i|z) + E(\alpha_i|z)
\]
\[
E(y_i^0|z) = E(\beta_i|z)
\]

are both continuous functions of \( z \) at \( z = z^* \). Condition (DD3) is an independence assumption but only applies locally. Under (DD3) we can write in the neighbourhood of \( z^* \)

\[
E(y_i|z) = E(\beta_i|z) + P(d_i = 1|z) E(\alpha_i|d = 1, z)
\]

since the gain \( \alpha \) is independent of \( d \) for \( z \) sufficiently close to \( z^* \). This ensures that \( E(y|z) \) is discontinuous at \( z^* \) as a consequence of the discontinuity in the odds of participation at that point only.

Using a small \( \delta > 0 \), we can now write

\[
E(y_i|z + \delta) - E(y_i|z - \delta)
\]

\[
= [E(\beta_i|z + \delta) - E(\beta_i|z - \delta)] + [P(d_i = 1|z + \delta) E(\alpha_i|z + \delta) - P(d_i = 1|z - \delta) E(\alpha_i|z - \delta)]
\]

which taking the limits as \( \delta \to 0 \) at \( z = z^* \) yields

\[
\lim_{z \to z^+} E(y_i|z) - \lim_{z \to z^-} E(y_i|z) = E(\alpha_i|z^*) \left[ \lim_{z \to z^+} P(d_i = 1|z) - \lim_{z \to z^-} P(d_i = 1|z) \right].
\]

The DD estimator of the impact of treatment at \( z = z^* \) is the sample counterpart of,

\[
\alpha^{DD}(z^*) = \frac{\lim_{z \to z^+} E(y_i|z) - \lim_{z \to z^-} E(y_i|z)}{\lim_{z \to z^+} P(d_i = 1|z) - \lim_{z \to z^-} P(d_i = 1|z)}
\]

which identifies the local average treatment effect, \( E(\alpha_i|z^*) \). It measures the impact of treatment on individuals with characteristics \( z \) close to \( z^* \). This is an average treatment
effect at the local level since selection on idiosyncratic gains is excluded at the local level by assumption (DD3). The importance of assumptions (DD1) and (DD2) is made clear from the derivation of the DD estimator above. Would $\beta$ or $\alpha$ had a discontinuity at $z^*$ and we would not be able to separate the average impact.

In the *sharp* design case, the denominator in the expression above reduces to 1 and the parameter identified by the discontinuity design approach is simply,

$$\alpha^{\text{DD}} (z^*) = E(\alpha_i | z^*)$$

$$= \lim_{z \to z^+} E(y_i | z) - \lim_{z \to z^-} E(y_i | z)$$

Notice how (DD1)-(DD3) recover randomisation under discontinuity in the odds of participation at the discontinuity point. Assumption (DD3) is precisely a local version of (R2), meaning that ATE, is identifiable locally by discontinuity design. Note however, that under (DD3), 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 $z$ at $z^*$. Continuity ensures that it vanishes by differencing and taking the limits, thus ceasing to be a problem.

### 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 and it is close to the local effect identified by discontinuity design. To see this, replace assumptions (DD2) and (DD3) by the LATE-type of conditions:

**DD2-3’**: *(i) $(\alpha_i, d_i(z))$ is jointly independent of $z$ close to $z^*$; (ii) $d$ is monotonic in $z$ close to $z^*$, say $d(z + \delta) > d(z)$ for a small $\delta > 0$ and $z$ close to $z^*$.*

Using the same notation as before while discussing LATE, write

$$y_i(z) = \beta_i + d_i(z) \alpha_i$$
Under (DD2-3'), we can write that for \( z \) close enough to \( z^* \)

\[
E[y_i(z) | z] = E(\beta_i | z) + P(d_i(z) = 1 | z) E[\alpha_i | z, d_i(z) = 1] \\
= E(\beta_i | z) + P(d_i(z) = 1) E[\alpha_i | d_i(z) = 1]
\]

where the second equality holds since both \( \alpha \) and \( d(z) \) are locally independent of \( z \) by assumption (DD2-3').

