

Alternative Approaches to Evaluation in Empirical Microeconomics

Richard Blundell* and Monica Costa Dias

Institute for Fiscal Studies

December 2007

Abstract

This paper reviews a range of the most popular policy evaluation methods in empirical microeconomics: 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 labor market policy evaluation literature.

Keywords: Evaluation methods, policy evaluation, matching methods, instrumental variables, social experiments, natural experiments, difference-in-differences, discontinuity design, control function.

JEL Classification: J21, J64, C33.

Acknowledgements: We would like to thank the editor and referees as well as graduate students and researchers at UCL and IFS for their helpful comments. 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	9
2.3	A ‘running’ evaluation example: returns to education	10
2.3.1	Homogeneous treatment effects	13
2.3.2	Heterogeneous treatment effects	13
3	Social Experiments	14
3.1	Random assignment	14
3.2	Recovering the average return to education	16
4	Natural Experiments	16
4.1	The difference-in-differences (DID) estimator	16
4.2	A DID Application: The New Deal Gateway in the UK	19
4.3	Weaknesses of DID	20
4.3.1	Selection on idiosyncratic temporary shocks: ‘Ashenfelter’s dip’	20
4.3.2	Differential macro trends	21
4.4	DID with Repeated Cross-sections: compositional changes	23
4.5	Non-linear DID models	24
4.6	Using DID to estimate returns to education	27
4.6.1	Monte-Carlo results	32
5	Matching Methods	36
5.1	The matching estimator (M)	36
5.2	Propensity score matching	40
5.2.1	The linear regression model and the matching estimator	43

5.3	Weaknesses of matching	44
5.4	Using matching to estimate the returns to education	45
5.4.1	Monte-Carlo results	46
5.5	Combining matching and DID (MDID)	49
6	Instrumental Variables	52
6.1	The instrumental variables (IV) estimator	52
6.2	Weaknesses of IV	54
6.3	The LATE parameter	55
6.3.1	The LATE assumptions	59
6.3.2	What does LATE measure?	60
6.4	The Marginal Treatment Effect	61
6.5	Using IV to estimate the returns to education	65
7	Discontinuity Design	68
7.1	The discontinuity design estimator (DD)	68
7.1.1	The sharp design	69
7.1.2	The fuzzy design	71
7.2	The link between discontinuity design and IV	74
7.3	Weaknesses of discontinuity design	75
7.4	Using discontinuity design to estimate the returns to education	76
8	Control Function Methods	78
8.1	The Control Function Estimator (CF)	78
8.2	Weaknesses of the control function method	81
8.3	The link between the control function and the instrumental variables approach	81
8.4	Using the control function approach to estimate the returns to education	84
9	Summary	85

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 subject to the intervention 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.

The choice of evaluation method 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. The last of these is typically labeled the ‘assignment rule’ and will be a key 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 economic theory. Alternative methods exploit different assumptions concerning assignment and differ according to the type of assumption made. Unless there is a convincing case for the reliability of the assignment mechanism being used, the results of the evaluation are unlikely to convince the thoughtful skeptic. Just as an experiment needs to be carefully designed, a structural economic model needs to be carefully argued.

In this review we consider six distinct, but related, approaches: (i) social experiment methods, (ii) natural experiment methods, (iii) discontinuity design methods, (iv) matching methods, (v) instrumental variable 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 modeling the assignment rule in order to fully control for selection in observational data.¹ The other methods

¹The 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.

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 randomization 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 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. The instrumental variable approach 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 evaluations it is frequently difficult to ensure that the experimental conditions have been met. Since programs are typically voluntary, those individuals ‘randomized in’ may decide not to participate in the treatment. The measured program impact will therefore recover an ‘intention to treat’ parameter, rather than the actual treatment effect. Further, unlike in many clinical trials, it is not possible to offer the control group a placebo in economic policy evaluations. Consequently individuals who enter a program and then are ‘randomized out’ may suffer a ‘disappointment’ effect and alter their behavior. Nonetheless, well designed experiments have much to offer in enhancing our knowledge of the possible impact of policy reforms. Indeed, a comparison of results from non-experimental data can help assess appropriate methods where experimental data is 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 labeled “difference-in-differences” since it is usually implemented by comparing the difference in average behavior before and after the reform for the eligible group with the before and after contrast for a comparison group. This 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 by relying on two critically important identifying assumptions of (i) *common time effects across groups*, and (ii) *no systematic composition changes within each group*. The evaluation of the ‘New Deal for the Young Unemployed’ in the UK is a good example of a policy 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 years old in non-pilot areas as a comparison 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 remove 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 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. Lalonde (1986) and Heckman, Ichimura and Todd (1998) demonstrate that experimental data can help in evaluating the choice of matching variables.

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 constant parameter 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 parameter can differ in unobservable ways across individuals, the IV estimator will only identify the average treatment effect under strong assumptions and ones that are unlikely to hold in practise. Work by Imbens and Angrist (1994) and Heckman and Vytlačil (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 endogenous selection. This approach relates directly to the selectivity estimator of Heckman (1979).

As already noted, structural microeconomic 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. However, results from evaluation approaches described above can be usefully adapted to assess the validity of a structural evaluation model. 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.

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

The rest of paper is organized as follows. In the next section we ask what are we trying to measure in program evaluation?² We also develop an education evaluation model 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 non-experimental 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 organizing 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.

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).

2 Which Treatment Parameter?

2.1 Average Treatment Effects

Are individual responses to a policy homogeneous or do responses differ across individuals? If the responses differ, do they differ in a systematic way? The distinction between homogenous and heterogeneous treatment responses is central to understand what parameters alternative evaluation methods measure. In the homogeneous linear model, common in elementary econometrics, there is only one impact of the program and it is one that would be common to all participants and nonparticipants alike. In the heterogeneous model, the treated and non-treated may benefit differently from program participation. In this case, the average treatment effect among the treated will differ from the average value overall or on the untreated individuals. 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 here and throughout the whole paper, we reserve Greek letters to denote the unknown parameters of the model and use upper case to denote vectors of random variables and lower case to denote random variables.

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 in the past and 0 otherwise. The potential outcomes for individual i at any time t are denoted by y_{it}^1 and y_{it}^0 for the treated and non-treated scenarios, respectively. They are specified as

$$\begin{aligned} y_{it}^1 &= \beta + \alpha_i + u_{it} \\ y_{it}^0 &= \beta + u_{it} \end{aligned} \tag{1}$$

where β is the intercept parameter, α_i is the effect of treatment on individual i and u is the unobservable component of y . The observable outcome is then

$$y_{it} = d_{it}y_{it}^1 + (1 - d_{it})y_{it}^0. \tag{2}$$

so that

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

Notice that this is a very general model as, for now, we have not yet imposed any functional form or distributional assumptions on the components of the outcome. In what follows we will show how different estimators use different sets of restrictions.

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 summarized by the observable variables, Z_k , and unobservable, v_k . Assignment to treatment is then assumed to be made on the basis of a selection rule

$$d_{it} = \begin{cases} 1 & \text{if } d_{ik}^* \geq 0 \text{ and } t > k, \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

where d^* is a function of Z and v

$$d_{ik}^* = g(Z_{ik}, v_{ik}) \quad (4)$$

A popular specification for the selection rule is based on the assumption of a linear index:

$$d_{it} = \mathbf{1}(Z_{ik}\gamma + v_{ik} \geq 0) \quad (5)$$

where γ is the vector of coefficients.

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

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

On the other hand, the appropriate parameter to identify the impact of the program on individuals of a certain type that were assigned to treatment is the ATT. Using the model specification above, we can express these three average parameters at time $t > k$ as follows

$$\alpha^{ATE} = E(\alpha_i) \tag{6}$$

$$\alpha^{ATT} = E(\alpha_i | d_{it} = 1) = E(\alpha_i | g(Z_{ik}, v_{ik}) \geq 0) \tag{7}$$

$$\alpha^{ATNT} = E(\alpha_i | d_{it} = 0) = E(\alpha_i | g(Z_{ik}, v_{ik}) \geq 0). \tag{8}$$

An increasing interest on the distribution of treatment effects has led to the study of additional treatment effects in the recent literature (Bjorklund and Moffitt, 1987, 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 , meaning that it changes with Z . Now suppose there exist two distinct values of Z , say Z' and Z'' , for which only a subgroup of participants under Z'' will also participate if having experienced Z' . The average impact of treatment on individuals that move for non-participants to participants when Z changes from Z' to Z'' is the LATE parameter

$$\alpha^{LATE}(Z', Z'') = E(\alpha_i | d_i(Z'') = 1, d_i(Z') = 0)$$

where $d_i(Z)$ is a dichotomous random variable representing the treatment status for an individual i drawing observables Z .

The MTE measures the change in aggregate outcome due to an infinitesimal change in the participation rate,

$$\alpha^{MTE}(P) = \frac{\partial E(y|P)}{\partial P}.$$

Under certain conditions, to be explored later, the MTE is a limit version of LATE.

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*

bias may be present.

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

$$\begin{aligned} y_{it} &= \beta + \alpha^{ATE} d_{it} + (u_{it} + d_{it} (\alpha_i - \alpha^{ATE})) \\ &= \beta + \alpha^{ATE} d_{it} + e_{it}. \end{aligned} \tag{9}$$

Non-random selection occurs if the unobservable term e in (9) 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 untreated 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 α 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. Here and throughout the whole paper, the discussion revolves around the identification of different treatment effects' parameters by alternative methods while sample analog estimators will also be presented.

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

$$y_{it} = \beta + \alpha d_{it} + u_{it}$$

where α is the impact of treatment on any individual and, therefore, equals α^{ATE} and any of the other parameters defined above since α is constant across the population in this case. The OLS estimator will then identify

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

which is in general different from α 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, $\alpha_i - \alpha^{ATE}$ (selection on gains). The parameter identified by the OLS estimator will now be

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

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

2.3 A ‘running’ evaluation example: returns to education

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

In the model, individuals differ with respect to educational attainment, which is determined by a number of observable and unobservable factors. Later on we will introduce an education subsidy and explore its use in the context of natural experiment, control function and instrumental variables methods. 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 appendix A.

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

$$c_i = \delta_0 + \delta_1 z_i + v_i \quad (10)$$

where δ_0 and δ_1 are some parameters.

In the second stage of life, $t = 2$, the individual is working. Lifetime earnings are realized, depending on ability, θ , educational attainment, d , and the unobservable u . We assume that u is unobservable to the researcher and is (partly) unpredictable by the individual at the time of deciding about education ($t = 1$). The logarithm of lifetime earnings is modeled as follows

$$\ln y_i = \beta_0 + \alpha_0 d_i + \alpha_1 \theta_i d_i + u_i \quad (11)$$

where β_0 is the intercept parameter and α_0 and α_1 are the treatment effect parameters for the general and ability-specific components, respectively.

The returns to high education are heterogeneous in this model for as long as $\alpha_1 \neq 0$, in which case such returns depend on ability. The individual-specific return is

$$\alpha_i = \alpha_0 + \alpha_1 \theta_i$$

We assume θ_i is known by individual i but not observable by the econometrician. The educational decision of individual i will be based on the comparison of expected lifetime earnings in the two

alternative scenarios

$$\begin{aligned} E[\ln y_i | d_i = 1, \theta_i, v_i] &= \beta_0 + \alpha_0 + \alpha_1 \theta_i + E[u_i | v_i] \\ E[\ln y_i | d_i = 0, \theta_i, v_i] &= \beta_0 + E[u_i | v_i]. \end{aligned}$$

with the cost of education in equation (10). Notice that we are assuming that z does not explain the potential outcomes except perhaps indirectly, through the effect it has on the education investment.

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 observed cost.

In this simple model, the education decision can be expressed by a threshold rule. Let \tilde{v} be the point at which an individual is indifferent between investing and not investing in education. It depends on the set of other information available to the individual at the point of deciding, namely (θ, z) . Then \tilde{v} solves the implicit equation

$$\tilde{v}(\theta_i, z_i) = E[y_i | d_i = 1, \theta_i, \tilde{v}(\theta_i, z_i)] - E[y_i | d_i = 0, \theta_i, \tilde{v}(\theta_i, z_i)] - \delta_0 - \delta_1 z_i.$$

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 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,

$$d_i = \begin{cases} 1 & \text{if } v_i < \tilde{v}(\theta_i, z_i) \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

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 $\alpha_1 = 0$ or $\theta_i = \theta$ over the whole population. In this case, the outcome equation (11) reduces to,

$$\ln y_i = \beta_0 + \alpha_0 d_i + u_i$$

and $\alpha^{ATE} = \alpha^{ATT} = \alpha^{ATNT} = \alpha_0$ while α_0 also equals α^{LATE} and α^{MTE} for any choice of z . In this case, the selection mechanism simplifies to $\tilde{v}(z_i)$.

If, in addition, v and u are mean independent, the selection process will be exclusively based on the cost of education. In this case, OLS will identify the true treatment effect α_0 .

2.3.2 Heterogeneous treatment effects

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

$$\begin{aligned}\alpha^{ATE} &= \alpha_0 + \alpha_1 E[\theta_i] \\ \alpha^{ATT} &= \alpha_0 + \alpha_1 E[\theta_i | v_i < \tilde{v}(\theta_i, z_i)]\end{aligned}$$

If α_1 is positive, then high ability individuals will have higher returns to education and the threshold rule \tilde{v} will be increasing in θ . This is the case where higher ability individuals are also more likely to invest in education. It implies that the average ability among educated individuals is higher than the average ability in the population because \tilde{v} will be increasing in θ and so $E[\theta_i | v_i < \tilde{v}(\theta_i, z_i)] > E[\theta_i]$. In this case it is also true that $\alpha^{ATT} > \alpha^{ATE}$.

Assuming θ is not observable by the analyst, the outcome equation (11) can be re-written as,

$$\ln y_i = \beta_0 + (\alpha_0 + \alpha_1 E(\theta)) d_i + (u_i + \alpha_1 d_i (\theta_i - E(\theta))).$$

and OLS identifies

$$\begin{aligned} E \left[(\widehat{\alpha_0 + \alpha_1 E(\theta)})^{OLS} \right] &= (\alpha_0 + \alpha_1 E(\theta)) + \alpha_1 E[\theta_i - E(\theta) | d_i = 1] + E[u_i | d_i = 1] - E[u_i | d_i = 0] \\ &= \alpha_0 + \alpha_1 E[\theta_i | d_i = 1] + E[u_i | d_i = 1] - E[u_i | d_i = 0] \end{aligned}$$

This is the ATT if (u, v) and (u, z) are two pairs of mean independent random variables, while the ATE will not be identified by OLS.⁴ Indeed, as will become clear from the discussion below, the ATE is much harder to identify.

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 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 we consider in this paper and described in equations (1)-(2), randomization 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 randomization ‘assumptions’ are required for recovering the average treatment effect (ATE).

⁴One could always think of controlling for z in the OLS regression if u and z are not mean independent. This is the motivation of the matching method, which will be discussed later in this paper.

Experiments are frequently impossible to implement. In many cases, such as in the education case, it may not be possible to convince a government to agree to exclude/expose individuals from/to a given treatment at random. But even when possible, experimental designs have 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 drop-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 not be observable, and this will determine the identifiable parameter.

To consider noncompliance further, 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, α_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 individuals 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.

Another possible problem results from the complexity of contemporaneous policies in developed countries and the availability of similar alternative treatments accessible to experimental controls.

The experiment itself may affect experimental controls as, for instance, excluded individuals may be “compensated” with detailed information about other available treatments, which in some cases is the same treatment but accessed through different channels. This would amount to another form of non-compliance, whereby controls obtain the treatment administered to experimental treatments.

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 enrolled in an 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,

$$E \left[(\alpha_0 + \widehat{\alpha_1 E(\theta)})^{OLS} \right] = \alpha_0 + \alpha_1 E(\theta)$$

which is the ATE.⁵

4 Natural Experiments

4.1 The difference-in-differences (DID) estimator

The natural experiment method makes use of naturally occurring phenomena that can be argued to induce some form of randomization across individuals in the eligibility or the assignment to treatment.

⁵Notice 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.

