Blundell, Richard W.; Dias, Monica Costa

Working Paper

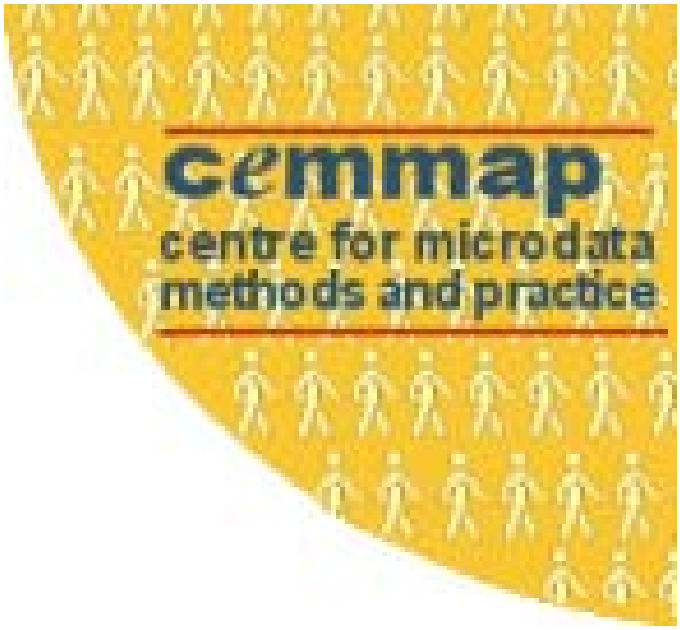
Alternative approaches to evaluation in empirical microeconomics

cemmap working paper, No. CWP10/02

Provided in Cooperation with:
Institute for Fiscal Studies (IFS), London

Suggested Citation: Blundell, Richard W.; Dias, Monica Costa (2002) : Alternative approaches to evaluation in empirical microeconomics, cemmap working paper, No. CWP10/02, Centre for Microdata Methods and Practice (cemmap), London, http://dx.doi.org/10.1920/wp.cem.2002.1002

This Version is available at:
http://hdl.handle.net/10419/79305



ALTERNATIVE APPROACHES TO EVALUATION IN EMPIRICAL MICROECONOMICS

Richard Blundell
Monica Costa Dias

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP10/02

Alternative Approaches to Evaluation in Empirical Microeconomics

Richard Blundell* and Monica Costa Dias

University College London and Institute for Fiscal Studies

March 2002

Abstract

Four alternative but related approaches to empirical evaluation of policy interventions are studied: social experiments, natural experiments, matching methods, and instrumental variables. In each case the necessary assumptions and the data requirements are considered for estimation of a number of key parameters of interest. These key parameters include the average treatment effect, the treatment of the treated and the local average treatment effect. Some issues of implementation and interpretation are discussed drawing on the labour market programme evaluation literature.

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

Acknowledgements: This review was prepared for the special 'microeconometrics' PEJ. Comments from the editors, the referee and participants at the CeMMAP conference at which the papers for this volume were presented are gratefully acknowledged. The research is part of the program of the ESRC Centre for the Microeconomic Analysis of Fiscal Policy at IFS. Financial support from the ESRC is gratefully acknowledged. The second author also acknowledges the financial support from Sub-Programa Ciência e Tecnologia do Segundo Quadro Comunitário de Apoio, grant number PRAXIS XXI/BD/11413/97. The usual disclaimer applies.

*Address: University College London and Institute for Fiscal Studies, 7 Ridgemount Street, London, WC1E 7AE. r.blundell@ucl.ac.uk, http://www.ifs.org.uk

1. Introduction

In this review we consider four distinct but closely related approaches to the evaluation problem in empirical microeconomics: (i) social experiments, (ii) natural experiments, (iii) matching methods, and (iv) instrumental methods. The first of these approaches is closest to the 'theory' free method of medical experimentation since it relies on the availability of a randomised control. The last approach is closest to the structural econometric method since it relies directly on exclusion restrictions. Natural experiments and matching methods lie somewhere in between in the sense that they attempt to mimic the randomised control of the experimental setting but do so with non-experimental data and consequently place reliance on independence and/or exclusion assumptions.