We can now use a small \( \delta > 0 \) to introduce variation around \( z^* \)

\[
E(y_i(z) | z = z^* + \delta) - E(y_i(z) | z = z^* - \delta)
\]  

(33)

Before proceeding, notice that the monotonicity assumption in (DD2-3'(ii)) implies that \( d(z^* + \delta) = 1 \) whenever \( d(z^* - \delta) = 1 \). This means that

\[
P(d_i(z^* + \delta) = 1) = P(d_i(z^* - \delta) = 1) + P(d_i(z^* + \delta) > d_i(z^* - \delta))
\]

But then the average treatment effect among participants at \( z = z^* + \delta \) can be decomposed as follows

\[
E(\alpha_i | d_i(z^* + \delta) = 1) \\
= \frac{P[d_i(z^* - \delta) = 1]}{P(d_i(z^* + \delta) = 1)} E[\alpha_i | d_i(z^* - \delta) = 1] \\
+ \frac{P[d_i(z^* + \delta) = 1]}{P(d_i(z^* + \delta) = 1)} - \frac{P[d_i(z^* - \delta) = 1]}{P(d_i(z^* + \delta) = 1)} E[\alpha_i | d_i(z^* + \delta) > d_i(z^* - \delta)]
\]

This expression can now be replaced in equation (33) above to yield

\[
E(y_i(z) | z = z^* + \delta) - E(y_i(z) | z = z^* - \delta) \\
= [E(\beta_i | z^* + \delta) - E(\beta_i | z^* - \delta)] \\
+ [P(d_i(z^* + \delta) = 1) - P(d_i(z^* - \delta) = 1)] E(\alpha_i | d_i(z^* + \delta) > d_i(z^* - \delta))
\]  

63
which, taking the limit as $\delta \to 0$ is

$$
\lim_{z \to z^*+} E(y_i \mid z) - \lim_{z \to z^*-} E(y_i \mid z) \\
= \lim_{\delta \to 0} E(\alpha_i \mid d_i(z^* + \delta) > d_i(z^* - \delta)) \left[ \lim_{z \to z^*+} P(d_i = 1 \mid z) - \lim_{z \to z^*-} P(d_i = 1 \mid z) \right]
$$

and can be used to identify the LATE parameter (or, more correctly, the MTE parameter),

$$
\alpha^{DD-LATE}(z^*) = \lim_{\delta \to 0} E(\alpha_i \mid d_i(z^* + \delta) > d_i(z^* - \delta))
$$

From the comparison of (32) with (34) it follows that the parameters identified under the two sets of assumptions are not generally the same. The LATE parameter, identified using discontinuity design (DD-LATE), is the average impact on individuals that move treatment status in response to the change in $z$. This is a local version of ATT since individuals are still allowed to select on idiosyncratic gains from treatment that are not captured by $z$. The standard discontinuity design approach, in contrast, identifies a local ATE parameter since selection on idiosyncratic gains is locally ruled-out. Under this assumption ATE and ATT are the same and both are identified by continuity design under assumptions (DD1)-(DD3). The case where both approaches identify the same parameters is that of a *sharp* design, where all eligibles and no non-eligible participate.

### 7.3 Drawbacks to 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 $z$ is identifiable. As in the binary instrument case of LATE, the discontinuity design is restricted to the discontinuity point which is dictated by the design of the policy. As discussed in the interpretation of LATE with continuous instruments, this can be a problem whenever the treatment effect, $\alpha$, changes with $z$.

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. Thus, by restricting the analysis to the neighborhood of \( s \), we only consider 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 \( z \), 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 \( s \) 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.

In the above discussion we have out-ruled some local dependence of \( d \) on \( \alpha \) in a neighborhood of the discontinuity point. Although probably less serious at the local level, this assumption can nevertheless be quite strong. For instance, if \( \alpha \) is independent of \( z \), the dependence of \( \alpha \) and \( d \) at the local level should not differ from that at the global level. In such case, the assumptions underlying LATE may be more appropriate and the local ATT is identified instead of the local ATE.