Typically this method is implemented using a before and after comparison across groups. It is 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. The policy shift may refer to a change of law in one jurisdiction but not another, to some natural disaster which changes a policy of interest in one area but not another, or to a change in policy that makes a certain group eligible to some treatment but keeps a similar group 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 make use of longitudinal data, where the same individuals are followed over time, or 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 longitudinal data is available. We assume a change in policy occurs at time $t = k$ and each individual is observed before and after the policy change, at times $t = t_0 < k$ and $t = t_1 > k$, respectively. For simplicity of notation, we denote by d_i (without the time subscript) the treatment group to which individual i belongs to. This is identified by the treatment status at $t = t_1$:

$$d_i = \begin{cases} 1 & \text{if } d_{it} = 1 \text{ for } t > k \text{ (in particular, } d_{it_1} = 1) \\ 0 & \text{otherwise} \end{cases}$$

The DID estimator uses a common trend assumption to rewrite the outcome equation (2) as follows

$$y_{it} = \beta + \alpha_i d_{it} + u_{it} \tag{13}$$

$$\text{where } E(u_{it} | d_i, t) = E(n_i | d_i) + m_t.$$

In the above equation, n is an unobservable individual fixed effect and m is an aggregate macro shock. Thus, DID is based on the assumption that the randomization hypothesis (R1) holds in first differences

$$E[u_{it_1} - u_{it_0} | d_i = 1] = E[u_{it_1} - u_{it_0} | d_i = 0] = E[u_{it_1} - u_{it_0}].$$

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

Under the DID assumption we can write,

$$E[y_{it}|d_i, t] = \begin{cases} \beta + E[\alpha_i|d_i = 1] + E[n_i|d_i = 1] + m_{t_1} & \text{if } d_i = 1 \text{ and } t = t_1 \\ \beta + E[n_i|d_i] + m_t & \text{otherwise} \end{cases} \quad (14)$$

It is now clear that we can eliminate both β and the error components by sequential differences

$$\begin{aligned} \alpha^{ATT} &= E(\alpha_i|d_i = 1) \\ &= [E(y_{it}|d_i = 1, t = t_1) - E(y_{it}|d_i = 1, t = t_0)] \\ &\quad - [E(y_{it}|d_i = 0, t = t_1) - E(y_{it}|d_i = 0, t = t_0)] \end{aligned} \quad (15)$$

This is precisely the DID identification strategy. The sample analog of equation (15) is the DID estimator:

$$\hat{\alpha}^{DID} = [\bar{y}_{t_1}^1 - \bar{y}_{t_0}^1] - [\bar{y}_{t_1}^0 - \bar{y}_{t_0}^0] \quad (16)$$

where \bar{y}_t^d is the average outcome over group d at time t . DID measures the excess outcome change for the treated as compared to the non-treated, this way identifying the ATT,

$$E[\hat{\alpha}^{DID}] = \alpha^{ATT}.$$

Notice that, 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 (13) to obtain

$$y_{it_1} - y_{it_0} = \alpha_i d_{it_1} + (m_{t_1} - m_{t_0}) + (o_{it_1} - o_{it_0})$$

where o represents the transitory idiosyncratic shocks. Under the DID assumptions, the above regression equation can be consistently estimated using OLS. Notice also that the DID assumption implies that the transitory shocks, o_{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 (14) 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.

4.2 A DID Application: The New Deal Gateway in the UK

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 et al. (2004) study investigates the impact of the program on employment in the first 18 months of the scheme. In particular it exploits an important design feature by which the program was rolled out in certain pilot areas prior to the national roll out. 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. 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 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 being age-specific implies that using slightly older people of similar unemployment duration is a natural comparison group. Second, the program was first piloted for 3 months (January to March 1998) in selected areas before being implemented nation-wide (the ‘National Roll Out’ beginning April 1998). The same age group in non-pilot areas is not only likely to satisfy the quasi-experimental conditions more closely but also allows for an analysis of the degree to which the DID comparisons within the treatment areas suffer from both general equilibrium or market level biases and serious substitution effects. Substitution occurs if participants take (some of) the jobs that non-participants would have got in the absence of treatment. Equilibrium wage effects may occur when the program is wide enough to affect the wage pressure of eligible and ineligible individuals.

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

4.3 Weaknesses of DID

4.3.1 Selection on idiosyncratic temporary shocks: ‘Ashenfelter’s dip’

The DID procedure does not control for unobserved temporary individual-specific shocks that influence the participation decision. If o is not unrelated to d , DID is inconsistent for the estimation of

ATT and instead approximates the following parameter

$$E(\hat{\alpha}^{DID}) = \alpha^{ATT} + E(o_{it_1} - o_{it_0} \mid d_{it_1} = 1) - E(o_{it_1} - o_{it_0} \mid d_{it_1} = 0)$$

To illustrate the conditions such inconsistency might arise, suppose a training program is being evaluated in which enrolment is more likely if a temporary dip in earnings occurs just before the program takes place - the so-called ‘Ashenfelter’s dip’ (see Ashenfelter, 1978, and Heckman and Smith, 1999). A faster earnings growth is expected among the treated, even without program participation. Thus, the DID estimator is likely to over-estimate the impact of treatment.

4.3.2 Differential macro trends

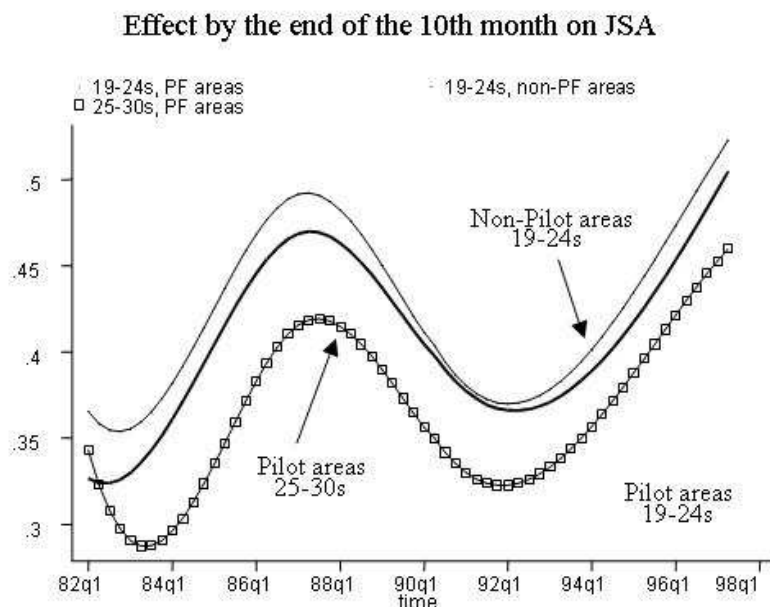
The identification of ATT using DID relies on the assumption that treated and controls experience common trends or, in other words, the same macro shocks. If this is not the case, DID will not consistently estimate the ATT. Differential trends might arise in the evaluation of training programs if treated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

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

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

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

Figure 1: Outflows from unemployment benefits (JSA) conditional on completing 6 months. Effect by the end of the 10th month after inflow. Men only.



Notes: PF stands for “Pathfinder” or “Pilot” areas. Figure plots the probability of leaving the unemployment claimant count by age and region of residence. From Blundell et al., 2004.

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

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

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

Given 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 τ_0 to τ_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 (τ_0, τ_1) one can separate the true impact of treatment from the differential trend.

More precisely, suppose one finds a pre-reform period, (τ_0, τ_1) for which the differential macro trend matches the bias term in the DID estimator, $(k^1 - k^0)[m_{t_1} - m_{t_0}]$. That is,

$$(k^1 - k^0)[m_{t_*} - m_{t_{**}}] = (k^1 - k^0)[m_{t_1} - m_{t_0}]$$

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 (τ_0, τ_1) , this condition allows one to identify the bias term $(k^1 - k^0)[m_{t_1} - m_{t_0}]$ 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 (τ_0, τ_1) . This is the differentially adjusted estimator proposed by Bell, Blundell and Van Reenen (1999), which will consistently estimate ATT,

$$\hat{\alpha} = \{[\bar{y}_{t_1}^1 - \bar{y}_{t_0}^1] - [\bar{y}_{t_1}^0 - \bar{y}_{t_0}^0]\} - \{[\bar{y}_{t_*}^1 - \bar{y}_{t_{**}}^1] - [\bar{y}_{t_*}^0 - \bar{y}_{t_{**}}^0]\}. \quad (17)$$

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.

4.4 DID with Repeated Cross-sections: compositional changes

Although DID does not require longitudinal data to identify the true ATT parameter, it does require similar treatment and control groups to be followed over time. In particular, in repeated cross-section surveys 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 illustrate this problem within our running education example.

4.5 Non-linear DID models

A restrictive feature of the DID method is the imposition of additive separability of the error term conditional on the observables, as specified in equation (13). Recent studies have proposed ways of relaxing this assumption. In their analysis of the New Deal for the Young People, Blundell et al. (2004) noted 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, DID loses much of its simplicity even under a very simple non-linear specification.

To extend DID to a non-linear setting, suppose the outcome equation is now:

$$y_{it} = \mathbf{1}(\beta + \alpha_i d_{it} + u_{it} > 0) \quad (18)$$

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

$$u_{it} = n_i + m_t - o_{it}$$

and the DID assumption holds,

$$E(u_{it}|d_i, t) = E(n_i|d_i) + m_t$$

where d_i represents the treatment group. Additional assumptions are required. We assume o follows a distribution F where F is invertible.⁶ Denote by F^{-1} the inverse probability rule. We simplify the model further by assuming a common group effect instead of allowing for an individual-specific effect: it is assumed that $n_i = n_d$ for $d = 0, 1$ being the post-program treatment status of individual i .⁷

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

⁶More precisely, we are assuming the transitory shocks, o , are *iid* continuous random variables with a strictly increasing cumulative density function, F , which is assumed known.

⁷This is generally required for non-linear discrete choice models (see Nickell, 1981).

this would identify some average of the treatment parameter α . One could then average the impact on y over the treated to recover the average treatment effect on the treated (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 (18) as follows:

$$y_{it} = \mathbf{1} (\beta + \alpha^{ATE} d_{it} + n_d + m_t - o_{it} + d_{it} (\alpha_i - \alpha^{ATE}) > 0)$$

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

Instead of imposing further restrictions in the model, we can progress by noticing that under our parametric setup,

$$E(y_{it}^0 | d_i = d, t) = F(\beta + n_d + m_t)$$

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 ,

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

Using this result it is obvious that the trend can be identified from the comparison of non-treated before and after treatment:

$$m_{t_1} - m_{t_0} = F^{-1} [E(y_{it}^0 | d_i = 0, t_1)] - F^{-1} [E(y_{it}^0 | d_i = 0, t_0)] \quad (19)$$

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

$$m_{t_1} - m_{t_0} = F^{-1} [E(y_{it}^0 | d_i = 1, t_1)] - F^{-1} [E(y_{it}^0 | d_i = 1, t_0)] \quad (20)$$

But then (19) and (20) can be combined to form the unobserved counterfactual as follows:

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

Let the average parameter of interest be α^{ATT} , which measures the average impact among the treated on the inverse transformation of the expected outcomes. Then ⁸

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

Rearranging, the missing counterfactual is

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

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

$$\widehat{ATT} = \bar{y}_{t_1}^1 - F [F^{-1} (\bar{y}_{t_1}^1) - \widehat{\alpha}^{ATT}]$$

⁸Notice that α^{ATT} is not the ATT since $F^{-1} [E (y_{it}^1 | d_i = 1, t_1)]$ is generally different from the average index for this group and time period (which is $\beta + \alpha^{ATT} + n_1 + m_{t_1}$) given the non-linearity of F^{-1} and the heterogenous nature of the treatment effect. To see why notice that,

$$E [y_{it}^1 | d_i = 1, t_1] = \int_{\mathcal{D}(\alpha)} F(\beta + \alpha + n_1 + m_{t_1}) dG_{\alpha|d}(\alpha | d_i = 1)$$

where $\mathcal{D}(\alpha)$ is the space of possible treatment effects, α , and $G_{\alpha|d}$ is the cumulation distribution function of α among individuals in treatment group d . Applying the inverse transformation yields,

$$F^{-1} (E [y_{it}^1 | d_i = 1, t_1]) = F^{-1} \left(\int_{\mathcal{D}(\alpha)} F(\beta + \alpha + n_1 + m_{t_1}) dG_{\alpha|d}(\alpha | d_i = 1) \right) \\ \neq \int_{\mathcal{D}(\alpha)} F^{-1} (F(\beta + \alpha + n_1 + m_{t_1})) dG_{\alpha|d}(\alpha | d_i = 1)$$

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

where

$$\widehat{\alpha}^{ATT} = [F^{-1}(\bar{y}_{t_1}^1) - F^{-1}(\bar{y}_{t_0}^1)] - [F^{-1}(\bar{y}_{t_1}^0) - F^{-1}(\bar{y}_{t_0}^0)]$$

Recently, Athey and Imbens (2006) have developed a general non-linear DID method specially suited for continuous outcomes: the “changes-in-changes” (CIC) estimator.⁹ 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).

4.6 Using DID to estimate returns to education

Since formal education occurs earlier in the life-cycle than labor market outcomes, it is generally not possible to evaluate the returns to education using earnings of treated and controls before and after the treatment. However, the comparison of different cohorts can be used but only when some exogenous change leads to differences in educational investments between cohorts. To explore this latter alternative, we consider a small extension to our simulation example with the introduction of an education subsidy.

In this extension 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, θ , an endogenously chosen level of effort, q , and an unpredictable (to the individual) and unobservable (to the researcher) component, w

$$s_i = \gamma_0 + \gamma_1 \theta_i q_i + w_i \tag{21}$$

where γ_0 and γ_1 are some parameters. Effort q carries some utility cost, as described in Appendix A. The (stochastic) payoff to this effort is the possibility of accessing subsidized education.

To investigate 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.

⁹An extension to the discrete case is also considered by the authors.

The new earnings equation for an individual i in the generation working at time t is,

$$\ln y_{it} = \beta_0 + \beta_1 x_i + (\alpha_0 + \alpha_1 \theta_i) d_i + u_{it} \quad (22)$$

where the distribution of u remains constant over time.

We first assume that the subsidy is launched at time k in region $x = 1$. From time k onwards, the cost of education in region $x = 1$ follows an adjusted form of equation (10),

$$c_i = \delta_0 + \delta_1 z_i - \mathbf{1}(s_i > \underline{s})S + v_i \quad (23)$$

where \underline{s} is the threshold rule defining eligibility to the subsidy and S is the new subsidy.

The question now is: Can we explore this regional change in policy to learn about 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. To put it shortly, there is non-compliance. 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 a monotonicity condition - that the introduction of the educational subsidy does not lead anyone to give up education. Instead, it makes education more attractive for all eligibles and does not change the incentives to invest in education among non-eligibles.¹⁰

Define the treatment and control groups as those living in regions affected ($x = 1$) and not affected ($x = 0$) by the policy change. Now suppose we have data on educational attainment and earnings

¹⁰We discuss this type of monotonicity assumption in more detail later on, along with the LATE parameter.

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 two regions over time using DID.

Designate by $\overline{\ln y_{xt}}$ the average log earnings in region x at time t . As before, d_{it} is a dummy variable indicating whether individual i in cohort t has acquired high education, and we define the probabilities

$$p_{xt} = P(d_{it} = 1 | t, x)$$

where i indexes individuals, x represents the region ($x = 0, 1$) and t represents time ($t = t_0, t_1$). Thus, p_{xt} is the odds of participation in region x 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 all i in region $x = 1$ and, therefore, $p_{1t_1} \geq p_{1t_0}$. In the control region we assume $p_{0t_1} = p_{0t_0}$ 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 (13),

$$u_{it} = n_i + m_t + o_{it}$$

yields under the DID assumptions,

$$E [\overline{\ln y_{1t_1}} - \overline{\ln y_{1t_0}}] = (m_{t_1} - m_{t_0}) + (p_{1t_1} - p_{1t_0}) E [\alpha_i | d_{it_1} = 1, d_{it_0} = 0, x_i = 1]$$

meaning that only the impact on the movers is identified. Similarly,

$$E [\overline{\ln y_{0t_1}} - \overline{\ln y_{0t_0}}] = (m_{t_1} - m_{t_0})$$

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

assumption we identify,

$$E [\widehat{\alpha}^{DID}] = (p_{1t_1} - p_{1t_0}) E [\alpha_i | d_{it_1} = 1, d_{it_0} = 0, x_i = 1] \quad (24)$$

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, $p_{1t_1} - p_{1t_0}$. This will identify the impact of education on individuals changing their educational status in response to a policy change - the LATE parameter.

Not correcting for the proportion of movers 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 $d_{it_1} \geq d_{it_0}$ for i in C . In this case, DID will identify

$$\begin{aligned} E [\widehat{\alpha}^{DID}] &= (p_{1t_1} - p_{1t_0}) E [\alpha_i | d_{it_1} = 1, d_{it_0} = 0, x_i = 1] + \\ &\quad (p_{0t_1} - p_{0t_0}) E [\alpha_i | d_{it_1} = 1, d_{it_0} = 0, x_i = 0] \end{aligned}$$

In this case, the ability to single out the impact of treatment on a subset of the movers (movers in $x = 1$ net of movers in $x = 0$) depends on two additional factors: *(i)* that movers in $x = 1$ in the absence of a policy change would have the same returns to education as movers in $x = 0$, which typically requires that they are similar individuals; and *(ii)* that different proportions of individuals

move in the two areas.