Our concern here is with the evaluation of a policy intervention at the microeconomic level. This could include training programmes, welfare programmes, wage subsidy programmes and tax-credit programmes, for example. At the heart of this kind of policy evaluation is a missing data problem since, at any moment in time, an individual is either in the programme under consideration or not, but not both. If we could observe the outcome variable for those in the programme had they not participated then there would be no evaluation problem of the type we discuss here. Thus, constructing the counterfactual is the central issue that the evaluation methods we discuss address. Implicitly, each of the four approaches provides an alternative method of constructing the counterfactual.

The literature on evaluation methods in economics is vast and continues to grow. There are also many references in the literature which document the development of the analysis of the evaluation problem in economics. In the labour 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).

In many ways the *social experiment method* is the most convincing method of evalu-

ation since it directly constructs a control (or comparison) group which is a randomised 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 Cochrane and Rubin (1973) and Fisher (1951), for example). A properly defined social experiment can overcome the missing data problem. For example, in the design of the impressive study of the Canadian Self Sufficiency Project reported in Card and Robbins (1998), the labour supply responses of approximately 6000 single mothers in British Columbia to an in-work benefit programme, in which half those eligible were randomly excluded from the programme, were recorded. This study has produced invaluable evidence on the effectiveness of financial incentives in inducing welfare recipients into work.

Of course, social experiments have their own drawbacks. They are rare in economics and typically expensive to implement. They are not amenable to extrapolation. That is, they cannot easily be used in the ex-ante analysis of policy reform proposals. They also require the control group to be completely unaffected by the reform, typically ruling out spillover, substitution, displacement and equilibrium effects on wages etc. None-the-less, they have much to offer in enhancing our knowledge of the possible impact of policy reforms. Indeed, a comparison of results from non-experimental data to those obtained from 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 (1997) and Heckman, Smith and Clements (1997) use experimental data to assess the reliability of comparison groups used in the evaluation of training programmes. We draw on the results of these studies below.

It should be noted that randomisation can be implemented by area. If this corresponds to a local (labour) market, then general equilibrium or market level spillover effects will be accounted for. The use of control and treatment areas is a feature of the design of the New Deal evaluation data base in the UK. In the discussion below the

area to area comparisons are used to comment on the likely size of general equilibrium and spillover effects.

The *natural experiment approach* considers the policy reform itself as an experiment and tries to find a naturally occurring comparison group that can mimic the properties of the control group in the properly designed experimental context. This method is also often labelled "difference-in-differences" since it is usually implemented by comparing the difference in average behaviour before and after the reform for the eligible group with the before and after contrast for the comparison group. In the absence of a randomised experiment and under certain very strong conditions, this approach can be used to recover the average effect of the programme on those individuals entered into the programme - or those individuals "treated" by the programme. Thus measuring the average effect of the treatment on the treated. It does this by removing unobservable individual effects and common macro effects. However, it relies on the two critically important assumptions of (i) *common time effects across groups*, and (ii) *no systematic composition changes within each group*. These two assumptions make choosing a comparison group extremely difficult. For example, in their heavily cited evaluation study of the impact of Earned Income Tax Credit (EITC) reforms on the employment of single mothers in the US, Eissa and Liebman (1996) use single women without children as one possible control group. However, this comparison can be criticized for not satisfying the common macro effects assumption (i). In particular, the control group is already working to a very high level of participation in the US labour market (around 95%) and therefore cannot be expected to increase its level of participation in response to the economy coming out of a recession. In this case all the expansion in labour market participation in the group of single women with children will be attributed to the reform itself. In the light of this criticism the authors also use low education childless single women as a control group for which nonparticipation is much more common and who have other similar characteristics to those single parents eligible to EITC.

The *matching method* has a long history in non-experimental statistical evaluation (see Heckman, Ichimura and Todd (1997), Rosenbaum and Rubin (1985) and Rubin (1979)). The aim of matching is simple. It is to select sufficient observable factors that any two individuals with the same value of these factors will display no systematic differences in their reaction to the policy reform. Consequently, if each individual undergoing the reform can be matched with an individual with the same matching variables that has not undergone the reform, the impact on individuals of that type can be measured. It is a matter of prior assumption as to whether the appropriate matching variables have been chosen. If not, the counterfactual effect will not be correctly measured. Again experimental data can help here in evaluating the choice of matching variables and this is precisely the motivation for the Heckman, Ichimura and Todd (1997) study. As we document below, matching methods have been extensively refined in the recent evaluation literature and are now a valuable part of the evaluation toolbox.