### 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.

Table 6: Monte Carlo experiment - DD estimates of local ATE and bias

<table>
<thead>
<tr>
<th>True parameter</th>
<th>global ATE</th>
<th>local ATE</th>
<th>estimate</th>
<th>bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A: Estimates in levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) corr ((u, v) = 0)</td>
<td>1.676</td>
<td>2.248</td>
<td>2.057</td>
<td>0.085</td>
</tr>
<tr>
<td>(2) corr ((u, v) &lt; 0)</td>
<td>1.673</td>
<td>2.297</td>
<td>2.388</td>
<td>0.040</td>
</tr>
<tr>
<td>Panel B: Estimates in logs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) corr ((u, v) = 0)</td>
<td>0.354</td>
<td>0.449</td>
<td>0.420</td>
<td>0.065</td>
</tr>
<tr>
<td>(4) corr ((u, v) &lt; 0)</td>
<td>0.354</td>
<td>0.465</td>
<td>0.477</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”.

As expected, this method performs well, independently of whether \(v\) and \(u\) are correlated or not 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 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 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 (1) 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). 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 becomes additively separable when the logarithm transformation is performed.

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. 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). The control function approach is based on the following assumption:

**CF1:** Conditional on $v$, $u$ is independent of $d$ and $Z$.

Assumption (CF1) 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 \( v \).

\[
\text{CF1': } E[u|v, d, z] = E[u|v] = g(v)
\]

where \( g \) is a function of \( v \), the control function.

Both (CF1) and (CF1’) 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 for the assignment rule 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 (5)-(4), 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 (CF1’) when applied to the evaluation model being
considered becomes

\[ E [u|d = 1, Z] = \rho \lambda_1 (Z\gamma) \]
\[ E [u|d = 0, Z] = \rho \lambda_0 (Z\gamma) \]

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

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

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

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

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

where \( \delta \) is what remains of the error term in the outcome equation

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

which is mean independent from \( d \). It is clear from the regression equation (36) that all that is identified when the impact of treatment is heterogeneous is \( \alpha^{ATT} = \bar{\alpha} + E [\alpha_i - \bar{\alpha}|d_i = 1] \).

There are two key assumptions underlying specification (36): (i) the parametric assumption for the joint distribution of unobservables and (ii) the linear index assumption \( Z\gamma \). These assumptions can be relaxed and the control function approach in a nonparametric setting can be shown to have the local IV interpretation discussed above. To see this we first draw the relationship between the CF and IV methods.
8.2 The link between the control function and the instrumental variables approach

For the model being analyzed here we can show that the CF assumptions hold under the IV conditions for the heterogenous effects model. Consider the selection model (4)-(5) together with the outcome equation (11) under the IV assumption (IV2') with respect to $Z$. Conditional on $Z$, an individual participates whenever $v > -Z\gamma$ so that

$$P[d = 1|Z = z] = P[v > -z\gamma|Z = z] = P[v > -z\gamma] = P(z)$$

where the second equality exploits the independence of $v$ and $Z$. We can now write

$$E(y_i|z_i, d_i) = \beta + d_i [\bar{\alpha} + E(\alpha_i - \bar{\alpha}|d_i = 1)] + d_i E(u_i|z_i, d_i = 1) + (1 - d_i)E(u_i|z_i, d_i = 0)$$

$$= \beta + d_i [\bar{\alpha} + E(\alpha_i - \bar{\alpha}|d_i = 1)] + d_i E(u_i|v_i > -z_i\gamma) + (1 - d_i)E(u_i|v_i > -z_i\gamma)$$

$$= \beta + d_i [\bar{\alpha} + E(\alpha_i - \bar{\alpha}|d_i = 1)] + d_i \Lambda_1(z_i\gamma) + (1 - d_i)\Lambda_0(z_i\gamma)$$

where the second equality uses the independence of $u$ and $Z$ conditional on the treatment status. In the joint normality case we have discussed above, $\Lambda_j$ would be $\rho \lambda_j$, where $\rho$ is as defined in (35) and $j = 0, 1$ for the non-treated and treated scenarios, respectively. This shows that the CF regression model can be derived using the IV assumption, (IV2').