Now suppose that instead of a local policy, we are exploring the use of a global policy change, simultaneously introduced in the whole country. Instead of using treated and non-treated regions, one can think of using the eligibility rules as the source of randomization. Let us now define the treatment and control groups as composed by individuals scoring above and below the eligibility threshold, \underline{s} . Let \tilde{s} denote eligibility: \tilde{s} is 1 if $s \geq \underline{s}$ and is 0 otherwise. Again, we assume data is available on two cohorts, namely those affected and unaffected by the policy change.

The use of the eligibility rule instead of regional variation suffers, in this case, from one additional problem: the identification of the eligibility group before the introduction of the program. The affected generations will react to the new rules, adjusting their behavior even before their treatment status is revealed (which amounts to becoming eligible to the subsidy). In our model, future eligibility status can be influenced in anticipation by adjusting effort 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 eligibility groups over time. This means that eligibles and non-eligibles 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 (2003) when the dynamic nature of labor market decisions is acknowledged. Individuals may react in anticipation of treatment, trying to explore the policy rules. If the rules change, anticipatory behavior may also change, thus rendering individuals with similar characteristics incomparable when such characteristics are affected by the endogenous selection behavior that is not explicitly modeled. Reactions in anticipation to treatment are not observable and tend to change over time. Their occurrence may create a problem similar to the Ashenfelter dip described above as their potential impact on the outcome will be absorbed by the transitory unobservable component. Treated and controls with similar pre-treatment characteristics and outcomes will be inherently different as observables are endogenously affected by the individuals' prospects about treatment.

In our example, individuals may react to the new subsidy by increasing effort in the test, raising

test performance on average and increasing the odds of becoming eligible to subsidized education. Thus, the ability distribution of eligibles will be affected by the policy change, not only the educational choice.

4.6.1 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.¹¹ We then estimate the impact of education in the alternative cases using DID and both correcting and not correcting for the fact that not all treated actually take up education.

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

Rows 1 to 4 in table 1 display some measures of eligibility and educational attainment among different groups of individuals and policy scenarios. These numbers show that when individuals are aware of the new subsidy they respond by significantly increasing their effort in period $t = 0$ in the intent of becoming eligible to subsidized education. The log-linear functional form for wages adopted in this example implies that this is specially true in region 1, where wages are higher. As a consequence, an additional 12% of the population invests in education if the subsidy is available (rows 3-4, column 1), amounting to almost 100% of the eligibles (row 4, column 2). However, if the subsidy is not announced in advance the change in educational attainment is much more modest as fewer individuals become eligible to subsidized education: 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 individuals investing in education gain more from the investment than the average individual (rows 5-6). Row 7 shows that individuals moving into education in

¹¹Non-zero correlation between v and u implies that some selection on non-treated outcomes is expected.

Table 1: Monte Carlo experiment - description assuming u and v are independent

	Expected policy change			Unexpected policy change	
	All	Eligibles after	Population	All	Eligibles after
	population	policy change	in region 1	population	policy change
	(1)	(2)	(3)	(4)	(5)
<i>Eligibility and education take up</i>					
(1) % eligibles before policy	0.061	0.219	0.061	0.061	1.000
(2) % eligibles after policy	0.276	1.000	0.303	0.061	1.000
(3) % educated before policy	0.223	0.525	0.280	0.223	0.382
(4) % educated after policy	0.344	0.962	0.388	0.253	0.876
<i>True parameters</i>					
(5) ATE	0.354	0.502	0.355	0.354	0.495
(6) ATT	0.471	0.505	0.469	0.462	0.508
(7) LATE	0.492	0.492	0.484	0.479	0.479
(8) Aggregate effect	0.059	0.215	0.052	0.014	0.236

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.

Numbers 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.

response to the policy change benefit more from the investment than individuals investing in education in the absence of the subsidy. This is because the selection process in the absence of the subsidy is strongly affected by the cost of education (or family background), and less so 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 individuals unaffected by the policy (those not changing their educational decision in response to the subsidy) and the *effect of education* on individuals that invest only if the subsidy is available. It is, therefore, much lower than any of the other parameters as it measures the impact of the *subsidy*, not the impact of high education.

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

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

Rows 3 and 4 show similar results for the corrected DID method. Bias is now measured with respect to the LATE parameter and again, region and eligibility for an unexpected policy change can be used to identify the correct parameter (columns 2 and 3) but the identification conditions are not met by the use of eligibility within the context of an expected policy change (column 1).

All estimates are uninformative about the returns to education for the average individual or for the

Table 2: Monte Carlo experiment - DID estimates assuming u and v are independent

		Expected policy change		Unexpected policy change
		Comparison by		Comparison by
		eligibility status	region	eligibility status
		(1)	(2)	(3)
(1)	uncorrected estimates	0.348	0.056	0.241
(2)	bias	62.3%	6.5%	2.1%
(3)	corrected estimates	0.595	0.531	0.479
(4)	bias	21.0%	9.8%	0.1%

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 1. Row 4 displays the relative bias of corrected estimates as compared to LATE in row 7 of table 1.

educated. Instead, they use the change in policy to identify the impact of education on a particular group of individuals: those at some margin of participating that respond to the extra incentive by becoming educated.

The results obtained under the alternative assumption of (negatively) correlated residuals resemble the ones presented.¹² Since this additional source of selection does not affect our ability to identify treated and controls before the policy change, an unbiased estimator under the setup discussed above will remain unbiased under the alternative setup.

¹²These are available from the authors under request.

5 Matching Methods

5.1 The matching estimator (M)

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

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

As a starting point we incorporate observable regressors X in the outcome equation in a reasonably general way. The covariates X explain part of the residual term u in (1) and part of the idiosyncratic gains from treatment:

$$\begin{aligned} y_i^1 &= \beta + u(X_i) + \alpha(X_i) + [(u_i - u(X_i)) + (\alpha_i - \alpha(X_i))] \\ y_i^0 &= \beta + u(X_i) + (u_i - u(X_i)) \end{aligned} \tag{25}$$

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

To estimate the ATT, the matching method assumes that the set of observables, X , contain all the information about the potential outcome in the absence of treatment, y^0 , that was available to the individual at the point of deciding about whether to become treated, d . This means that the econometrician has all the relevant information, namely the information that simultaneously characterizes the participation rule and the non-treated outcome. This is called the Conditional

Independence Assumption (CIA) and can be formally stated as follows

$$y_i^0 \perp d_i \mid X_i \tag{26}$$

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

Within model (25), the CIA can be restated in terms of the unobservable in the non-treated outcome equation,

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

meaning that the unobservable component of the non-treated outcomes is independent of participation into treatment or, which is the same, that there is no selection on the unobservable part of u_i in (25).

The CIA in (26) obviously implies a conditional version of the randomization hypothesis (R1),

$$E[u_i|d_i, X_i] = E[u_i|X_i] \tag{27}$$

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

The implication of (26) or (27) is that treated and non-treated individuals are comparable in respect to the non-treated outcome, y^0 , conditional on X . Thus, for each treated observation (y^1) we can look for a non-treated (set of) observation(s) (y^0) with the same X -realization and be certain that such y^0 is a good predictor of the unobserved counterfactual.

Thus, matching is explicitly a process of re-building an experimental data set. Its ability to do so, however, depends on the availability of the counterfactual. That is, we need to ensure that each treated observation can be reproduced among the non-treated. This is only possible if the observables X do not predict participation exactly, leaving some room for unobserved factors to influence the treatment

status. This is the second matching assumption, required to ensure that the region of X represented among participants is also represented among non-participants. Formally, it can be stated as follows

$$P(d_i = 1 | X_i) < 1 \quad (28)$$

Given assumptions (27) and (28), we can now define the matching estimator. Let S represent the subspace of the distribution of X that is both represented among the treated and the control groups. S is known as the common support of X . Under (28), S is the whole domain of X . The ATT over the common support S is

$$\begin{aligned} \alpha^{ATT}(S) &= E[y^1 - y^0 | d = 1, X \in S] \\ &= \frac{\int_S E(y^1 - y^0 | X, d = 1) dF_{X|d}(X | d = 1)}{\int_S dF_{X|d}(X | d = 1)} \end{aligned}$$

where $F_{X|d}$ is the cumulative distribution function of X conditional on d and $\alpha^{ATT}(S)$ is the mean of impact on participants with observable characteristics X in S .

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

$$\hat{\alpha}^M = \sum_{i \in T} \left\{ y_i - \sum_{j \in C} \varpi_{ij} y_j \right\} \omega_i \quad (29)$$

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

Identification of ATE requires a strengthened version of assumption (27) because the correct counterfactual needs to be constructed for both the treated and the non-treated. This means that both $(u_i - u(X_i))$ and $(\alpha_i - \alpha(X_i))$ need to be (mean) independent from d conditional on X . That is, selection on unobserved expected gains must also be excluded for matching to identify the correct

ATE. In its weaker version, the CIA is now formally:

$$\begin{aligned} E[u_i|d_i, X_i] &= E[u_i|X_i] \\ E[\alpha_i|d_i, X_i] &= E[\alpha_i|X_i] \end{aligned} \tag{30}$$

Estimation of ATE also requires a modification of the overlapping support assumption (28) to ensure that both the treated and the non-treated are represented within the alternative group. Formally,

$$0 < P(d_i = 1 | X_i) < 1 \tag{31}$$

Under (30) and (31), the ATE over the common support S is

$$\begin{aligned} \alpha^{ATE}(S) &= E[y^1 - y^0 | X \in S] \\ &= \frac{\int_S E(y^1 - y^0 | X) dF_X(X)}{\int_S dF_X(X)} \end{aligned}$$

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

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

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

However, it is rarely clear what sort of information is in the information set at assignment. What is clear is that matching variables should be determined before the time of assignment and not after as this could compromise the CIA by having matching variables affected by the treatment

status. A structural model can shed some light on what the correct set of matching variables should be. However, such models are likely to include some unobservable variables and are more naturally used to motivate Instrumental Variables and Control Function methods described further below. Nevertheless, they can suggest possible variables that capture the key determinants of selection. For example, in studies about the impact of training on labor market outcomes, previous labor market history could contain all the relevant information on the unobservable ability and job-readiness as it is partly determined by such factors.

5.2 Propensity score matching

A serious limitation to the implementation of matching is the dimensionality of the space of the matching variables, X . Even if all variables are discrete with a finite domain, the dimensionality of the combined space increases exponentially with the number of variables in X , making it virtually impossible to find a match for each observation within a finite (even if large) sample when more than a few variables are being controlled for.

A popular alternative is to match on a function of X . Usually, this is carried out on the probability of participation given the set of characteristics X . Let $P(X)$ be such probability, known as the “propensity score”. It is defined as

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

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

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

The balancing property of the propensity score implies that, if $P(X)$ is known, it can be used to replace X in the matching procedure.¹³ But then, knowledge of $P(X)$ reduces the matching problem

¹³More 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.

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

When using propensity score matching, the comparison group for each treated individual is chosen with a pre-defined criteria (established in terms of a pre-defined metric) of proximity between the propensity scores for treated and controls. Having defined the neighborhood for each treated observation, the next step is that of choosing the appropriate weights to associate the selected set of non-treated observations for each participant. Several possibilities are commonly used. We briefly refer the most commonly applied alternatives and refer the interested reader to Leuven and Sianesi (2003) and Becker and Ichino (2002) for a more detailed practical guide to alternative matching procedures.

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

Kernel Matching defines a neighborhood for each treated observation and constructs the counterfactual using all control observations within the neighborhood, not only the closest observation. It assigns a positive weight to all observations within the neighborhood while the weight is zero otherwise. Different weighting schemes define different estimators. For example, uniform kernel attributes the same weight to each observation in the neighborhood while other forms of kernel make the weights dependent on the distance between the treated and the control being matched, where the weighting function is decreasing in distance. By using more observations per treated, Kernel matching reduces the variability of the estimator as compared to the Nearest Neighbor and produces less bias than

Nearest Neighbor with many matches per treated. However it still introduces significant bias at the edges of the distribution of $P(X)$. When this is a problem, *Local Linear Matching* will effectively deal with this sort of bias.¹⁴

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

The general form of the matching estimator is not altered by the sort of weights one decides to apply. As before, it is given by $\hat{\alpha}^M$ in (29).

While propensity score matching is affected by the same problems as fully non-parametric matching in choosing the right set of controlling variables, it also faces the additional problem of finding a sufficiently flexible specification for the propensity score to ensure that the distribution of observables is indeed the same among treated and matched controls. That is, one wants to ensure that if (27) holds then (26) also holds. The evaluation literature has proposed a few balancing tests to assess whether the specification for the propensity score is statistically sound. For example, Rosenbaum and Rubin (1985) propose a test based on the comparison of means for each covariate between treated and matched controls. If the difference in means is too large, the test rejects the hypothesis that the samples (of treated and matched controls) are balanced with respect to the covariates when they are balanced with respect to the propensity score.

¹⁴For a discussion of non-parametric estimators including Kernel and Local Linear Regression methods see, Heckman, Ichimura and Todd (1997).

5.2.1 The linear regression model and the matching estimator

The linear regression model is often seen as a matching estimator and also relies on selection on the observables. It amounts to impose a fully parametric structure to model (25) by assuming that u_i and α_i are linear functions of X :

$$\begin{aligned} u(X_i) &= X_i\mu \\ \alpha(X_i) &= \xi_0 + X_i\xi_1 \end{aligned}$$

where (μ, ξ_0, ξ_1) are the unknown coefficients. The model can then be written as

$$y_i^0 = \beta^0 + X_i\gamma^0 + e_i^0 \tag{32}$$

$$y_i^1 = \beta^1 + X_i\gamma^1 + e_i^1 \tag{33}$$

where

$$\begin{aligned} \beta^d &= \beta + d\xi_0 \\ \gamma^d &= \mu + d\xi_1 \\ e_i^d &= (u_i - X_i\mu) + d(\alpha_i - \xi_0 - X_i\xi_1) \end{aligned}$$

and d is the treatment indicator.

Estimation of the ATT requires knowledge of the model for the untreated outcomes, (32). Under the CIA and the assumption of exogeneity of the covariates X , that is $E(e^0|X) = E(e^0)$, the ATT can be estimated using OLS. The common support assumption (28) is not required as the parametric specification can be used to extrapolate y^0 outside the observable range of X when predicting the counterfactual for each treated observation.

The imposition of a parametric specification is not as restrictive as it might first seem. In fact, by including many interactions between the variables and higher order polynomials in the (continuous) regressors, one will closely approximate any smooth function y^0 (see the Blundell, Dearden and Sianesi

(2005) application, for example). The main requirement is then to use a flexible enough functional form for y^0 .

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

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

5.3 Weaknesses of matching

The main weaknesses of matching are data driven: its availability and our ability to select the right information. The common support assumption (28) ensures that the missing counterfactual can be constructed from the population of non-treated. What (28) 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 (26) (or (27)) and (28) 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 (28). 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 available set 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 right hand side 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 mis-weighting over the common support. Through the process of choosing and re-weighting observations, matching corrects for the latter two sources of bias and selection on the unobservables is assumed to be zero by the CIA.

5.4 Using matching to estimate the returns to education

In this section we return to education evaluation example. We assume that earnings change with region and adopt the earnings specification (22) which we reproduce here:

$$\ln y_i = \beta_0 + \beta_1 x_i + (\alpha_0 + \alpha_1 \theta_i) d_i + u_i$$

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

5.4.1 Monte-Carlo results

We ran some monte-carlo experiments under different assumptions about the relationship between d and u , the source of endogeneity in evaluation problems. We estimated both the ATT and the ATNT using different sets of conditioning variables. Table 3 details the results obtained using log earnings when an education subsidy is available.¹⁵

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

Notice that true effects in columns (1) and (4) change with the set of conditioning variables due to changes in the overlapping support. In our example, this is never a serious problem mainly because the state space being considered is small. Nevertheless, the more conditioning variables are included to perform matching, the more the identifiable effect differs from the population one displayed in row (1).

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

¹⁵Estimates in levels and under the no-subsidy scenario show similar patterns to the ones presented here and are available from the authors under request.