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

The distinction between homogenous and heterogeneous treatments effects that is highlighted in this recent instrumental variable literature is central to the definition of a 'parameter of interest' in the evaluation problem. In the homogeneous linear model there is only one impact of the programme and it is one that would be common to participants

and nonparticipants a like. In the heterogeneous model, those that are treated may have a different mean impact of the programme from those not treated. Certainly this is likely to be the case in a non-experimental evaluation where participation provides some gain to the participants. In this situation we can define a treatment on the treated parameter that is different from a treatment on the untreated parameter or the average treatment effect. One central issue in understanding evaluation methods is clarifying what type of treatment effect is being recovered by these different approaches.

We should note that we do not consider fully structural econometric choice models in this review. These have been the cornerstone of nonexperimental evaluation (and simulation) of tax and welfare policies. They provide a comprehensive analysis of the choice problem facing individuals deciding on programme participation. They explicitly describe the full constrained maximisation problem and are therefore perfectly suited for ex-ante policy simulation. Blundell and MaCurdy (1999) provide a comprehensive survey and a discussion of the relationship of the structural choice approach to the evaluation approaches presented here.

The rest of the paper is organized as follows. In the next section we lay out the different definitions of treatment parameters and ask: what are we trying to measure? Section 3 considers the types of data and their implication for the choice of evaluation method. Section 4 is the main focus of this paper as it presents a detailed comparison of alternative methods of evaluation for non-experimental data. In section 5 we illustrate these methods drawing on recent applications in the evaluation literature. Section 6 concludes.

2. Which Parameter of Interest?

We begin by presenting a general model of outcomes which can then assume particular forms depending on the amount of structure one wishes, or needs, to include. There are several important decisions to be taken when specific applications are consid-

ered, the one we are especially concerned with is whether the response to the treatment is homogeneous across individuals or heterogeneous. Typically, we do not expect all individuals to be affected by a policy intervention in exactly the same way - there will be heterogeneity in the impact across individuals. Consequently, there are different potential questions that evaluation methods attempt to answer, the most commonly considered being the average effect on individuals of a certain type. This includes a wide range of parameters such as 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 (TTE), the effect of treatment on agents that are indifferent to participation, which is the marginal version of the local average treatment effect (LATE) discussed below, or the effect of treatment on the untreated (TU) which is typically an interesting measure for decisions about extending some treatment to a group that was formerly excluded from treatment. Under the homogeneous treatment effect assumption all these measures are identical, but this is clearly not true when treatment effects depend on individual's characteristics. From now onwards, except if explicitly mentioned, anywhere we discuss heterogeneous treatment effects the analysis pertains the TTE parameter.

To make things more precise, suppose there is a policy reform or intervention at time k for which we want to measure the impact on some outcome variable, Y. This outcome is assumed to depend on a set of exogenous variables, X, the particular relationship being dependent on the participation status in each period t. Let D be a dummy variable representing the treatment status, assuming the value 1 if the agent has been treated and 0 otherwise. The outcome's equations can be generically represented as follows,

$$\begin{array}{rcl} Y_{it}^{1} & = & g_{t}^{1}(X_{i}) + U_{it}^{1} \\ Y_{it}^{0} & = & g_{t}^{0}(X_{i}) + U_{it}^{0} \end{array} \tag{2.1}$$

where the superscript stands for the treatment status and the subscripts i and t identify the agent and the time period, respectively. The functions g^0 and g^1 represent the

relationship between the potential outcomes (Y^0, Y^1) and the set of observables X and (U^0, U^1) stand for the error terms of mean zero and assumed to be uncorrelated with the regressors X. The X variables are not affected by treatment (or pre-determined) and are assumed known at the moment of deciding about participation. For this reason we have excluded the time subscript from X. For comparison purposes, this means that agents are grouped by X before the treatment period and remain in the same group throughout the evaluation period. This is a general form of the switching regimes or endogenous selection model.