In order to illustrate the links and differences between the two approaches, notice that the CF identification strategy uses the relation derived above

$$y_i = \beta + d_i [\bar{\alpha} + E(\alpha_i - \bar{\alpha}|d_i = 1)] + d_i \Lambda_1(z_i\gamma) + (1 - d_i)\Lambda_0(z_i\gamma) + \delta_i$$

(37)

where $\delta$ is, by construction, mean independent of $d$. This equation is estimated in a two step procedure that uses the first step estimates of $\gamma$.

In comparison, the IV approach uses the following conditional mean equation

$$E[y_i|z_i] = \beta + P(d_i = 1|z_i) [\bar{\alpha} + E(\alpha_i - \bar{\alpha}|d_i = 1)].$$

Here identification uses variation in $Z$ to identify the impact of treatment on individuals that change treatment status as a consequence of a change in $Z$. This is the approach followed by Imbens and Angrist (1994) to define the LATE parameter discussed earlier. Indeed, both methods coincide in the interpretation of local average treatment effects.
8.3 A non-parametric CF approach: the local IV method

Recall from the discussion of the MTE that, under the assumption of a continuously distributed error term $v$, participation is established by the condition

$$v > -z\gamma \Leftrightarrow V < p(z)$$

where $V = 1 - F_v(v)$ and $p(z) = P(d = 1|z) = 1 - F_v(-z\gamma)$. But then

$$d(z, v) = 1 * (v > -z\gamma)$$

$$= 1 * (V < p(-z)) = d(V, p).$$

The non-parametric identification of treatment effects using CF requires an exclusion restriction in addition to assumption (CF1')

**CF2:** $E(y|z, d) = E(y|z)$

Now suppose we know $V$. Under this exclusion assumption we can write

$$E(\alpha|d, V) = E(\alpha|p, V) = E(\alpha|V)$$

and under the CF1’ assumption

$$E(u_i|d, p, V) = E(u_i|V).$$

But then

$$E(y|d, V) = \beta + E(\alpha|V)d + E(u|V)$$

$$= \beta + E(\alpha|V)d(p, V) + E(u|V)$$

$$= E(y|p, V).$$

\[17\] In the case where the treatment is continuous and the excluded instrument is also continuous, estimation can proceed under a number of more general models. For example, Blundell and Powell (2004), use the control function approach to deal with the endogeneity of a continuous treatment variable in a semipararnetric binary choice (Probit) model.
We can now integrate over $V$ to obtain,

$$E(y|p) = \beta + \int_0^1 E(\alpha|V)d(p,V)f_V(V)dV + \int_0^1 E(u|V)f_V(V)dV$$

$$= \beta + \int_0^p E(\alpha|V)f_V(V)dV + E(u)$$

$$= \beta + E(\alpha|V < p)p + E(u)$$

which is precisely the equation used to estimate the local treatment effect using IV. That is, the local IV approach can be derived from the CF approach and could be interpreted as a CF method.

### 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.\(^{18}\) 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).

In most cases in this example, the parametric CF method does not perform well. The problem arises from an incorrect specification of the selection rule: the linear index assumption is not consistent with the structural model that generated the data. This is true despite the structural shocks being normally distributed, and independently of whether estimates use earnings in levels or in logs (the latter represents the correct specification for the outcomes’ equation according to our structural model).