Table 3: Monte Carlo experiment - Matching estimates and bias in logs

		$corr(u, v) = 0$			$corr(u, v) < 0$		
		true effect	estimate	bias	true effect	estimate	bias
		(1)	(2)	(3)	(4)	(5)	(6)
<i>Panel A: Estimates of the ATT</i>							
(1)	Simple difference	0.474	0.494	0.043	0.453	0.991	1.188
Matching using conditioning variables:							
(2)	x (region)	0.474	0.469	0.009	0.453	0.969	1.140
(3)	z (family background)	0.474	0.494	0.042	0.453	1.065	1.353
(4)	s (test score)	0.444	0.518	0.167	0.444	1.210	1.726
(5)	θ (ability)	0.473	0.534	0.129	0.452	1.084	1.399
(6)	(x, z, s, θ)	0.455	0.487	0.070	0.442	1.256	1.840
<i>Panel B: Estimates of the ATNT</i>							
(7)	Simple difference	0.296	0.494	0.672	0.315	0.991	2.144
Matching estimates using conditioning variables:							
(8)	x (region)	0.296	0.470	0.590	0.315	0.970	2.080
(9)	z (family background)	0.295	0.585	0.984	0.315	1.163	2.691
(10)	s (test score)	0.294	0.332	0.131	0.315	1.153	2.656
(11)	θ (ability)	0.296	0.317	0.068	0.318	1.027	2.231
(12)	(x, θ)	0.296	0.290	0.020	0.317	1.004	2.168
(13)	(x, z, s, θ)	0.296	0.254	0.142	0.319	1.263	2.964

Notes: Simulated data based on 200 Monte-Carlo replications of 2000 observations each. All estimates obtained under the assumption that the true specification of the outcomes equation is additively separable in logs and that subsidized education places are available for those with high test scores. Columns (1) to (3) present results obtained for independent error terms, u and v . Columns (4) to (6) present results obtained for (negatively) correlated error terms, u and v , with a correlation coefficient of -0.5. Bias estimates result from the comparison of the average estimate with the true effect in column (1) and are measured in relative terms. ATT stands for “average treatment on the treated”. ATNT stands for “average treatment on the non-treated”. Estimates in rows (2)-(6) and (8) to (12) are based on propensity score matching using Epanechnikov kernel weights. Different matching variables are used in each row. Estimates in rows (1) and (7) are based on simple differences.

Matching on other characteristics will induce some bias. For instance, suppose we decide to match on the test score - row (4). Individuals with the same score may decide differently about education because they expect different gains from the investment. In our case, one of the determinants of

earnings and gains from education investment is region. Individuals in the high-returns region are more likely to participate, but they would also enjoy from higher earnings if remained uneducated then their counterparts in the low-returns region. This means that the distribution of education conditional on s differs across regions.¹⁶ Thus, when comparing educated and uneducated individuals conditional on s we are over-sampling treated from the high-returns region and non-treated from the low-returns/low-earnings region, leading to biased estimates.

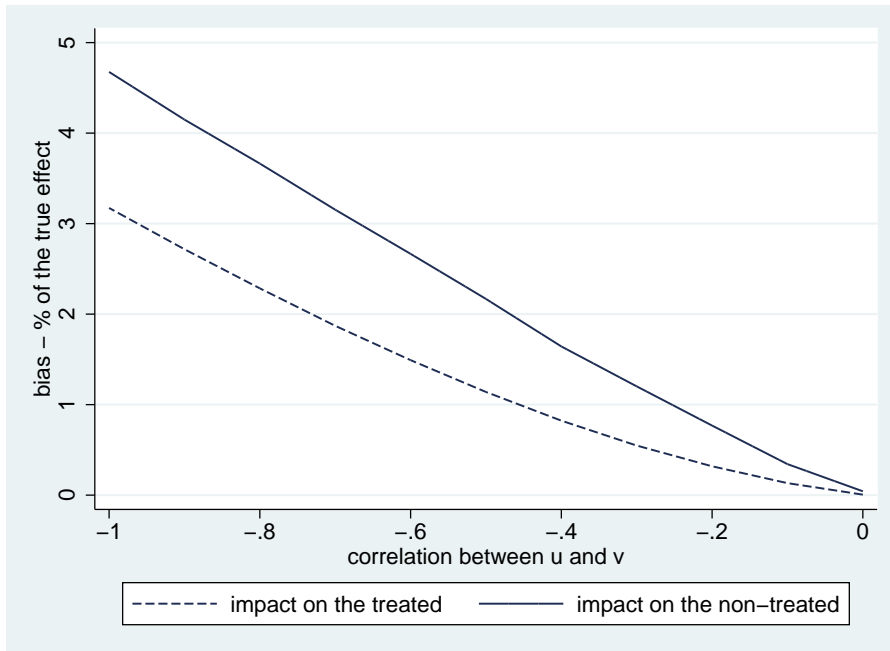
Rows (7)-(13) show how much more difficult it can be estimating ATNT than ATT. The conditional independence assumption underlying the matching estimation of the ATNT requires the treated outcome y^1 to be independent of the treatment status, d conditional on the covariates Z . This excludes selection on the non-treated outcome, y^0 , and on the gains from treatment, α . y_0 depends on region, x , and the gains from treatment depend on ability, θ , and region, x . Thus, the correct conditioning set is now (θ, x) and the first 3 columns of row 12 show the results that confirm this.

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

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

¹⁶It will also be the case that the distribution of ability conditional on s varies with region since individuals take into account potential earnings when deciding about effort. However, this will have no adverse consequences to the estimation of ATT because ability affects the potential gains only.

Figure 2: Relative bias in matching estimates by level of correlation between the unobservables u and v



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 forecast outcome as data is rarely rich enough to describe the relevant available information. However, the combination of matching with DID can accommodate unobserved determinants of the non-treated outcome affecting participation for as long as these are constant over time.

To discuss MDID, we start by decomposing the unobservable term u_i in (25) into a fixed effect (n), macro shock (m) and an idiosyncratic transitory shock (o). MDID can be applied when treated and non-treated are observed over time with at least one observation before and one after the treatment. For simplicity we consider two time periods, (t_0, t_1) , where $t_0 < k < t_1$ and k is the time of treatment. MDID compares the evolution of treated outcomes with that of non-treated over the observation period (t_0, t_1) and assigns any difference to the impact of treatment. To do so, MDID makes a

common trends assumption - had the treated remained non-treated and they would have experienced a change in outcomes equal to that observed among the actual non-treated.

More formally, the model can now be written as

$$\begin{aligned} y_{it}^1 &= \beta + u(X_i) + \alpha(X_i) + [(n_i + m_t + o_{it} - u(X_i)) + (\alpha_i - \alpha(X_i))] \\ y_{it}^0 &= \beta + u(X_i) + (n_i + m_t + o_{it} - u(X_i)) \end{aligned} \quad (34)$$

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

$$(u_{it_1} - u_{it_0}) \perp d_{it_1} \mid X_i \quad (35)$$

The main matching hypothesis is now stated in terms of the before-after evolution instead of in 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. We continue to consider time invariant covariates, X , even though MDID explicitly explores the time-series dimension of the data. The discussion on the choice of covariates at the end of section 5.1, where we argued that the appropriate covariates should reflect the information available to the individual at the time of making a participation decision, explains this choice.

Assumption (35) is not enough to ensure identification of ATT. Just as in the matching case, we also need to impose a common support hypothesis. This will be the same as (28) when longitudinal data is available. If we only have available repeated cross-section data we will need to strengthen it to ensure that the treated group can be reproduced in all three control groups characterized by treatment status before and after the program. This version of the common support assumption states that all treated individuals have a counterpart on the non-treated population before and after the treatment

$$P(d_{it_1} = 1 \mid X_i, t) < 1 \quad (36)$$

where $P(d_{it_1} = 1 | X_i, t)$ is the probability that an individual observed at time t with characteristics X_i would belong to the treatment group at time t_1 .

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

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

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 .¹⁷ In this case, the matching-DID estimator would be

$$\hat{\alpha}^{MDID,RCS} = \sum_{i \in T_1} \left\{ \left[y_{it_1} - \sum_{j \in T_0} \varpi_{ijt_0}^T y_{jt_0} \right] - \left[\sum_{j \in C_1} \varpi_{ijt_1}^C y_{jt_1} - \sum_{j \in C_0} \varpi_{ijt_0}^C y_{jt_0} \right] \right\} \omega_i$$

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

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

¹⁷As with the DID estimator, our ability to correctly separate treated from non-treated at t_0 is determinant for the quality of the estimates.

6 Instrumental Variables

6.1 The instrumental variables (IV) estimator

In this section we continue considering the model described by equations (2)-(4) with potential outcomes being partly explained by the observables X as in equation (25). The variables Z in the decision rule (4) may include all observables X in the outcome equation plus additional regressors. For simplicity, we implicitly condition on X and omit it from the discussion below. We also consider only one additional variable in Z , which we denote by z . Finally, we 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.

In contrast to the matching method, the method of Instrumental Variables (IV) deals directly with selection on the *unobservables*. The IV approach requires the existence of at least one regressor exclusive to the decision rule. In our notation, this is the variable z , which is known as the instrument. It affects participation only and so is not in X . This is known as the *exclusion restriction*. It implies that the potential outcomes do not vary with z and any difference in the mean observed outcomes of two groups differing only with respect to z can only be due to consequent differences in the participation rates and composition of the treatment group with respect to potential gains from treatment. When the treatment effect is homogeneous, so that $\alpha^{ATE} = \alpha^{ATT} = \alpha_i = \alpha$, only differences in participation rates subsist and these can be used together with resulting differences in mean outcomes to identify the impact of treatment.

To see this more clearly, we formalize these three assumptions below. The latter assumption states that the treatment effect is homogeneous:

$$\alpha_i = \alpha \text{ for all } i \tag{37}$$

The first two assumptions define the dependence of the outcome y and the participation status d on the instrument z . They can be stated as:

$$P[d = 1|z] \neq P[d = 1] \tag{38}$$

and

$$E[u|z] = E[u]. \quad (39)$$

Under conditions (37)-(39), the instrument z is the source of exogenous variation used to approximate randomization. It provides variation correlated with the participation decision only.

Under assumptions (37) and (39), the dependence of y on the instrument arises through the index (propensity score) $P(z) = P(d = 1|z)$ as follows:

$$\begin{aligned} E(y_i | z_i) &= \beta + \alpha P(z_i) + E(u_i|z_i) \\ &= \beta + \alpha P(z_i) + E(u_i) \\ &= E(y_i | P(z_i)). \end{aligned} \quad (40)$$

Assumption (38) then ensures that two different values of z exist that induce variation in $P(z)$ and allow for the identification of α . Let z^* and z^{**} be such values. Then

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

and the treatment effect is identified from the ratio:

$$\alpha = \frac{E(y_i | z_i = z^*) - E(y_i | z_i = z^{**})}{P(z^*) - P(z^{**})}. \quad (41)$$

This is the standard IV (or Wald) identification strategy. It is designed to explore discrete instruments or discrete changes in continuous instruments.

In the standard continuous instrument case it is more efficient to explore the whole variation in z . The IV conditions discussed above ensure that

$$\begin{aligned} \text{cov}(y, z) &= \alpha \text{cov}(x, z) + \text{cov}(u, z) \\ &= \alpha \text{cov}(x, z) \end{aligned}$$

and the IV estimator is

$$\hat{\alpha}^{IV} = \frac{\widehat{cov}(y, z)}{\widehat{cov}(d, z)}.$$

6.2 Weaknesses of IV

A key issue in the implementation of IV is the choice of the instrument. It is frequently very difficult to find an observable variable that satisfies assumption (39), in which case IV is of no practical use. This will happen when participation is mainly driven by the determinants of potential outcomes. In other cases, the instrument has insufficient variation or causes insufficient variation in the propensity score. Instruments with this characteristic are known as weak instruments. Although (38) may still hold with a weak instrument, the consequent (small) size of the denominator in (41) leads to very imprecise estimates of the treatment effect.

Identification using classical IV still relies on the additional homogeneity assumption in equation (37). If (37) does not hold, the exclusion restriction is also unlikely to hold. To see why, notice that the unobservable in the outcome equation is now

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

and the new exclusion restriction needs to be expressed in terms of e :

$$\begin{aligned} E[e_i|z] &= E[u_i|z] + P(z)E(\alpha_i - \alpha^{ATE}|d_i = 1, z) \\ &= E(e_i) \end{aligned}$$

But since z explains d , the second equality above is generally not satisfied.

The one exception occurs when there is no selection on the idiosyncratic gains. This means that the idiosyncratic gain, $\alpha_i - \alpha^{ATE}$, and the unobservable in the selection rule, v , are not related. In such case, $E(\alpha_i - \alpha^{ATE}|d_i = 1, z) = 0$ and $E(e_i|z) = E(e_i)$ under (39). Thus, classical IV will still identify ATE (and, which is the same, ATT) if individuals do not have or do not act on unobservable information related to (expected) gains to decide about treatment status.

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 will certainly make a more informed participation decision. The resulting selection process breaks the independence between α and z conditional on selection since both variables affect the selection process. A change in z will drive into treatment individuals with an expected return different from α^{ATE} .

To illustrate the problem, consider the education example we have been using. Assume that the returns to education are partly determined by the child's unobservable ability. Suppose the instrument is some measure of the cost of education (say distance to college) under the assumption that it is uncorrelated with the child's potential earnings and, therefore, ability (this is possibly a strong assumption but serves only our illustration purposes). However, the selection process will create a relationship between distance to college and returns to college education in the data. This is because individuals facing a relatively low cost of education (live closer to college) may be more likely to invest in college education, even if expecting comparatively small returns, than individuals facing higher education costs. Under our simplistic setup, this means that the distribution of ability among college graduates who live far from college is more concentrated on high ability levels than that for college graduates who live close to college. Such compositional differences will then affect the distribution of returns to college in the data for the two groups.

If the homogeneity assumption (37) fails to hold, IV will not generally identify ATE or ATT. This happens because the average outcomes of any two groups differing only on the particular z -realizations are different for two reasons: (i) different participation rates and (ii) compositional differences in the treated/non-treated groups with respect to the unobservables. The latter precludes identification of ATE or ATT. However, a different "local" average parameter can be identified under slightly modified hypothesis - the LATE parameter, 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 the effect is heterogeneous. The rationale is that, under certain conditions, a change in z reproduces random assignment locally by inducing individuals to

alter their participation status without affecting the potential outcomes, (y^0, y^1) . As with standard IV, the difference in average outcomes between two groups differing only in the realization of z results exclusively from the consequent difference in participation rates. Unlike standard IV, the identifiable effect will not correspond to the ATE or the ATT. Instead, it will depend on the particular values of z used to make the comparison and the identifiable effect is the average impact on individuals that change their participation status when faced with the change in z used to estimate the effect of treatment.

As with classical IV, the validity of an instrument z depends on whether it determines participation and can be excluded from the outcomes' equation except through its effect on participation. In an heterogeneous effect framework the exclusion condition requires that: *(i)* z has no joint variation with v , or otherwise changes in z would not separate changes in the participation rates unrelated to outcomes as simultaneous changes in v could be related with changes in the unobservable components of the potential outcomes, particularly gains from treatment; and *(ii)* z is unrelated to the unobserved determinants of potential outcomes.

The LATE assumptions can now be formally established. The first two assumptions are identical to the classical IV assumptions (38) and (39):

$$P[d = 1|z] \neq P[d = 1] \tag{42}$$

$$E[u|z] = E[u]. \tag{43}$$

However LATE requires stronger identification assumptions than standard IV to compensate for the relaxation of the homogeneity hypothesis. The additional assumption pertains the relationship between z and the remaining unobservables:¹⁸

$$(\alpha, v) \perp z \tag{44}$$

¹⁸A slightly stronger version of assumptions (43)-(44) is frequently imposed: $(u, \alpha, v) \perp z$ or, which is the same $(y^0, y^1, v) \perp z$

Under the notation of the selection rule in (3)-(4), assumption (44) ensures that

$$\begin{aligned} P(d_i = 1 | z) &= P(g(z, v) > 0 | z) \\ &= P(g(z, v) > 0) \end{aligned}$$

meaning that z is exogenous in the selection rule. Furthermore, joint independence of idiosyncratic gains and v from z guarantees that

$$E(\alpha_i | z, d_i = 1) = E(\alpha_i | g(z, v) > 0)$$

Taken together, the three LATE assumptions are sufficient to ensure that $P(z)$ contains all the information in z that explains y :

$$\begin{aligned} E(y_i | z) &= \beta + P(d_i = 1 | z)E(\alpha_i | z, d_i = 1) \\ &= \beta + P(g(z, v) > 0)E(\alpha_i | g(z, v) > 0) \end{aligned} \tag{45}$$

This means that all relevant information in z results from what can be inferred from z and d about the location of v , the unobservable in the selection rule that is correlated with the unobserved components of the outcomes and that is at the root of the selection problem.