We assume that the participation decision can be parameterised in the following way: For each individual there is an index, IN, depending on a set of variables W, for which enrolment occurs when this index raises above zero. That is:

$$IN_i = f(W_i) + V_i \tag{2.2}$$

where V_i is the error term, and,

$$\begin{array}{ll} D_{it} = 1 & \text{if } IN_i > 0 \text{ and } t > k \\ D_{it} = 0 & \text{otherwise} \end{array} . \tag{2.3}$$

Except in the case of experimental data, assignment to treatment is most probably not random. As a consequence, the assignment process is likely to lead to a non-zero correlation between enrolment in the programme - represented by D_{it} - and the outcome's error term - $\left(U^0, U^1\right)$. This happens because an individual's participation decision is probably based on personal unobservable characteristics that may well affect the outcome Y as well. If this is so, and if we are unable to control for all the characteristics affecting Y and D simultaneously, then some correlation between the error term and the participation variable is expected. Any method that fails to take such problem into account is not able to identify the true parameter of interest.

Under the above specification, one can define the individual-specific treatment effect, for any X_i, to be

$$\alpha_{it}(X_i) = Y_{it}^1 - Y_{it}^0 = \left[g_t^1(X_i) - g_t^0(X_i)\right] + \left[U_{it}^1 - U_{it}^0\right] \qquad \text{with } t > k. \tag{2.4}$$

The different parameters of interest measured in period $t > k$, can then be expressed as:

Average Treatment Effect:

$$\alpha_{\mathsf{ATE}} = E(\alpha_{\mathsf{it}}|X = X_{\mathsf{i}}),$$

Average Treatment on the Treated Effect:

$$\alpha_{\mathsf{TTE}} = E\left(\alpha_{\mathsf{it}}|X = X_{\mathsf{i}}, D_{\mathsf{t}} = 1\right),$$

Average Treatment on the Untreated Effect:

$$\alpha_{\mathsf{TU}} = E\left(\alpha_{\mathsf{it}}|X = X_{\mathsf{i}}, D_{\mathsf{t}} = 0\right).$$

2.1. Homogeneous Treatment Effects

The simplest case is when the effect is assumed to be constant across individuals, so that

$$\alpha_{\mathsf{t}} = \alpha_{\mathsf{it}}\left(X_{\mathsf{i}}\right) = g_{\mathsf{t}}^{1}\left(X_{\mathsf{i}}\right) - g_{\mathsf{t}}^{0}\left(X_{\mathsf{i}}\right) \qquad \text{with } t > k$$

for any i. But this means that g^1 and g^0 are two parallel curves, only differing in the level, and the participation-specific error terms are not affected by the treatment status. The outcome's equation (2.1) can therefore be re-written as

$$Y_{\mathsf{it}} = g_{\mathsf{t}}^{0}\left(X_{\mathsf{i}}\right) + \alpha_{\mathsf{t}} D_{\mathsf{it}} + U_{\mathsf{i}}. \tag{2.5}$$

2.2. Heterogeneous Treatment Effects

However, it seems reasonable to assume that the treatment impact varies across individuals. These differentiated effects may come systematically through the observables' component or be a part of the unobservables. Without loss of generality, the

outcome's equation (2.1) can be re-written as follows

$$\begin{aligned} Y_{it} &= D_{it} Y_{it}^1 + (1 - D_{it}) Y_{it}^0 \\ &= g_t^0(X_i) + \alpha_{it}(X_i) D_{it} + U_{it}^0 \\ &= g_t^0(X_i) + \alpha_t(X_i) D_{it} + \left[U_{it}^0 + D_{it}\left(U_{it}^1 - U_{it}^0\right)\right] \end{aligned} \tag{2.6}$$

where

$$\alpha_t(X_i) = E\left[\alpha_{it}(X_i)\right] = g_t^1(X_i) - g_t^0(X_i) \tag{2.7}$$

is the expected treatment effect at time t among agents characterised by X_i.[1]

Two issues are particularly important under this more general setting. The first relates to the observables and their role in the identification of the parameter of interest. It is clear that the common support problem is central to this setting:[2] contrary to the homogeneous treatment effect, this structure does not allow extrapolation to areas of the support of X that are not represented at least among treated (if a particular parameterisation of g^0 is assumed one may be able to extrapolate among non-treated).

The second problem concerns the form of the error term, which differs across observations according to the specific treatment status. If there is selection on the unobservables, the OLS estimator after controlling for the covariates X is inconsistent for $\alpha_t(X)$, identifying the following parameter,

$$E\left(\hat{\alpha}_t(X)\right) = \alpha_t(X) + E(U_t^1 \mid X, D_t = 1) - E(U_t^0 \mid X, D_t = 0) \qquad \text{with } t > k.$$

3. Experimental And Non-Experimental Data

3.1. Experimental Data

Under ideal conditions to be discussed below, experimental data provides the correct missing counterfactual, eliminating the evaluation problem. The contribution of

[1] Specification (2.6) obviously includes (2.5) as a special case.