Having said that, it is also worth noting that the importance of the functional form misspecification depends on the explanatory variables considered in the selection equation. Leaving out important explanatory variables in the selection rule will accentuate the bias. Such effect can be observed from the comparison of rows (3) and (4) or (7) and (8), particularly in what regards to the no-subsidy scenario (columns (2) to (5)). Results in row (3) and (7) restrict the explanatory variables in the selection equation to the test

---

\(^{18}\)For the non-parametric CF estimates, we re-direct the reader to section 6.4, where local IV is discussed.
Table 7: Monte Carlo experiment - CF estimates of ATT and respective bias using the Heckit estimator

<table>
<thead>
<tr>
<th></th>
<th>no subsidy</th>
<th></th>
<th>positive subsidy</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>corr ((u, v) = 0)</td>
<td>corr ((u, v) &lt; 0)</td>
<td>corr ((u, v) = 0)</td>
<td>corr ((u, v) &lt; 0)</td>
</tr>
<tr>
<td></td>
<td>estimate</td>
<td>bias</td>
<td>estimate</td>
<td>bias</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
<td>(4)</td>
</tr>
<tr>
<td></td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
<td>(8)</td>
</tr>
<tr>
<td>(1) True parameters</td>
<td>2.334</td>
<td>3.308</td>
<td>2.356</td>
<td>2.993</td>
</tr>
<tr>
<td>CF estimates using the selection variables:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) (z)</td>
<td>1.838</td>
<td>0.213</td>
<td>1.280</td>
<td>0.613</td>
</tr>
<tr>
<td>(3) (score)</td>
<td>3.940</td>
<td>0.688</td>
<td>5.733</td>
<td>0.733</td>
</tr>
<tr>
<td>(4) (z, score, \theta, region)</td>
<td>2.679</td>
<td>0.148</td>
<td>2.305</td>
<td>0.303</td>
</tr>
</tbody>
</table>

Panel B: Estimates in logs

<table>
<thead>
<tr>
<th></th>
<th>no subsidy</th>
<th></th>
<th>positive subsidy</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>corr ((u, v) = 0)</td>
<td>corr ((u, v) &lt; 0)</td>
<td>corr ((u, v) = 0)</td>
<td>corr ((u, v) &lt; 0)</td>
</tr>
<tr>
<td></td>
<td>estimate</td>
<td>bias</td>
<td>estimate</td>
<td>bias</td>
</tr>
<tr>
<td></td>
<td>(5)</td>
<td>(6)</td>
<td>(7)</td>
<td>(8)</td>
</tr>
<tr>
<td>(5) True parameters</td>
<td>0.459</td>
<td>0.434</td>
<td>0.471</td>
<td>0.453</td>
</tr>
<tr>
<td>CF estimates using the selection variables:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(6) (z)</td>
<td>0.395</td>
<td>0.139</td>
<td>0.359</td>
<td>0.173</td>
</tr>
<tr>
<td>(7) (score)</td>
<td>0.717</td>
<td>0.562</td>
<td>0.796</td>
<td>0.834</td>
</tr>
<tr>
<td>(8) (z, score, \theta, region)</td>
<td>0.519</td>
<td>0.131</td>
<td>0.507</td>
<td>0.168</td>
</tr>
</tbody>
</table>

Notes: Estimates based on the parametric CF approach under the assumption of joint normality of the residuals (Heckit estimator). Variables in the outcomes equation are education and region. The selection into education is modeled as a linear index model of the variables described in the first column of this table. Simulations are based on 200 Monte-Carlo replications using samples of size \(N = 2000\).
score. In the no educational subsidy scenario, individuals do not directly select on test scores. The role of test scores in this case is that of a proxy for ability, with small explanatory power of the decision process. The bias is always quite large in these cases. The inclusion of the remaining determinants of the participation decision significantly reduces bias, as can be observed from rows (4) and (8), columns (2) to (5). This is also generally true in the positive educational subsidy scenario but differences are much less pronounced.

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,

\[
s_i = \gamma_0 + \gamma_1 \theta_i e_i + w_i \tag{38}
\]

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,

\[
c_i = \delta_0 + \delta_1 z_i + \mathbf{1}(s_i > s) S + v_i \tag{39}
\]

where \(c_i\) is the cost of education faced by individual \(i\), \(s\) is the threshold on the test score to determine eligibility, \(S\) is the educational subsidy available to eligible individuals, the function \(\mathbf{1}(A)\) 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 (38)) and its impact on the cost of education (equation (39)), the individual chooses effort in period \(t = 0\) to maximise lifetime utility,

\[
V_{0i}(\theta_i, z_i, x_i) = \max_{e_i} \{-\lambda e_i + \beta E_{s,v} [V_{1i}(\theta_i, z_i, x_i, s_i, v_i)]\} \tag{40}
\]