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

$$\begin{aligned} &E[y_i | z = z^{**}] - E[y_i | z = z^*] \\ &= P(g(z^{**}, v) > 0)E(\alpha_i | g(z^{**}, v) > 0) - P(g(z^*, v) > 0)E(\alpha_i | g(z^*, v) > 0) \\ &= P(g(z^{**}, v_i) > 0, g(z^*, v_i) < 0)E(\alpha_i | g(z^{**}, v_i) > 0, g(z^*, v_i) < 0) - \\ &\quad P(g(z^{**}, v_i) < 0, g(z^*, v_i) > 0)E(\alpha_i | g(z^{**}, v_i) < 0, g(z^*, v_i) > 0) \end{aligned}$$

The intuition behind the above expression is that 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. The last equality shows

two treatment parameters that one may be willing to identify: the impact of treatment on the treated under z^{**} but not treated under z^* and the impact of treatment on the treated under z^* but not treated under z^{**} . In practice there are frequently strong arguments to eliminate one of the alternatives. For example, it may be the case that every participant at $z = z^*$ also participates at $z = z^{**}$ but not the reverse. This is the substance of the monotonicity assumption, the last of the LATE assumptions. Formally:

$$g(z, v) \text{ is a } \textit{monotonic} \text{ function of } z. \quad (46)$$

Suppose (46) holds. In particular, suppose g is increasing in z and $z^{**} > z^*$ so that $P(g(z^{**}, v_i) < 0, g(z^*, v_i) > 0) = P(z^*) - P(z^{**}) = 0$. In such case

$$E[y_i | z^{**}] - E[y_i | z^*] = P(g(z^{**}, v_i) > 0, g(z^*, v_i) < 0)E[\alpha_i | g(z^{**}, v_i) > 0, g(z^*, v_i) < 0].$$

and this equation can be rearranged to obtain the LATE parameter:

$$\begin{aligned} \alpha^{LATE}(z^*, z^{**}) &= E[\alpha_i | g(z^{**}, v_i) > 0, g(z^*, v_i) < 0] \\ &= \frac{E[y_i | z^{**}] - E[y_i | z^*]}{P[g(z^{**}, v_i) > 0, g(z^*, v_i) < 0]} \\ &= \frac{E[y_i | z^{**}] - E[y_i | z^*]}{P(z^{**}) - P(z^*)}. \end{aligned} \quad (47)$$

The first equality clarifies the meaning of the LATE parameter: it measures the impact of treatment on individuals that move from non-treated to treated when z changes from z^* to z^{**} .

The LATE approach can also be illustrated within our running example on education investment. As before, suppose z is a measure of cost, say distance to college, with participation assumed to become less likely as z increases. To estimate the effect of college education, consider a group of individuals that differ only in z . Among those that invest in further education when distance z equals z^* some would not do so if $z = z^{**}$ where $z^* < z^{**}$. In this case, LATE measures the impact of college education on the “movers” by assigning any difference on the average outcomes of the two groups to

the different enrollment rates caused by the difference in the cost of investing.¹⁹

6.3.1 The LATE assumptions

The independence assumptions (43)-(44) are required to establish the result (45) on which derivation of LATE hinges. It states that z is independent of the unobservable components in the outcomes and participation rules, namely u_i (mean independent), $(\alpha_i - \alpha^{ATE})$ and v_i . This means that z should not affect the observed outcomes through any effect on the potential outcomes or any relation with the unobserved components of the model. While the former is easy to understand the later requires some explanation. Suppose z is related with v in the participation rule and v is related with u in the outcome equation. Then the potential outcome will generally be related with z .

The education example can be used again to illustrate the conditions under which the independence assumption may not apply. As before, suppose z is a measure of the cost of education, say distance to college. A family that values education may take some steps to facilitate the investment, for example by taking it into account when deciding about residence. Such family may also be particularly interested in encouraging learning and developing a taste for it in the children. So children raised in such environment may benefit both from lower education costs *and* higher taste for education. The later is unobservable, included in v , and is likely to be related with the future taste for working, also unobservable and included in u . In this case, although z has no direct impact on potential outcomes, the selection of home location will create a dependence between z and the potential outcomes arising through a dependence between v and u . That is, if z is not exogenous in the participation equation then two groups with different realizations of z will represent two different populations in terms of the distribution of v and, thus, two different populations in terms of the distribution of u (a similar argument could be constructed in relation to α). In such case, even if one could observe both potential outcomes they would not be independent of z in the data.

The monotonicity assumption 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

¹⁹Abadie, Angrist and Imbens (2002) extend this approach to the evaluation of *quantile treatment effects*. The goal is to assess how different parts of the outcome's distribution are affected by the policy. As with LATE, a local IV procedure is used, making the estimated impacts representative only for the sub-population of individuals changing their treatment status in response to the particular change in the instrument being considered.

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 is due to individuals moving *in and out* of participation. However, it would not be 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).

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

6.3.2 What does LATE measure?

Although analytically very similar to the IV estimator in (41), 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. Whether the parameter is of policy interest or not depends on the instrument and the specific values of the instrument used in the estimation (see, for example, the discussion in Heckman, Lalonde and Smith, 1999). When a discrete variable, namely a change in policy, is used to instrument participation, LATE will measure the effect of treatment on individuals changing their treatment status in response to the policy change. In this case, LATE focus on an important subpopulation and may provide an important measure of the impact of the policy. If, on the other hand, a continuous variable measuring some individual characteristic is used to instrument participation, LATE will generally be much less informative.

In our education example, notice that we discussed two alternative instruments to measure a local effect. In the first case, in the context of DID, we used a change in policy to measure the impact of education on individuals moving into education. DID differs from the standard LATE estimator based on a change in policy only by allowing the aggregate conditions to vary over time (although it requires treated and controls to be similarly affected by the market conditions). In the second case, we discussed the use of family background or cost of education to instrument participation. Clearly, the former is much more informative for the policy maker than the later. The estimated

parameter based on our continuous variable will depend on the specific values being compared, may not represent a specific population that can be easily targeted and is more likely to raise arguments about the validity of the instrument (just as illustrated by the discussion in the previous section).

6.4 The Marginal Treatment Effect

Heckman and Vytlacil (1999, 2001, 2006) and Carneiro, Heckman and Vytlacil (2005) reinterpret the local IV methods and the local treatment effect parameters within a selection model. These authors consider the estimation of the impact of treatment over the whole distribution of a continuous instrument. To do so, they use infinitesimal changes in the participation probabilities to measure the limit of LATE as the change in these probabilities becomes arbitrarily small. As the whole distribution of local treatment effects is possibly identified, all more aggregate parameters can also be estimated by integration over the distribution of the probability of participation.

These authors consider a version of the selection model (3)-(4) which assumes additive separability of the unobservable, v . As before, selection follows a latent variable specification where

$$d_i^* = g(z_i) - v_i$$

and

$$d_i = \begin{cases} 1 & \text{if } v_i \leq g(z_i) \\ 0 & \text{otherwise} \end{cases}$$

The propensity score as a function of z is,

$$\begin{aligned} P(z_i) &= P[d_i = 1 | z_i] \\ &= P[v_i \leq g(z_i)] \\ &= F_v(g(z_i)) \end{aligned}$$

where F_v is the distribution function of the unobservable v . Heckman and Vytlacil (1999) use a general alternative representation of the above preferences when v is an absolutely continuous random variable

(meaning that v has no mass points). This is obtained by transforming the selection rule

$$v_i \leq g(z_i)$$

by a monotonically increasing function such as F_v to yield:

$$\begin{aligned} F_v(v_i) &\leq F_v(g(z_i)) \text{ which can be written as} \\ \mathbf{v}_i &\leq P(z_i). \end{aligned}$$

Given continuity of v , the transformed unobservable \mathbf{v} will follow a uniform distribution in $[0, 1]$ and the data-equivalent selection model is

$$\mathfrak{d}_i^* = P(z_i) - \mathbf{v}_i \tag{48}$$

with

$$d_i = \begin{cases} 1 & \text{if } \mathbf{v}_i \leq P(z_i) \\ 0 & \text{otherwise.} \end{cases} \tag{49}$$

The advantage of the latter representation is the connection between the propensity score and the newly defined unobservable \mathbf{v} : an individual with characteristic z is indifferent about participation when drawing $\mathbf{v} = P(z)$ and will participate with probability $P(z)$ or when drawing $\mathbf{v} \leq P(z)$.

Using this representation of the decision process, the marginal treatment effect (MTE) parameter at a point \mathbf{v}^* of the distribution of the unobservable can now be written as

$$\alpha^{MTE}(\mathbf{v}^*) = E(\alpha | \mathbf{v} = \mathbf{v}^*).$$

This parameter measures the impact of treatment on individuals with unobservable characteristics \mathbf{v}^* affecting the decision process. Under the LATE assumptions (43)-(44), which is now imported to the analysis of identification and estimation of the MTE parameter, notice that the instrument z does not bring further information about the expected gains from treatment after conditioning for \mathbf{v} . This is because \mathbf{v} contains all the information in d that may be related with the potential outcomes. So, individuals with the same \mathbf{v} but experiencing different values of z expect to gain the same from

treatment. Thus,

$$\alpha^{MTE}(\mathbf{v}^*) = E(\alpha|\mathbf{v} = \mathbf{v}^*, z)$$

for any possible value of z . In particular, one may try to evaluate the MTE at the point(s) in the distribution of z where individuals are indifferent about participation under the assumption that $P(z)$ is a non-trivial function of z :

$$\alpha^{MTE}(\mathbf{v}^*) = E(\alpha|\mathbf{v} = \mathbf{v}^*, \mathbf{v}^* = P(z)).$$

This specification explains the alternative definition of MTE that is more commonly encountered, namely the average effect of treatment on individuals just indifferent about participation at $P(z)$. (See also the important early work by Bjorklund and Moffitt (1987) and the recent nonparametric application by Moffitt (2007)). It is this definition that is explored to identify MTE.

Assumptions (43)-(44) together with the additive separability of v can be used to show:

$$\begin{aligned} E(y|z) &= \beta + P(z)E(\alpha|z, d = 1) \\ &= \beta + P(z)E(\alpha|\mathbf{v} < P(z)) \\ &= E(y|P(z)) \end{aligned}$$

Using this formulation, the LATE parameter can be expressed as:

$$\begin{aligned} \alpha^{LATE}(z^*, z^{**}) &= \frac{E(y|z^{**}) - E(y|z^*)}{P(z^{**}) - P(z^*)} \\ &= \frac{E(y|P(z^{**})) - E(y|P(z^*))}{P(z^{**}) - P(z^*)} \\ &= \alpha^{LATE}(P(z^*), P(z^{**})) \end{aligned}$$

Within this framework, LATE measures the impact of treatment on individuals with unobservable characteristics \mathbf{v} in the interval $[P(z^*), P(z^{**})]$. Again, this parameter does not change with the particular values of z selected for as long as \mathbf{v} remains on the interval $[P(z^*), P(z^{**})]$.

The MTE can be defined from LATE by considering an arbitrarily small interval in \mathbf{v} . The

limit can also be taken on the estimator of LATE to define the estimator of the MTE. Notice that the definition of LATE in (47) determines the LATE estimator by identifying the movers using the instrument. This will not be independent of z because it relies on the specific values of z chosen to assess the impact, which will determine the specific population of movers used in estimating the impact (or the margin to which the parameter corresponds to).

The Local Instrumental Variables (LIV) is precisely an estimator of the MTE obtained by taking the limit of the LATE estimator (47) as $P(z^*)$ becomes arbitrarily close to $P(z^{**})$:

$$\alpha^{LIV}(P(z)) = \frac{\partial E(y|\mathbf{v} = P(z))}{\partial P(z)}$$

This estimator is also dependent on z in the sense that, for a particular chosen value of the instrument, the specific margin at which MTE is being estimated is the specific value of the unobservable at the indifference point, namely $\mathbf{v} = P(z)$. Comparatively to LATE, however, the use of MTE is usually associated with the intention to recover the full distribution of treatment effects. MTE uses a continuous instrument to recover the whole (or an interesting part) of the distribution of participation probabilities from 0 to 1 for as long as all individuals have strictly positive probabilities of being treated and non-treated.

If data is rich enough to explore changes in treatment status over the whole distribution of \mathbf{v} then all the average parameters, namely ATE, ATT, ATNT and LATE, can be expressed as averages of MTE using different weights (see appendix B for details). For example, the estimation of ATT using MTE with a continuous instrument z requires the space of \mathbf{v} , $[0, 1]$, to be finely discretized using the distribution of $P(z)$. Estimation may use some non-parametric regression procedure to identify the slope of y with respect to $P(z)$ at each of the points on the grid - say a Local Quadratic Regression. This is the MTE at each point \mathbf{v} . The ATT among individuals with a probability of participation equal to p , $\alpha^{ATT}(p)$, may then be obtained by integrating the MTE's over the space of \mathbf{v} up to p - these are the participants among those with a probability of participation equal to p . The overall ATT may now be obtained by integrating $\alpha^{ATT}(p)$ over the whole distribution of p (see Carneiro and Lee, 2007, or Carneiro Heckman and Vytlacil, 2005, for details on the implementation procedure).

However, data may not be rich enough to allow for the estimation of MTE over the whole distri-

bution of \mathbf{v} , in which case LATE may be the best available option. This is clearly the case when the instrument is binary as, for example, a specific change in policy, in which case LATE is particularly suited to identify a parameter of interest.

6.5 Using IV to estimate the returns to education

Under the IV conditions, the variables in the selection process which do not enter 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 or distance to college (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 estimates for the two scenarios depending on the availability (panel B) or not (panel A) of an education subsidy and consider both uncorrelated (columns 1 and 2) and negatively correlated (columns 3 and 4) 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 (37) 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) and (2) show, as expected, that z is a valid instrument while s always induces significant bias when the disturbances in the cost of education (v) and earnings (u) equations are uncorrelated.

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, notice that the local IV technique estimates the ATT by integrating the MTE over the population of participants (see appendix B). The MTE at \mathbf{v} measures the impact of treatment on individuals that draw this specific value for the unobservable component in the selection rule. Because this unobservable contains all information about potential outcomes in the selection rule, changing z conditional on \mathbf{v} will not change (expected) gains. So the MTE can, in particular, be

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

		corr(u, v) = 0		corr(u, v) < 0	
		estimate	bias	estimate	bias
		(1)	(2)	(3)	(4)
<i>Panel A: No subsidy</i>					
(1)	True parameters	0.459		0.434	
IV estimates using the instruments:					
(2)	z (family background)	0.421	0.083	0.392	0.097
(3)	s (test score)	0.654	0.425	0.652	0.502
<i>Panel B: Positive subsidy</i>					
(4)	True parameters	0.471		0.453	
IV estimates using the instruments:					
(5)	z (family background)	0.404	0.120	0.417	0.039
(6)	s (test score)	0.537	0.170	0.583	0.343

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”.

interpreted as the impact on individuals with observable characteristics z that make them indifferent about participation at \mathbf{v} . These are the individuals that draw $\mathbf{v} = P(z)$.

Estimation of the MTE relies on these movers to feed a (local) IV estimator. However, the ATT cannot be recovered if $P(z)$ is not observed to vary over the whole unit interval. More precisely, the identification of the ATT will be affected in the absence of observations for $P(z)$ in $[0, \underline{p}]$ for some \underline{p} significantly larger than zero. In this case, we know that individuals experiencing $\mathbf{v} < \underline{p}$ will always prefer to participate within the observable range of $P(z)$. But then we never observe these individuals at their indifference point between participation and non-participation. Unfortunately, these individuals are unlikely to be a random sample of the population: they prefer to participate even at low levels of $P(z)$, which may indicate they expect to earn more from participation than most of the population. Thus, the estimated effect will not be the ATT but the average treatment for

Table 5: Monte Carlo experiment - Local IV estimates of ATT and respective bias in logs

		$\text{corr}(u, v) = 0$		$\text{corr}(u, v) < 0$	
		estimate	bias	estimate	bias
		(1)	(2)	(3)	(4)
<i>Panel A: No subsidy</i>					
(1)	True parameters	0.459		0.434	
IV estimates using the instruments:					
(2)	z (family background)	0.491	0.070	0.381	0.122
(3)	s (test score)	0.676	0.473	0.731	0.684
<i>Panel B: Positive subsidy</i>					
(4)	True parameters	0.471		0.453	
IV estimates using the instruments:					
(5)	z (family background)	0.484	0.028	0.384	0.152
(6)	s (test score)	0.401	0.147	0.382	0.157

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”.

individuals indifferent between participation and non-participation at the values of \mathbf{v} in the observable interval, $[\underline{p} > 0, \bar{p}]$.