[2] By *common support* it is meant the subspace of individual characteristics that is represented both among treated and non-treated.

experimental data is to rule out bias from self-selection as individuals are randomly assigned to the programme. To see why, imagine an experiment that randomly chooses individuals from a group to participate in a programme - these are administered the treatment. It means that assignment to treatment is completely independent from a possible outcome or the treatment effect. Under the assumption of no spillover (general equilibrium) effects, the group of non-treated is statistically equivalent to the treated group in all dimensions except treatment status. The ATE within the experimental population can be simply measured by

$$\hat{\alpha}_{\mathsf{ATE}} = \bar{Y}_{\mathsf{t}}^{1} - \bar{Y}_{\mathsf{t}}^{0} \qquad t > k \tag{3.1}$$

where $\bar{Y}_{\mathsf{t}}^{(1)}$ and $\bar{Y}_{\mathsf{t}}^{(0)}$ stand for the treated and non-treated average outcomes at a time t after the programme.

However, a number of disrupting factors may interfere with this type of social experiments, invalidating the results. First, we expect some individuals to dropout, and the process is likely to affect treatments and controls unevenly and to occur non-randomly. The importance of the potential non-random selection may be assessed by comparing the observable characteristics of the remaining treatments and controls and treatment groups. Second, given the complexity of the contemporaneous welfare systems, truly committed experimental controls may actively search for alternative programmes and are likely to succeed. Moreover, observed behaviour of the individuals may also change as a consequence of the experiment itself as, for instance, the officers may try to "compensate" excluded agents by providing them detailed information about other programmes. It is even possible that some controls end up receiving precisely the same treatment being enrolled through other programme, which plainly invalidates the simple experimental approach as presented above.

3.2. Non-Experimental Data

Despite the above comments, non-experimental data is even more difficult to deal with and requires special care. When the control group is drawn from the population at large, even if satisfying strict comparability rules based on observable information, we cannot rule out differences on unobservables that are related to programme participation. This is the *econometric selection problem* as commonly defined, see Heckman (1979). In this case, using the estimator (3.1) results in a fundamental non-identification problem since it approximates (abstracting from other regressors in the outcome equation),

$$E\left(\hat{\alpha}_{\mathsf{ATE}}\right) = \alpha + [E(U_{\mathsf{i}\mathsf{t}} \mid d_{\mathsf{i}} = 1) - E(U_{\mathsf{i}\mathsf{t}} \mid d_{\mathsf{i}} = 0)].$$

Under selection on the unobservables, $E(U_{\mathsf{i}\mathsf{t}} d_{\mathsf{i}}) \neq 0$ and $E\left(\widehat{\alpha}_{\mathsf{ATE}}\right)$ is expected to differ from α unless, by chance, the two final r.h.s. terms cancel out. Thus, alternative estimators are needed, which motivates the methods discussed in section 4 below.

4. Methods For Non-Experimental Data

The appropriate methodology for non-experimental data depends on three factors: the type of information available to the researcher, the underlying model and the parameter of interest. Data sets with longitudinal or repeated cross-section information support less restrictive estimators due to the relative richness of information. Not surprisingly, there is a clear trade-off between the available information and the restrictions needed to guarantee a reliable estimator.

This section starts by discussing the Instrumental Variables (IV) estimator, a frequent choice when only a single cross-section is available. IV uses at least one variable that is related with the participation decision but otherwise unrelated with the outcome. Under the required conditions, it provides the required randomness in the assignment rule. Thus, the relationship between the instrument and the outcome for different participation groups identifies the impact of treatment avoiding selection problems.

If longitudinal or repeated cross-section data is available, Difference in Differences (DID) can provide a more robust estimate of the impact of the treatment (Heckman and Robb (1985, 1986)[3]). We outline the conditions necessary for DID to reliably estimate the parameter of interest and discuss a possible extension to generalise the common trends assumption.