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 with 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,

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

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

$$V_{i1}(\theta_i, z_i, x_i, s_i, v_i) = \max_d \{-c_i d_i + \beta E_u[y_i(\theta_i, d_i, x_i, u_i)|v_i]\}$$

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

Under the model specification in equation (42), the education decision follows a reservation rule defined in the cost of education,

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

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

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

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

77
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 \( E(\theta_i|d_i = 1) \) 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 (43) in terms of the cost of education \( c \) can now be expressed in terms of the unobservable component, \( v \). We denote it by \( \tilde{\nu} \) 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, \( \tilde{\nu}(\theta, z, x, s) \) but since \( s = \gamma_0 + \gamma_1 \theta e(\theta, z, x) + u^s \) it is equivalent to write it as \( \tilde{\nu}(\theta, z, x, u^s) \). The reservation policy \( \tilde{\nu} \) fully characterizes the educational decision: whenever the individual draws a shock \( v > \tilde{\nu} \) the decision will be not to participate while the opposite happens when \( v < \tilde{\nu} \). Thus, the decision rule (43) can be re-written as,

\[
d_i = \begin{cases} 
1 & \text{if } v_i < \tilde{\nu}(\theta_i, z_i, x_i, u^s_i) \\
0 & \text{otherwise}
\end{cases}
\]

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

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

which is just the cumulative density function of \( v \) at the reservation point, \( \tilde{\nu}(\theta, z, x, u^s) \). 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,

\[
E(\theta|d = 1) = \int_\Theta \int_Z \int_X \int_{-\infty}^{+\infty} \theta F^v(\tilde{\nu}(\theta, z, x, u^s)) f^{\theta, z, x, u^s}(\theta, z, x, u^s) du^s dx dz d\theta
\]

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
  \[ \gamma_0 : 1.0 \]
  \[ \gamma_1 : 2.5 \]
  \[ w : N(0, 1) \]
- Cost of education
  \[ \delta_0 : 3.0 \]
  \[ \delta_1 : -1.2 \]
  \[ \xi : 4.0 \]
  \[ S : 2.5 \]
  \[ v : N(0, 1) \]
- Earnings equation (41)
  \[ \beta_0 : 0.70 \]
  \[ \beta_1 : 0.30 \]
  \[ \alpha_1 : 0.01 \]
  \[ \alpha_2 : 0.70 \]
  \[ u_i = \rho v_i + \xi_i \text{ where} \]
  \[ \rho : -0.5 \text{ in the correlated case or} \]
  \[ 0 \text{ in the non-correlated case} \]
  \[ \xi : N(0, \sigma^2 = 0.75) \text{ in the correlated case or} \]
  \[ N(0, 1) \text{ in the non-correlated case} \]
- The state variables, $\theta$, $z$ and $x$ are drawn from the following distributions,
  \[ \theta : N(0.5, \sigma = 0.25) \text{ truncated at } 0 \text{ and } 1 \]
  \[ z : N(0, 1) \text{ truncated at } -2 \text{ and } 2 \]
  \[ x : \text{Bernoulli} \ p = 0.4 \]
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,

$$\alpha^{ATT} = \int_0^1 \int_0^1 \alpha^{ATT}(p) f_p(p|d = 1) \, dp \, dV \, dp$$

Similarly, ATE, ATNT and LATE are,

$$\alpha^{ATE} = \int_0^1 \int_0^1 \alpha^{MTE}(V) f_p(p) \, dV \, dp$$

$$\alpha^{ATNT} = \int_0^1 \int_0^1 \alpha^{MTE}(V) \frac{f_p(p|d = 0)}{1 - p} \, dV \, dp$$

$$\alpha^{LATE}(p^*, p^{**}) = \frac{1}{p^{**} - p^*} \int_{p^*}^{p^{**}} \alpha^{MTE}(V) \, dV$$

80
References