The lack of support affects the results in columns (3) and (4), when the disturbances are correlated, where only values of $P(z)$ above 0.06 are observable. The expected outcome is that the obtained results are downward biased estimates of the ATT. In the uncorrelated disturbance case, however, the range of observable $P(z)$ starts very close to zero. In this case, displayed in columns (1) and (2), we are able to identify the impact of education even among individuals that show a strong preference towards education.²⁰

²⁰Not observing the top of the distribution of $P(z)$ does not affect the identification of ATT since agents with $\mathbf{v} > \bar{p}$ will never participate for the range of $P(z)$ observable. They are always non-participants and for as long as this is also

The above discussion is closely related to the literature on the ability of IV to produce interpretable parameters (see Heckman, 1997, Heckman and Vytlacil, 1998). The local parameters estimated by IV depend on the ability of the used instrument to induce a change of treatment status in each individual. Even the estimation based on the MTE, which explicitly attempts to run over the whole distribution of \mathbf{v} and uses the correct weights to aggregate the local parameters, may produce estimates that are not global. Instead, such estimates may be dependent on the particular instrument being used and apply only to the subpopulation of individuals that would switch treatment status at observable values of the instrument. Whether or not the identified parameter is of interest depends on the specific policy/evaluation question.

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 individual's response to the introduction of an education subsidy. Contrary to the no subsidy scenario, many individuals will make a positive effort to score better on the test if a subsidy is available. Such effort is related to z . Thus, the relationship between s and θ will be reduced while z will become related with the test score due to the endogenous effort decision. 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.

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 recent attention occurs when the probability of enrollment into treatment changes discontinuously with some continuous variable z . The variable z is an observable instrument, typically used to determine eligibility. It is, therefore, in matrix Z in the selection model (3). The discontinuity design (DD)

true in the population (as it happens to be the case in our example) they just do not belong to the population of interest for the evaluation of the ATT.

approach uses the discontinuous dependence of d on z to identify a local average treatment effect even when the instrument does not satisfy the IV assumptions discussed before. Instead of some exclusion or independence assumption like (39) and (44), DD relies on a continuous relationship between the instrument z and all the determinants of the outcome except participation in treatment. Any discontinuity in y as a function of z is, therefore, attributed to a discontinuous change in the participation rate as a function of z . As will be discussed, the parameter identified by DD is a local average treatment effect like the LATE parameter discussed under IV but is not necessarily the same parameter.²¹

As before, we assume participation d is determined by z and the unobservables v in a completely flexible way: $d = \mathbf{1}(g(z, v) > 0)$. The dependence of d on z means that the participation probability changes with z . The main source of identification used by DD is a discontinuity in such probability at a given point in the distribution of z . The discontinuity may be *sharp* or *fuzzy*, depending on whether participation is a deterministic function of z or not. We now discuss these two cases.

7.1.1 The sharp design

The most popular case, although empirically less frequent, is what is known by *sharp design*. This occurs when z fully determines participation on the basis of a threshold, z^* . The treated (non-treated) are individuals with values of z , say, above (below) the threshold. In this case, participation status changes at z^* for all individuals, from being deterministically equal to 0 to being deterministically equal to 1. Thus the probability of participation changes discontinuously at z^* from 0 to 1. The identification condition with sharp design can be stated as follows,

$$\begin{aligned} \lim_{z \rightarrow z^{*-}} P(d = 1|z) &= P(z^{*-}) = 0 \\ \lim_{z \rightarrow z^{*+}} P(d = 1|z) &= P(z^{*+}) = 1 \end{aligned} \tag{50}$$

where, to simplify the notation, $P(z^{*-})$ ($P(z^{*+})$) represents the limit of the propensity score ($P(d = 1|z) = P(z)$) as z approaches z^* from below (above). Both limits are assumed to exist.

The fact that participation is locally a deterministic function of z means that individuals do not

²¹For an insightful discussion of DD see Hahn, Todd and Van der Klaauw, 2001; more recently, Imbens and Lemieux, 2007, provide a detailed discussion of DD and implementation issues.

contribute to the decision process.²² The sharp design implies that the decision process is exogenously determined by z and all the selection is on the observables. Thus, the impact of treatment is probably independent from the selection process, at least locally. Although selection occurs only on the observables, matching is not feasible given the absence of overlap between treated and controls once z is included in the set of covariates. Instead of the common support assumption used in matching, DD is based on the additional hypothesis of continuity of the remaining determinants of outcomes as functions of z at z^* . Under a sharp design, all that is required is that y^0 is continuous at z^* to guarantee that the information from the untreated side of the threshold constitutes the correct counterfactual for the treated side of the threshold. Within our model of outcomes, (1)-(2), this is equivalent to the condition

$$E(u_i|z^{*+}) = E(u_i|z^{*-}) \quad (51)$$

where $E(u|z^{*+})$ and $E(u|z^{*-})$ are the limits of $E(u|z)$ when z approaches z^* from above and below, respectively.

Under assumptions (50) and (51), any observed discontinuity in y at z^* results exclusively from the discontinuity in the participation rate. The DD parameter is in this case:

$$\alpha^{DD}(z^*) = E(y_i|z^{*+}) - E(y_i|z^{*-}).$$

where $E(y|z^{*+})$ and $E(y|z^{*-})$ are the limits of $E(y|z)$ when z approaches z^* from above and below, respectively. $\alpha^{DD}(z^*)$ measures the impact of treatment on a randomly selected individual with observable characteristics z just above z^* :

$$\alpha^{DD}(z^*) = E(\alpha|z^{*+}).$$

A continuity assumption on α similar to (51) will allow for a more natural and precise interpretation

²²The possibility that individuals adjust z in response to the eligibility criteria in the intent of changing their participation status is ruled-out from the DD analysis.

of $\alpha^{DD}(z^*)$. We state it as:

$$E(\alpha_i|z^{*+}) = E(\alpha_i|z^{*-}) \quad (52)$$

Under (51) and (52), $\alpha^{DD}(z^*)$ identifies the impact of treatment on a randomly selected individual with observable characteristics $z = z^*$. That is:

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

There are a few examples of economic studies that fall in the category of sharp design. They typically involve some exogenously imposed eligibility rule with a cut-off point. One example is the New Deal evaluation discussed above. Among other things, eligibility is based on age. Eligibles are those individuals that have not completed 25 years of age when reaching 6 months in unemployment. This rule has been used to estimate the impact of the New Deal on the oldest participants (see De Giorgi (2005)).

7.1.2 The fuzzy design

Possibly more common in economics is a *fuzzy design*. It refers to the situation in which the conditional probability of participation, $P(d = 1|z)$, is discontinuous at z^* . A fuzzy design occurs when dimensions other than z , particularly unobserved dimensions, also determine participation. To illustrate a possible fuzzy design consider our education example and suppose a subsidy is available for individuals scoring above a certain threshold in a test. The university intake will include both subsidized and unsubsidized individuals. However, the threshold-rule is expected to create a discontinuity in the probability of enrollment given the discontinuous change in the cost of education at the threshold.

One case that is empirically relevant is that of a treatment only available but not mandatory on one side of the threshold (say z^{*+}). This is the case, for example, of the Swedish Youth Practice, a subsidized employment program available for unemployed individuals under the age of 25.²³ Participation is not compulsory among eligibles but is not possible for anyone aged 25 or above. This

²³See Larsson, 2003, for further details on this program.

case turns out to be identical to the sharp design in terms of necessary identification assumptions. Together with this special discontinuity design, assumption (51) is enough to identify the impact of treatment of participants just above the threshold using DD. Formally:²⁴

$$\begin{aligned}\alpha^{DD}(z^*) &= E(\alpha_i | d_i = 1, z^{*+}) \\ &= \frac{E(y_i | z^{*+}) - E(y_i | z^{*-})}{P(d_i = 1 | z^{*+})}.\end{aligned}$$

Under the additional continuity assumption (52), the DD parameter is more naturally interpreted as the impact of treatment on participants at the margin:

$$\alpha^{DD}(z^*) = E(\alpha_i | d_i = 1, z^*)$$

In the general case where participation and non-participation occur on both sides of the threshold, DD is based on a stronger set of assumptions. To start with, notice that one implication of fuzzy design is that $d(z)$ does not change deterministically anymore. Thus, assumption (50) needs to be adjusted accordingly:

$$P(d = 1 | z^{*+}) \neq P(d = 1 | z^{*-}) \quad (53)$$

Just as in the sharp design case, identification of the treatment effect parameter requires continuity of the remaining determinants of the outcomes. But since treated and non-treated exist on both sides of the threshold in a general fuzzy design, continuity of the outcomes is now required for both y^0 and y^1 . Thus, both assumptions (51) and (52) are required. Under (51) to (53), any discontinuity of y at z^* can only be linked to the discontinuous change in the participation rate at that point.

Unlike the sharp or special fuzzy designs discussed above, the general fuzzy design requires an additional (local) independence assumption. For any z in a small neighborhood of z^* (say $z \in [z^{*-}, z^{*+}]$), the independence assumption can be written as:

$$E(\alpha_i | d, z) = E(\alpha_i | z) \quad (54)$$

²⁴See Battistin and Rettore, 2007, for further detail.

This assumption is less credible here than it would be in the sharp design case since unobserved factors are expected to determine participation along with z , and such factors may be related with the potential gains from treatment when gains are heterogeneous.

Under assumption (54), the conditional mean outcome y at a point z close to z^* (specifically, $z \in [z^{*-}, z^{*+}]$) can be written as:

$$E(y_i|z) = \beta + E(\alpha_i|z)P(d_i = 1|z) + E(u_i|z)$$

and the additional (dis)continuity assumptions (51) to (53) imply

$$\alpha^{DD}(z^*) = \frac{E(y_i|z^{*+}) - E(y_i|z^{*-})}{P(d_i = 1|z^{*+}) - P(d_i = 1|z^{*-})} \quad (55)$$

As before, $\alpha^{DD}(z^*)$ is the local average treatment effect, $E(\alpha_i|z = z^*)$. It measures the mean impact of treatment on a randomly selected individual with characteristic $z = z^*$. This is an average treatment effect at the local level since selection on idiosyncratic gains is locally excluded.

The local continuity and independence assumptions recover randomization under discontinuity in the odds of participation at the discontinuity point. The independence assumption is precisely a local version of (R2), meaning that ATE is identifiable locally by DD. Note also that, under the independence assumption (54), 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.

The DD estimator is the sample analog of (55):

$$\hat{\alpha}^{DD}(z^*) = \frac{\bar{y}^+ - \bar{y}^-}{\hat{p}^+ - \hat{p}^-} \quad (56)$$

where \bar{y}^+ and \bar{y}^- are sample averages of the outcomes at each side of the threshold and \hat{p}^+ and \hat{p}^- are estimators of the participation probability at each side of the threshold.

A non-parametric version of DD is simple to implement. It only requires running non-parametric regressions of y and d on z at each side of the discontinuity point. The expected value of y and d at z^*

can then be predicted at each side of the discontinuity and used to estimate the impact of treatment using (56).

7.2 The link between discontinuity design and IV

Interestingly, we have discussed the average treatment effect at a local level before, under IV. This was the LATE parameter or, when taking the limits using a continuous instrument, the MTE. To understand the similarities and differences between DD and local IV we consider the fuzzy design case and notice that both methods will identify the same parameter in a sharp design framework, namely the mean effect of treatment on a randomly selected individual among the treated close to the eligibility cutoff point.

The fuzzy design case is slightly more complex. DD relies on continuity and the local independence assumption in equation (54). The latter determines the parameter identified by DD as being the average impact of treatment on a randomly selected individual with a value of z at the threshold.

On its turn, LATE relies on the independence assumption (??) and on the monotonicity assumption (46). Under these conditions, LATE identifies the average impact of treatment on a randomly selected individual from the group of agents that change participation status as the value of the instrument changes from α^{*-} to α^{*+} .

One could think of replacing (54) by a monotonicity assumption like (46) to identify the mean effect on the movers when (54) raises suspicions. The monotonicity assumption requires that $d_i(z^{*+}) \geq d_i(z^{*-})$ for all individuals i . Together with the continuity assumption (51), monotonicity implies that the difference in mean outcomes is:

$$\begin{aligned} E(y_i | z^{*+}) - E(y_i | z^{*-}) &= E(\alpha_i | z^{*+}, d_i(z^{*+}) = 1, d_i(z^{*-}) = 0) P(d_i(z^{*+}) = 1, d_i(z^{*-}) = 0 | z^{*+}) + \\ &E(\alpha_i | z^{*+}, d_i(z^{*+}) = 1, d_i(z^{*-}) = 1) P(d_i(z^{*+}) = 1, d_i(z^{*-}) = 1 | z^{*+}) - \\ &E(\alpha_i | z^{*-}, d_i(z^{*-}) = 1) P(d_i(z^{*-}) = 1 | z^{*-}) \end{aligned}$$

The first term in the above equation measures the mean impact of treatment among individuals above the threshold that move into participation when z changes from just below to just above the threshold weighted by the probability of being a mover in the z^{*+} population; the second term measures the

mean impact of treatment among individuals above the threshold that would have participated even if below the threshold weighted by the probability of being one such individual in the z^{*+} population; finally, the third term measures the mean impact of treatment among participants below the threshold weighted by the probability of being one of these.

Given the above expression, one could think of measuring the impact of treatment on the movers when the second and third terms cancel each other out. Under monotonicity, this would require

$$\begin{aligned} P\left(d_i(z^{*-}) = 1 \mid z^{*+}\right) &= P\left(d_i(z^{*-}) = 1 \mid z^{*-}\right) \\ E\left(\alpha_i \mid z^{*+}, d_i(z^{*-}) = 1\right) &= E\left(\alpha_i \mid z^{*-}, d_i(z^{*-}) = 1\right) \end{aligned}$$

which would be guaranteed under the LATE independence assumption:

$$(\alpha, d(z)) \perp z \text{ in a neighborhood of } z^*$$

Thus, the interpretation of the DD parameter (55) depends on the set of assumptions one is willing to accept. It could be either the average effect for a randomly selected individual with characteristics z^* (DD) or the average effect for a randomly selected individual from the group of compliers in a neighborhood of z^* (those that move into treatment as z changes from just below to just above the cutoff point).

7.3 Weaknesses of discontinuity design

An obvious drawback of discontinuity design is its dependence on discontinuous changes in the odds of participation. In general this implies that only a local average parameter is identifiable. As in the binary instrument case of local IV, the discontinuity design is restricted to the discontinuity point which is dictated by the design of the policy. As discussed before under LATE with continuous instruments, the interpretation of the identified parameter can be a problem whenever the treatment effect, α , 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 \underline{s} in a given test. The subsidy creates a discontinuity in the cost of education at the threshold and, therefore, a discontinuity in the participation rates. On the other hand, the test score, s , and the returns to education, α , are expected to be (positively) correlated if both depend on, say, ability. But then, the local analysis will 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, at best the returns to education are estimated at a certain margin and other more general parameters cannot be inferred.

However, we could also suspect that neither the DD nor the LATE assumptions hold in this example. The former requires local independence of the participation decision from the potential gains conditional on the test score. But at any given level of the test score there is a non-degenerate distribution of ability levels. If higher ability individuals expect to gain more from treatment and are, for this reason, more likely to participate, then assumption (54) cannot be supported. The latter requires local independence of, in particular, $d(s)$ from s . But again, if both s and α depend on ability and individuals use information on expected gains to decide about participation, then s will not be exogenous in the selection rule. However, while this may be a serious problem to the use of LATE more generally, the infinitesimal changes considered here may reduce its severity when applied in the context of a DD framework.

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

7.4 Using discontinuity design to estimate the returns to education

Returning to our education returns example, estimation using the discontinuity design method is only possible when the educational subsidy is available and the score eligibility rule is used for assignment.

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

As expected, this method performs well at the local level when combined with relatively small

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

		True parameter			Estimated effects			
		global	local	local effect	bandwidth=0.5		bandwidth=1.5	
		ATE	ATE	on movers	estimate	bias	estimate	bias
<i>Panel A: Estimates in levels</i>								
(1)	$\text{corr}(u, v) = 0$	1.676	2.244	2.095	2.057	0.018	2.338	0.116
(2)	$\text{corr}(u, v) < 0$	1.673	2.801	1.948	2.388	0.226	2.588	0.329
<i>Panel B: Estimates in logs</i>								
(3)	$\text{corr}(u, v) = 0$	0.354	0.462	0.449	0.420	0.065	0.472	0.051
(4)	$\text{corr}(u, v) < 0$	0.354	0.485	0.470	0.477	0.015	0.542	0.153

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. Estimates were performed using alternative values for the bandwidth. The table presents estimates for a bandwidth of 0.5 (columns 4 and 5) and 1.5 (columns 6 and 7). The true local ATE represents the impact of education for agents scoring between 3.99 and 4.01. The true local effects on movers represents the impact on agents scoring between 4.0 and 4.1 when the subsidy becomes available and changing their education decision as a consequence of the subsidy. This is the parameter against which the estimates are contrasted. 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 presents estimates in levels. Panel B presents estimates in logs. ATE stands for “average treatment effect”.

bandwidth. Increasing the bandwidth increases the bias of the estimates. This is independent from the outcome being measured in levels or logs and v and u being correlated or not. The robustness of DD is to be expected from the weak set assumptions on which it is based.