An alternative approach is the method of matching, which can be adopted with either cross section or longitudinal data although typically detailed individual information from the before the programme period is required (see Heckman, Ichimura and Todd (1997), Rosenbaum and Rubin (1985) and Rubin (1979)). Matching deals with the selection process by constructing a comparison group of individuals with observable characteristics similar to the treated. A popular choice that will be discussed uses the probability of participation to perform matching - the so called Propensity Score Matching.

Finally, a joint DID and matching approach may significantly improve the quality of non-experimental evaluation results and is the last estimator discussed[4].

4.1. The Instrumental Variables (IV) Estimator

The IV method requires the existence of, at least, one regressor exclusive to the decision rule, Z, satisfying the two following conditions:[5]

A1: Conditional on X, Z is not correlated with the unobservables $\left(V, U^0\right)$ and $\left(V, U^1\right)$.

A2: Conditional on X, the decision rule is a non-trivial (non-constant) function of Z.

Assumption (A2) means that there is independent (from X) variation in Z that affects programme participation, or, in other words, that under a linear specification of

[3] This idea is further developed in Blundell, Duncan and Meghir (1998), Bell, Blundell and Van Reenen (1999).

[4] This is applied in Blundell, Costa Dias, Meghir and Van Reenen, 2001

[5] The time subscript is omitted from the IV analysis since we are assuming only one time period is under consideration.

the decision rule, the Z coefficient(s) is(are) non-zero. Thus, in general

$$E(D \mid X, Z) = P(D = 1 \mid X, Z) \neq P(D = 1 \mid X)$$

Assumption (A1) means that Z has no impact on the outcomes equation through the unobservable component. The only way Z is allowed to affect the outcomes is through the participation status, D. Under homogeneous treatment effects, this means Z affects the level only, while under heterogeneous treatment effects how much Z affects the outcome depends on the particular values of X. The variable(s) Z is called the instrument(s), and is a source of exogenous variation used to approximate randomised trials: it provides variation that is correlated with the participation decision but does not affect the potential outcomes from treatment directly.

4.1.1. The IV Estimator: Homogeneous Treatment Effect

Under conditions (A1) and (A2), the standard IV procedure identifies the treatment effect α using only the part of the variation in D that is associated with Z ($\widehat{\alpha}_{IV} = cov(y_i, Z_i)/cov(d_i, Z_i)$). An alternative is to use both Z and X to predict D, building a new variable $\widehat{D}$ that is used in the regression instead of D. A third possibility is directly derived by noting that, given assumption (A1) and equation (2.5),

$$E(Y \mid X, Z) = g^0(X) + \alpha P(D = 1 \mid X, Z)$$

and since, from assumption (A2), there are at least two values of Z, say z and $z + \delta$ ($\delta \neq 0$), such that $P(D = 1 \mid X, Z = z) \neq P(D = 1 \mid X, Z = z + \delta)$,

$$\begin{aligned} \alpha_{IV} &= \frac{\int_{S(X)} [E(Y \mid X, Z = z) - E(Y \mid X, Z = z + \delta)] \, dF(X \mid X \in S(X))}{\int_{S(X)} [P(D = 1 \mid X, Z = z) - P(D = 1 \mid X, Z = z + \delta)] \, dF(X \mid X \in S(X))} \\ &= \frac{E(Y \mid Z = z) - E(Y \mid Z = z + \delta)}{P(D = 1 \mid Z = z) - P(D = 1 \mid Z = z + \delta)} \end{aligned} \tag{4.1}$$

where $S(X)$ stands for the support of X.

4.1.2. The IV Estimator: Heterogeneous Treatment Effects

Depending on the assumptions one is willing to accept, the heterogeneous framework may impose additional requirements on the data for the treatment effect to be identifiable. We start from the simpler case given by the following assumption,

A3: Individuals do not use information on the idiosyncratic component of the treatment effect when deciding about participation ($\alpha_i(X) - \alpha(X)$ where $\alpha(X) = E(\alpha_i|X)$).

Assumption (A3) is satisfied if potential participants have no *a priori* information apart from the one available to the researcher (X) and decision is based on the average treatment effect for the agent's specific group. In such case,

$$E\left[U_i^1 - U_i^0 \middle| X, Z, D\right] = E\left[D_i\left[\alpha_i(X) - \alpha(X)\right] \middle| X, Z\right] = 0$$

which together with (A1) and (A2) is sufficient to identify the average treatment effect $E[\alpha_i \mid X]$. Furthermore, there is no apparent reason for it to differ from the effect of treatment on the treated, $E[\alpha_i \mid X, D_i = 1]$ for as long as the estimated parameters are conditional on the observables, X.