Despite the relatively good performance of DD in this example, 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 were they eligible to a subsidy but not otherwise. Like the LATE parameter, the DD is especially well suited to address the questions related with potential extensions of the policy to a wider group by partially relaxing the eligibility rules.

The higher the uncertainty about gains from treatment at the moment of deciding about par-

ticipation, the further away the DD parameter (the local effect on the movers in column 3) will be to the local ATE (in column 2). In our example, individuals make a more informed participation decision under correlated error terms, v and u . In this case, 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 and the aggregate parameters will also display sharper differences. In this model, this is true in levels but not in logs, the reason being that the effect of treatment in log earnings does not depend on the shock u as it is additively separable.

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 (3)-(4) explicitly into consideration in the estimation process (see Heckman, 1976). The control function method does exactly this, treating the endogeneity of d as an omitted variable problem.

Consider the outcome equation (9) together with the selection rule (3)-(4). In this section we again omit the time subscript since only cross section data is generally required by the CF method. For simplicity of notation and exposition, we also drop the regressors X in the outcome equation (considered under matching) implicitly assuming that all the analysis is conditional on the observables X .

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

$$u \perp d, Z | v \tag{57}$$

That is, were we be able to control for v , d would become exogenous in the outcome equation. It 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 that

is required is mean independence of u from d and Z conditional on v :

$$E[u|v, d, Z] = E[u|v] = h(v) \tag{58}$$

where h is a function of v , the control function.

Both (57) and (58) recover the randomization hypothesis (R1) *conditional* on the unobservable term v . As discussed before with respect to other non-experimental approaches, assumption (R2) is harder to reproduce and is not considered in most empirical studies. Thus, only the ATT will be identified in general.

The control function method is close to a fully structural approach in the sense that it *explicitly* incorporates the decision process in the estimation of the impact of the treatment. The problem is how to specify and 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 parametric specification for the selection rule in (5), where all that can be inferred when the parameters γ are known is whether v is above or below $-Z\gamma$ depending on whether $d = 1$ or $d = 0$.

Early applications of the control function approach use a parametric assumption on the joint distribution of the error terms, u and v , and a functional form assumption for the decision rule. The most commonly encountered set of assumptions impose joint normality and linearity. The selection model of outcomes becomes:

$$\begin{aligned} y_i &= \beta + \alpha_i d_i + u_i \\ d_i &= \mathbf{1}(Z_i \gamma + v_i) \\ (u, v) &\sim \mathcal{N}(0, \Sigma) \end{aligned}$$

When applied to this setup, the control function assumption (58) becomes

$$\begin{aligned} E[u|d = 1, Z] &= \rho\lambda_1(Z\gamma) \\ E[u|d = 0, Z] &= \rho\lambda_0(Z\gamma) \end{aligned} \tag{59}$$

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

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

where n and Φ stand for the standard normal pdf and cdf, respectively. Thus, joint normality implies that the conditional expectation of u conditional on d and Z is a known function of the threshold, $Z\gamma$, that determines the assignment propensity: $P(d_i = 1|Z_i) = P(v_i > -Z_i\gamma|Z_i)$.

This model can be estimated using the Heckit procedure, Heckman (1976, 1979). This is a two-step estimator. The first step generates predictions of the control functions specified above from a regression of d on Z . The second step estimates the enlarged outcome equation by OLS:

$$y_i = \beta + d_i(\alpha^{ATE} + E[\alpha_i - \alpha^{ATE}|d_i = 1]) + \left[\rho d_i \frac{n(Z_i\hat{\gamma})}{\Phi(Z_i\hat{\gamma})} + \rho(1 - d_i) \frac{-n(Z_i\hat{\gamma})}{1 - \Phi(Z_i\hat{\gamma})} \right] + \delta_i \tag{60}$$

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

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

It is clear from the regression equation (60) that only the ATT can be identified when the impact of treatment is heterogeneous: $\alpha^{ATT} = \alpha^{ATE} + E[\alpha_i - \alpha^{ATE}|d_i = 1]$.²⁵

²⁵An additional joint normality assumption between the idiosyncratic gains α_i and the unobservable v would further allow for the identification of the ATE but this assumption is not required to estimate the ATT.

8.2 Weaknesses of the control function method

The relative robustness of classical parametric CF comes from the structure it imposes on the selection process. This makes this approach particularly informative for the policy maker by allowing for selection on the unobservables and supporting the extrapolation of results to alternative policy scenarios. However, this same feature has been strongly criticized for being overly restrictive. A number of semi-parametric CF estimators have been proposed that deal (at least partially) with this problem (e.g. see the review by Powell (1994) and also Ahn and Powell, 1993, Andrews and Schafgans, 1998, and Das, Newey and Vella, 2003).

In its non-parametric setup, the CF approach has been shown to be equivalent to the LATE approach. While such advances deal with the main criticism to CF, they also reduce the usefulness of the CF approach to inform about possible changes in the policy scenario.

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

There are two key assumptions underlying the selection model specified in the previous section: *(i)* the parametric assumption for the joint distribution of unobservables and *(ii)* the linear index assumption on the selection rule. As noted above important recent developments have proposed new semi-parametric estimators that relax these assumptions. More recently, Vytlacil (2002) has shown that the LATE approach can be seen as an application of a selection model. To see this, we first compare the two methods and then briefly discuss the equivalence result of Vytlacil.

Consider the non-parametric selection model under additive separability of the error term. It can be written as:

$$d_i = \mathbf{1}(v_i < g(Z_i)) \tag{61}$$

Also consider the outcome model as specified in (1)-(2).

The CF method uses the additive separability in the selection rule to acquire all the required information about the unobservable v from the observation of d and Z . Estimation is based on the

following regression function:

$$E(y_i|Z_i, d_i) = \beta + d_i E[\alpha_i|Z_i, d_i = 1] + d_i \Lambda_1(Z_i) + (1 - d_i) \Lambda_0(Z_i) \quad (62)$$

where $\Lambda_1(Z_i) = E(u_i|d_i = 1, Z_i)$ and $\Lambda_0(Z_i) = E(u_i|d_i = 0, Z_i)$. On the top of additive separability of v in the selection model, the following assumptions are required to establish (62):

- Z is independent of (y^0, y^1, v) .
- d is a non-trivial function of Z .

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

The regression model (62) explicitly controls for the part of the error term u correlated with the participation status d , thus eliminating the endogeneity problem.

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

$$E(y_i|Z_i) = \beta + p(d_i = 1|Z_i) E[\alpha_i|Z_i, d_i = 1]$$

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

- Z is independent of $(y^0, y^1, d(Z))$.
- $d(Z)$ is a non-trivial function of Z , where $d(Z)$ is a random variable representing the treatment status and its dependence on Z means that the distribution of d depends on Z .
- $d(Z') \geq d(Z'')$ (or $d(Z') \leq d(Z'')$) for all individuals.

The first assumption is the exclusion restriction together with the exogeneity of Z in the selection rule. The later is required to ensure that Z does not affect the outcome other than through d , even if only indirectly through some relation with v that may itself be related with u . The second assumption ensures that the probability of participation changes with Z . The third assumption is

the monotonicity assumption. It is required to ensure that the difference in the average outcomes evaluated at two distinct realizations of Z , say Z' and Z'' , is solely due to the move into treatment of some individuals (or out of treatment, depending on the direction of the monotonic relation) and not to a complete change in the composition of the treatment group due to some individuals moving in and some others moving out of treatment.

Notice that the LATE approach does not impose the selection model (61) to identify the treatment effect. It does not require any functional form or distributional assumptions, instead relying on the general form for the decision process as specified in (3)-(4). LATE uses the monotonicity assumption to replace the selection structure. As for the rest of the CF and LATE assumptions, they are equivalent - the first and second assumptions in each set are the same.

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

The reverse implication, however, is not necessarily true. However, when taken together the LATE assumptions are equivalent to the CF assumptions. Vytlačil (2002) shows that under the LATE assumptions it is always possible to construct a selection model $\tilde{d}(Z)$ of the type (61) satisfying the CF assumptions and such that $\tilde{d}(Z) = d(Z)$ almost everywhere. This means that under the LATE assumptions stated above, we can always find a selection model that rationalizes the data at hand. This equivalence result shows that the LATE approach can be seen as an application of a non-parametric version of the CF method.

Also notice that the local IV method of Heckman and Vytlačil (1999 and 2001) discussed earlier withdraws the monotonicity assumption of LATE and is instead based on the additive separability of the selection rule, as in (61). Thus, it is explicitly a CF method.

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

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

Rows (2)-(3) and (9)-(10) of table (7) show the estimates of the ATT. In general, estimates exhibit large bias, not identifying the correct parameter, despite including the correct observables in the selection process. There are two reasons for this. The first is the problem of specification. The decision rule in (12) is a non-linear function of the observables z and x . The second relates to the underlying assumption of joint normality of the disturbances in the selection rule and outcome equations. Although (u, v) are jointly normally distributed, these are not the only unobservable components of the two equations. The permanent (unobservable) ability also affects the selection rule through its effect on the returns to education. The specification in levels suffers from the additional problem that $\exp(u)$, not u , is part of the disturbance.

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

Results for the specification in levels (panel A) improve significantly with the inclusion of the new variables. Test score s seems to help in the identification of ATT when ability is not observed. If available, the inclusion of ability drives the bias towards zero. Allowing for a more general specification of the selection rule does not improve results. Results for the specification in logs are not as bright. The inclusion of s does not help and even ability may only improve the results slightly. The remaining problem in this case is the specification of the selection rule.

Table 7: Monte Carlo experiment - CF estimates of ATT and respective bias. Simulations under the assumption of negatively correlated residuals, (u, v)

		no subsidy		positive subsidy	
		estimate	bias	estimate	bias
		(1)	(2)	(3)	(4)
<i>Panel A: Estimates in levels</i>					
(1)	True parameters	3.308		2.993	
CF estimates using the selection variables:					
(2)	z	1.280	0.613	1.408	0.530
(3)	(z, x)	1.485	0.551	1.528	0.489
(4)	(z, s, x)	2.070	0.374	2.548	0.149
(5)	(z, s, x) +interactions	2.205	0.333	2.633	0.120
(6)	(z, θ, x)	3.299	0.003	2.919	0.025
(7)	(z, θ, x) +interactions	3.353	0.014	2.944	0.016
<i>Panel B: Estimates in logs</i>					
(8)	True parameters	0.434		0.453	
CF estimates using the selection variables:					
(9)	z	0.359	0.173	0.379	0.163
(10)	(z, x)	0.366	0.157	0.384	0.152
(11)	(z, s, x)	0.555	0.279	0.549	0.212
(12)	(z, s, x) +interactions	0.562	0.294	0.552	0.218
(13)	(z, θ, x)	0.426	0.018	0.519	0.146
(14)	(z, θ, x) +interactions	0.444	0.023	0.528	0.165

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 outcome is modeled in levels (panel A) or logs (panel B). The selection into education is modeled as an linear index model of the variables described in the first column of this table. z stands for family background; x is region; s is the test score and θ is ability. The interactions include higher order terms of each continuous variable and the product of each combination of two different variables. Simulations are based on 200 Monte-Carlo replications using samples of size $N = 2000$.

9 Summary

This paper has presented an overview of alternative 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 assignment 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 (hopefully plausible) model of economic choice. Perfect experimental designs and fully plausible structural 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.

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 randomization 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: A simple dynamic model of investment in education

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. The prototypical individual in this model lives for three periods, which we denote by $t = 0, 1, 2$. In period $t = 0$ all individuals are in school. In period $t = 1$ some individuals will enrol in college and in period $t = 3$ all individuals are working. The problem of the individual is to decide optimally about educational investment when there is uncertainty about the future returns to the investment. We will now explain the model in more detail.

At birth ($t = 0$) each individual is characterized by three variables, which we denote by (θ, z, x) . For interpretation purposes, we assume θ measures ability and is observable to the individual but unobservable to the econometrician. z is observable to the individual and econometrician and measures the conditions faced by the individual while young that affect the cost of education. It will be interpreted as some measure of family background or some measure of cost like distance to college. Finally, x is another observable variable to both the individual and econometrician. It measures market conditions and we interpret it as region. All three variables (θ, z, x) are assumed to remain unaltered throughout the individual's life.

Based on this information, the individual decides at $t = 0$ about the level of effort in school. Combined with ability θ , the endogenous effort will determine performance in school. This is measured as a score in a test and is denoted by s :

$$s_i = \gamma_0 + \gamma_1 \theta_i q_i + w_i \quad (63)$$

where q is effort, w is the unpredictable component of the score and (γ_0, γ_1) are the parameters.

The test score is revealed in the next period, $t = 1$, after the effort choice being made. Depending on its value, it may give access to subsidized education if such subsidy exists. Eligibility is defined on a threshold rule: students scoring above \underline{s} will be eligible while students scoring below this level will not.

Investment in high education has a (utility) cost, denoted by c . c depends on the individual's characteristics as well as on the test score if an education subsidy is available. In the presence of a subsidy, c is defined as

$$c_i = \delta_0 + \delta_1 z_i + \mathbf{1}(s_i \geq \underline{s})S + v_i \quad (64)$$

where S is the education 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, v is the unpredictable part of the cost of education and (δ_0, δ_1) are the parameters.

The decision of whether or not to invest in education occurs 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 and used to inform the decision process. Thus, the precise cost of education is known at 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. These are specified as:

$$\ln y_i = \beta_0 + \beta_1 x_i + (\alpha_0 + \alpha_1 \theta_i) d_i + u_i \quad (65)$$

where y is earnings, d is a dummy variable representing the education decision, the β 's and α 's are the parameters of the earnings function and u is the unpredictable component of earnings. Notice that the returns to education are not known in advance, at $t = 1$, because u is unknown and y is nonlinear in its arguments.

We now formalize the problem of the individual in a dynamic framework. The individual chooses effort in period $t = 0$ to maximize lifetime utility. His/her choice is conditional on how effort affects the test score (equation (63)) and the impact of the test score on the cost of education (equation (64)). It can be formalized as

$$V_{0i}(\theta_i, z_i, x_i) = \max_{q_i} \{-\lambda q_i + \rho E_{s,v} [V_{1i}(\theta_i, z_i, x_i, s_i, v_i)]\} \quad (66)$$

where V_{ti} represents the discounted value of present and future utility for individual i in period t , ρ 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 to be computed. From the above equation the optimal level of effort is a function of θ , z and x : $q^*(\theta, z, x)$.

The problem of the individual at time $t = 1$ is that of choosing the educational level without knowing the returns to the investment with certainty. Conditional on the (known) form of the earnings equation (65), the problem can be formalized as

$$V_{1i}(\theta_i, z_i, x_i, s_i, v_i) = \max_d \{-c_i d_i + \rho E_u [y_i(\theta_i, d_i, x_i, u_i) | v_i]\} \quad (67)$$

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

Under the model specification in equation (67), the education decision follows a reservation rule defined in the cost of education. The optimal decision is a function of the information set at time

$t = 1$, $d_i = d^*(\theta_i, x_i, z_i, s_i, v_i)$. Formally:

$$d_i = \begin{cases} 1 & \text{if } E_u(y_i | d_i = 1, x_i, \theta_i, v_i) - E_u(y_i | d_i = 0, x_i, \theta_i, v_i) > c_i \\ 0 & \text{otherwise} \end{cases} \quad (68)$$

Finally, in period 2 the individual works and collects lifetime earnings as defined in equation (65). There is no decision to be taken at this stage.

Average parameters

The impact of high education on the logarithm of earnings for individual i is

$$\alpha_i = \alpha_0 + \alpha_1 \theta_i$$

We can use this expression to specify the ATE on log earnings as

$$\begin{aligned} \alpha^{ATE} &= \alpha_0 + \alpha_1 E(\theta_i) \\ &= \alpha_0 + \alpha_1 \int_{\Theta} \theta f_{\theta}(\theta) d\theta \end{aligned}$$

where f_{θ} is the probability density function of θ and Θ is the space of possible realizations or domain of θ .

In a similar way, the ATT on log earnings is just

$$\alpha^{ATT} = \alpha_0 + \alpha_1 E(\theta_i | d_i = 1)$$

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 ,

$$u_i = \mu v_i + r_i$$

where $\mu \leq 0$ is the slope parameter and r is a iid shock. In this case, the reservation policy described in equation (68) in terms of the cost of education c can now be expressed in terms of the unobservable component, v . We denote it by \tilde{v} and note that it is a function of the variables known at time $t = 1$ that impact either on the cost of education or on the expected future earnings. Thus, $\tilde{v}(\theta, z, x, s)$ but

since $s = \gamma_0 + \gamma_1 \theta q(\theta, z, x) + w$ it is equivalent to write it as $\tilde{v}(\theta, z, x, w)$. The reservation policy \tilde{v} fully characterizes the educational decision: whenever the individual draws a shock $v > \tilde{v}$ the decision will be not to participate while the opposite happens when $v < \tilde{v}$. Thus, the decision rule (68) can be re-written as,

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

Conditional on the set of variables (θ, z, x, w) , the size of the population investing in education will be given by,

$$\begin{aligned} P[d = 1 | \theta, z, x, w] &= F_v(\tilde{v}(\theta, z, x, w)) \\ &= \int_{-\infty}^{\tilde{v}(\theta, z, x, w)} f_v(v) dv \end{aligned}$$

which is just the cumulative density function of v at the reservation point, $\tilde{v}(\theta, z, x, w)$. Notice that to derive the above expression it is being assumed that the v is independent of (θ, z, x, w) .

We can now integrate ability over the whole educated population to obtain $E(\theta | d = 1)$:

$$E(\theta | d = 1) = \int_{\Theta} \int_{\mathcal{D}(z)} \int_{\mathcal{D}(x)} \int_{-\infty}^{+\infty} \theta F_v(\tilde{v}(\theta, z, x, w)) f_{(\theta, z, x, w)}(\theta, z, x, w) dw dx dz d\theta$$

where Θ , $\mathcal{D}(z)$ and $\mathcal{D}(x)$ stand for the domains of θ , z and x , respectively, and $f_{(\theta, z, x, w)}$ is the joint density function of (θ, z, x, w) .

Parameters used in the simulations

- Discount parameter: $\rho = 1$
- Utility cost of effort to prepare test: $\lambda = 0.9$
- Test score (equation (63))
 - γ_0 : 1.0
 - γ_1 : 2.5
 - w : $N(0, 1)$
- Cost of education (equation (64))

δ_0 : 3.0
 δ_1 : -1.2
 \underline{s} : 4.0
 S : 2.5
 v : $N(0, 1)$

- Earnings (equation (65))

β_0 : 0.70

β_1 : 0.30

α_0 : 0.01

α_1 : 0.70

$u_i = \mu v_i + r_i$ where

μ : -0.5 in the correlated case or

0 in the non-correlated case

r : $N(0, \sigma^2 = 0.75)$ in the correlated case or

$N(0, 1)$ in the non-correlated case

- The state variables, θ , z and x are drawn from the following distributions,

θ : $N(0.5, \sigma = 0.25)$ truncated at 0 and 1

z : $N(0, 1)$ truncated at -2 and 2

x : 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 \mathbf{v} are those that draw $\mathbf{v}_i < p$. Thus,

$$\begin{aligned}
 \alpha^{ATT}(p) &= \int_0^p \alpha^{MTE}(\mathbf{v}) f_{\mathbf{v}}(\mathbf{v} | \mathbf{v} < p) d\mathbf{v} \\
 &= \frac{1}{p} \int_0^p \alpha^{MTE}(\mathbf{v}) d\mathbf{v}
 \end{aligned}$$

where the second equality results from the fact that \mathbf{v} is uniformly distributed. Integrating over all the support of p yields the overall ATT,

$$\begin{aligned}\alpha^{ATT} &= \int_0^1 \alpha^{ATT}(p) f_p(p|d=1) dp \\ &= \int_0^1 \int_0^p \alpha^{MTE}(\mathbf{v}) \frac{f_p(p|d=1)}{p} d\mathbf{v} dp\end{aligned}$$

Similarly, the ATE, ATNT and LATE are,

$$\begin{aligned}\alpha^{ATE} &= \int_0^1 \int_0^1 \alpha^{MTE}(\mathbf{v}) f_p(p) d\mathbf{v} dp \\ \alpha^{ATNT} &= \int_0^1 \int_p^1 \alpha^{MTE}(\mathbf{v}) \frac{f_p(p|d=0)}{1-p} d\mathbf{v} dp \\ \alpha^{LATE}(p^*, p^{**}) &= \frac{1}{p^{**} - p^*} \int_{p^*}^{p^{**}} \alpha^{MTE}(\mathbf{v}) d\mathbf{v}\end{aligned}$$

References

- [1] Abadie, A., Angrist, J. and Imbens, G. (2002), “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings”, *Econometrica*, 70(1), 91-117
- [2] Abbring, J. and van den Berg, G. (2003), “The Nonparametric Identification of Treatment Effects in Duration Models”, *Econometrica*, 71(5), 1491-1517
- [3] Ahn, H. and Powell, J. (1993), “Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism”, *Journal of Econometrics*, 58, 3-29
- [4] Andrews, D. and Schafgans, M. (1998), “Semiparametric Estimation of the Intercept of a Sample Selection Model”, *Review of Economic Studies*, 65(3), 497-517
- [5] Ashenfelter, O. (1978), “Estimating the Effect of Training Programs on Earnings”, *Review of Economics and Statistics*, 60, 47-57
- [6] Ashenfelter, O. and Card, D. (1985), “Using the Longitudinal Structure of Earnings to Estimate the Effect of Training Programs”, *Review of Economics and Statistics*, 67, 648-660
- [7] Athey, S. and Imbens, G. (2006), “Identification and Inference in Nonlinear Difference-In-Differences Models”, *Econometrica*, 74(2), 431-497
- [8] Bassi L. (1983), “The Effect of CETA on the Post-Program Earnings of Participants”, *Journal of Human Resources*, 18, 539-556
- [9] Bassi, L. (1984), “Estimating the Effects of Training Programs with Nonrandom Selection”, *Review of Economics and Statistics*, 66, 36-43
- [10] Battistin, E. and Rettore, E. (2007), “Ineligibles and Eligible Non-Participants as a Double Comparison Group in Regression Discontinuity Designs”, *Journal of Econometrics*, forthcoming
- [11] Becker, S. and Ichino, A. (2002), “Estimation of average treatment effects based on propensity scores”, *The Stata Journal*, http://www.labor-torino.it/pdf_doc/ichino.pdf
- [12] Bell, B., Blundell, R. and Van Reenen, J. (1999), “Getting the Unemployed Back to Work: An Evaluation of the New Deal Proposals”, *International Tax and Public Finance*, 6, 339-360
- [13] Bjorklund, A. and Moffitt, R. (1987), “The Estimation of Wage Gains and Welfare Gains in Self-Selection Models”, *Review of Economics and Statistics*, 69(1), 42-49

- [14] Blundell, R., Costa Dias, M., Meghir, C. and Van Reenen, J. (2004), "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program", *Journal of the European Economics Association*, 2(4), 596-606
- [15] **Blundell, R. and Costa Dias M. (2000), "Evaluation Methods for Non-Experimental Data", *Fiscal Studies*, 21(4), 427-468**
- [16] **Blundell, R., Dearden, L. and Meghir, C. (1996), "The Determinants and Effects of Work-Related Training in Britain", London: Institute for Fiscal Studies**
- [17] Blundell, R., Dearden, L. and Sianesi, B. (2005), "Evaluating the Impact of Education on Earnings: Models, Methods and Results from the NCDS", *Journal of the Royal Statistical Society, Series A*, 168(3), 473-512
- [18] **Blundell, R., Duncan, A. and Meghir, C. (1998), "Estimating Labour Supply Responses using Tax Policy Reforms", *Econometrica*, 66, 827-861**
- [19] Blundell, R. and MaCurdy, T. (1999), "Labor Supply: A Review of Alternative Approaches", in A. Ashenfelter and D. Card (eds), *Handbook of Labour Economics*, vol. 3, Amsterdam: Elsevier Science
- [20] Blundell, R. and Powell, J. (2003), "Endogeneity in Nonparametric and Semiparametric Regression Models", in M. Dewatripont, L. Hansen and S. Turnovsky (eds.), *Advances in Economics and Econometrics*, Cambridge University Press
- [21] **Blundell, R. and Powell, J. (2004), "Endogeneity in Semiparametric Binary Response Models", *The Review of Economic Studies*, 71(3), 581-913**
- [22] **Burtless, G. (1985), "Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment", *Industrial and Labor Relations Review*, 39, 105-114**
- [23] Card, D. and Robins, P. (1998), "Do Financial Incentives Encourage Welfare Recipients To Work?", *Research in Labor Economics*, 17, 1-56
- [24] Carneiro, P., Hansen, K. and Heckman, J. (2001), "Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policy", *Swedish Economic Policy Review*, 8, 273-301
- [25] Carneiro, P., Hansen, K. and Heckman, J. (2003), "Estimating Distributions of Counterfactuals with an Application to the Returns to Schooling and Measurement of the Effect of Uncertainty on Schooling Choice", *International Economic Review*, 44(2), 361-422

- [26] Carneiro, P., Heckman, J. and Vytlacil, E. (2005), "Understanding What Instrumental Variables Estimate: Estimating the Average and Marginal Return to Schooling", unpublished manuscript
- [27] Cochran, W. and Rubin, D. (1973), "Controlling Bias in Observational Studies", *Sankhya*, 35, 417-446
- [28] Das, M., Newey, W. and Vella, F. (2003), "Nonparametric Estimation of Sample Selection Models" *Review of Economic Studies*, 70, 33-58
- [29] De Giorgi, G. (2005), "The New Deal for Young People Five Years On", *Fiscal Studies*, 26(3), 371-383
- [30] Eissa, N. and Liebman, J. (1996), "Labor Supply Response to the Earned Income Tax Credit", *Quarterly Journal of Economics*, CXI, 605-637
- [31] Devine, T. and Heckman, J. (1996), "Consequences of Eligibility Rules for a Social Program: A Study of the Job Training partnership Act (JTPA)", in S. Polachek (ed.), *Research in Labor Economics*, 15, CT: JAI Press, 111-170
- [32] Fan, J. (1992), "Design Adaptive Nonparametric Regression", *Journal of the American Statistical Association*, 87, 998-1004
- [33] Fisher, R. (1951), *The Design of Experiments*, 6th edition, London: Oliver and Boyd.
- [34] Gronau, R. (1974), "Wage Comparisons - A Selectivity Bias", *Journal of Political Economy*, 80, S74-S103
- [35] Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects", *Econometrica*, 66, 315-331
- [36] Hahn, J., Todd, P. and Van der Klaauw, W. (2001), "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design", *Econometrica*, Vol 69(1): 201-209
- [37] Hausman, J. and Wise, D. (1985), *Social Experimentation*, NBER, Chicago: University of Chicago Press
- [38] Heckman, J. (1976), "The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such methods", *Annals of Economic and Social Measurement*, 5, 475-492
- [39] Heckman, J. (1979), "Sample Selection Bias as a Specification Error", *Econometrica*, 47, 153-611

- [40] Heckman, J. (1990), “Varieties of Selection Bias”, *American Economic Review*, 80, 313-318
- [41] Heckman, J. (1992), “Randomization and Social program”, in C. Manski and I. Garfinkle (eds.), *Evaluating Welfare and Training Programs*, Cambridge, Mass: Harvard University Press
- [42] Heckman, J. (1996), “Randomization as an Instrumental Variable Estimator”, *Review of Economics and Statistics*, 56, 336-341
- [43] Heckman, J. (1997), “Instrumental Variables: A Study of the Implicit Assumptions underlying one Widely used Estimator for Program Evaluations”, *Journal of Human Resources*, 32, 441-462
- [44] Heckman, J. and Hotz, V. (1989), “Choosing among Alternatives Nonexperimental Methods for Estimating the Impact of Social programs”, *Journal of the American Statistical Association*, 84, 862-874
- [45] Heckman, J., Ichimura, H. and Todd, P. (1997), “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme”, *The Review of Economic Studies*, 64(4), 605-654
- [46] Heckman, J., Ichimura, H. and Todd, P. (1998), “Matching as an Econometric Evaluation Estimator”, *The Review of Economic Studies*, 65(2), 261-294
- [47] Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1998), “Characterising Selection Bias Using Experimental Data”, *Econometrica*, 66(5), 1017-1098
- [48] Heckman, J., LaLonde, R. and Smith, J. (1999), “The Economics and Econometrics of Active Labour Market Programs”, in A. Ashenfelter and D. Card (eds.), *Handbook of Labour Economics*, vol. 3, Amsterdam: Elsevier Science
- [49] Heckman, J. and Lozano, S. (2004), “Using Matching, Instrumental Variables and Control Functions to Estimate Economic Choice Models”, *The Review of Economics and Statistics*, 86, 30-57
- [50] Heckman, J. and Robb, R. (1985), “Alternative Methods for Evaluating the Impact of Interventions”, in J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labour Market Data*, New York: Wiley
- [51] Heckman, J. and Robb, R. (1986), “Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes”, in H. Wainer (ed.), *Drawing Inferences from Self-Selected Samples*, Berlin: Springer Verlag

- [52] Heckman, J. and Smith, J. (1999), “The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies”, *Economic Journal*, 109, 313-348
- [53] Heckman, J., Smith, J. and Clements, N. (1997), “Making the Most out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in program Impacts”, *Review of Economic Studies*, 64, 487-536
- [54] Heckman, J. and Vytlacil, E. (1998), “Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling”, *Journal of Human Resources*, 33(4), 974-987
- [55] Heckman, J. and Vytlacil, E. (1999), “Local Instrumental Variables and Latent Variable Models for Identifying the Bounding Treatment Effects”, *Proceedings of the National Academy of Sciences*, 96(8), 4730-4734
- [56] Heckman, J. and Vytlacil, E. (2001), “Local Instrumental Variables”, in C. Hsiao, K. Morimune, and J. Powell (eds.), *Nonlinear Statistical Modeling: Essays in Honor of Takeshi Amemiya*, New York: Cambridge University Press.
- [57] Heckman, J. and Vytlacil, E. (2006), “Econometric Evaluation of Social Programs”, in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6, Amsterdam: Elsevier
- [58] Horowitz, J. (2001), “The Bootstrap”, in J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Vol. 5, Amsterdam: Elsevier
- [59] Imbens, G. and Angrist, J. (1994), “Identification and Estimation of Local Average Treatment Effects”, *Econometrica*, 62(2), 467-75
- [60] Imbens, G. and Lemieux, T., (2007), “Regression discontinuity designs: A guide to practice”, *Journal of Econometrics*, forthcoming
- [61] **Imbens, G. and Newey, W. (2002), “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity”, NBER working paper 0285**
- [62] **Kemple, J., Dolittle, F. and Wallace, J. (1993), “The National JTPA Study: Site Characteristics in participation patterns”, New York: Manpower Demonstration Research Corporation**
- [63] LaLonde, R. (1986), “Evaluating the Econometric Evaluations of Training Programs with Experimental Data”, *American Economic Review*, 76, 604-620

- [64] Larsson, L. (2003), “Evaluation of Swedish Youth Labor Market Programs”, *Journal of Human Resources*, 38, 891–927
- [65] Leuven, E. and B. Sianesi (2003) “PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing”, Statistical Software Components S432001, Boston College Department of Economics, revised 28 December
- [66] Moffitt, R. (2007), “Estimating Marginal Treatment Effects in Heterogeneous Populations” , Johns Hopkins, Department of Economics, Working Papers
- [67] Nickell, Stephen (1981). “Biases in Dynamic Models with Fixed Effects”, *Econometrica*, 49(6): 1417-1426
- [68] **Orr, L., Bloom, H., Bell, S., Lin, W., Cave, G. and Dolittle, F. (1994), “The National JTPA Study: Impacts, Benefits and Costs of Title II-A” , A Report to the US Department of Labor, 132, Abt Associates: Bethesda, Maryland**
- [69] Powell, J. (1994), “Estimation of Semiparametric Models”, in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, Vol. 4, Amsterdam: North Holland
- [70] Rosenbaum, P. and Rubin, D. (1983), “The Central Role of the Propensity Score in Observational Studies for Causal Effects”, *Biometrika*, 70, 41-55
- [71] Rosenbaum, P. and Rubin, D. (1984), “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score”, *Journal of the American Statistical Association*, 79, 516-524
- [72] Rosenbaum, P. and Rubin, D. (1985), “Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score”, *American Statistician*, 39-38
- [73] **Rubin, D. (1978), “Bayesian Inference for Causal Effects: The Role of Randomization”, *Annals of Statistics*, 7, 34-58**
- [74] Rubin, D. (1979), “Using Multivariate Matched Sampling and regression Adjustment to Control Bias in Observational Studies”, *Journal of the American Statistical Association*, 74, 318-329
- [75] **Sianesi, B. (2001), “An evaluation of the Swedish system of active labour market programmes in the 1990s”, IFS WP W02/01**
- [76] Vytlačil, E. (2002), “Independence, Monotonicity, and Latent Index Models: An Equivalence Result”, *Econometrica*, 70(1), 331-341