If, however, agents are aware of their own idiosyncratic gains from treatment, they are likely to make a more informed participation decision. Selection on the unobservables is expected, making individuals that benefit more from participation to be the most likely to participate within each X-group. Such a selection process creates correlation between $\alpha_i(X)$ and Z. This is easily understood given that the instrument impacts on D, facilitating or inhibiting participation. For example, it may be that participants with values of Z that make participation more unlikely are expected to gain on average more from treatment than participants with values of Z that make participation more likely to occur. Take the case where distance from home to the treatment location is taken as an instrument. Though in general such a variable is unlikely to be related with

outcomes such as earnings or employment probabilities, it is likely to be related with the idiosyncratic component of the treatment effect since agents living closer incur less travelling costs and are, therefore, more likely to participate even if expecting lower gains from treatment. Such a relationship between the instrument Z and the idiosyncratic gain from treatment is immediately recognised formally since, from (2.1)

$$U_i^1 - U_i^0 = \left(Y_i^1 - Y_i^0\right) - \alpha\left(X_i\right) = \alpha_i\left(X_i\right) - \alpha\left(X_i\right)$$

Thus, the error term under heterogeneous treatment effect is

$$U_i = U_i^0 + D_i\left[\alpha_i\left(X_i\right) - \alpha\left(X_i\right)\right] \tag{4.2}$$

where D is, by assumption, determined by Z depending on the gain $\alpha_i\left(X_i\right) - \alpha\left(X_i\right)$.

Under such circumstances, assumptions (A1) and (A2) are no longer enough to identify the ATE or TTE. This happens because the average outcomes of any two groups differing on the particular Z-realisations alone are different not only as a consequence of different participation rates but also because of compositional differences in the participants (non-participants) groups according to the unobservables. Thus, the main concern relates to the existence and identification of regions of the support of X and Z where changes in Z cause changes in the participation rates unrelated with potential gains from treatment.

The solution advanced by Imbens and Angrist (1994) is to identify the impact of treatment from local changes of the instrument Z. The rationale is that some local changes in the instrument Z reproduce random assignment by inducing agents to decide differently as they face different conditions unrelated to potential outcomes. To guarantee that the groups being compared are indeed comparable, Imbens and Angrist use a strengthened version of (A2),

A2': Conditional on X, the decision rule is a non-trivial monotonic function of Z.

In what follows, suppose D is an increasing function of Z, meaning that an increase in Z leads some individuals to take up treatment but no one individual to give up treatment. In an hypothetical case, where Z changes from $Z = z$ to $Z = z + \delta$ ($\delta > 0$), the individuals that change their participation decisions as a consequence of the change in Z are those that choose not to participate under $Z = z$ excluding the ones that choose not to participate under $Z = z + \delta$, or, equivalently, those that decide to participate under $Z = z + \delta$ excluding the ones that prefer participation under $Z = z$. Thus, the expected outcome under treatment and non-treatment for those affected by the change in Z can be estimated as follows,

$$E\left[Y_i^1 \mid X_i, D_i(z) = 0, D_i(z+\delta) = 1\right] =$$
$$= \frac{E\left[Y_i^1 \mid X_i, D_i(z+\delta) = 1\right] P[D_i = 1 \mid X_i, z+\delta] - E\left[Y_i^1 \mid X_i, d_i(z) = 1\right] P[D_i = 1 \mid X_i, z]}{P[D_i = 1 \mid X_i, z+\delta] - P[D_i = 1 \mid X_i, z]}$$

and

$$E\left[Y_i^0 \mid X_i, D_i(z) = 0, D_i(z+\delta) = 1\right] =$$
$$= \frac{E\left[Y_i^0 \mid X_i, D_i(z) = 0\right] P[D_i = 0 \mid X_i, z] - E\left[Y_i^0 \mid X_i, D_i(z+\delta) = 0\right] P[D_i = 0 \mid X_i, z+\delta]}{P[D_i = 1 \mid X_i, z+\delta] - P[D_i = 1 \mid X_i, z]}.$$

The estimated treatment effect is given by,

$$\begin{aligned} \alpha_{\mathsf{LATE}}(X_i, z, z+\delta) &= E(Y_i^1 - Y_i^0 \mid X_i, D_i(z) = 0, D_i(z+\delta) = 1) \\ &= \frac{E[Y_i \mid X_i, z+\delta] - E[Y_i \mid X_i, z]}{P[D_i = 1 \mid X_i, z+\delta] - P[D_i = 1 \mid X_i, z]} \end{aligned} \tag{4.3}$$

which is the Local Average Treatment Effect (LATE) parameter. To illustrate the LATE approach, take the example discussed above on selection into treatment dependent on the distance to the treatment site. Participation is assumed to become less likely the longest the distance from home to the treatment location. To estimate the treatment effect, consider a group of individuals that differ only on the distance dimension. Among those that participate when the distance Z equals z some would stop participating if at distance $z + \delta$. LATE measures the impact of the treatment on the "movers" group

by attributing any difference on the average outcomes of the two groups defined by the distance to the treatment site to the different participation frequency.[6]

Though very similar to the IV estimator presented in (4.1), LATE is intrinsically different since it does not represent TTE or ATE. LATE depends on the particular values of Z used to evaluate the treatment and on the particular instrument chosen. The group of "movers" is not in general representative of the whole treated or, even less, the whole population. For instance, agents benefiting the most from participation are more unlikely to be observed among the movers. The LATE parameter answers a different question, of how much agents at the margin of participating benefit from participation given a change in policy. That is, it measures the effect of treatment on the sub-group of treated at the margin of participating for a given $Z = z$. This is more easily seen if taking the limits when $\delta \longrightarrow 0$, as in Heckman and Vytlacil (1999),

$$\alpha_{MTE}(X_i, z) = \left. \frac{\partial E[Y \mid X_i, Z]}{\partial P[D = 1 \mid X_i, Z]} \right|_{Z=z}$$

α_{MTE} is the Marginal Treatment Effect (MTE), and is by definition the LATE parameter defined for an infinitesimal change in Z. It represents TTE for agents that are indifferent between participating and not participating at $Z = z$. All the three parameters, namely ATE, TTE and LATE, can be expressed as averages of MTE over different subsets of the Z support. The ATE is the expected value of MTE over the entire support of Z, including the values where participation is nil or universal. The TTE excludes only the subset of the Z-support where participation does not occur. Finally, LATE is defined as the average MTE over an interval of Z bounded by two values for which participation rates are different.[7]

[6] Abadie, Angrist and Imbens (1998) extend this approach to the evaluation of *quantile treatment effects*. The goal is to assess how different parts of the outcome's distribution are affected by the policy. As with LATE, a local IV procedure is used, making the estimated impacts representative only for the sub-population of individuals changing their treatment status as a consequence of the particular change in the instrument considered.

[7] The importance of the monotonic assumption depends on the parameter of interest. It is not needed if one is willing to assess the effects of a change in policy on average outcomes, which includes both changes in participation and effects of participation see (see Heckman, 1997).

4.2. The Difference In Differences (DID) Estimator

If longitudinal or repeated cross-section information is available, the additional time dimension can be used to estimate the treatment effect under less restrictive assumptions. Without loss of generality, re-write model (2.6) as follows,

$$Y_{it} = g_t^0(X_i) + \alpha_{it}(X_i) D_{it} + (\phi_i + \theta_t + \varepsilon_{it}) \tag{4.4}$$

where the error term U_{it}^0, is being decomposed on an *individual-specific fixed effect*, ϕ_i, a *common macro-economic effect*, θ_t and a *temporary individual-specific effect*, ε_{it}. The main assumption underlying the DID estimator is the following,

A4: Selection into treatment is independent of the temporary individual-specific effect, ε_{it}, so that,

$$E(U_{it}^0 \mid X_i, D_i) = E(\phi_i \mid X_i, D_i) + \theta_t$$

where D_i distinguishes participants from non-participants and is, therefore, time-independent.

Assumption (A4) is sufficient because ϕ_i and θ_t vanish in the sequential differences.[8] To see why, suppose information is available for a pre- and a post-programme periods - denoted respectively by t_0 and t_1 ($t_0 < k < t_1$). DID measures the excess outcome growth for the treated compared to the non-treated. Formally, it can be presented as follows,

$$\hat{\alpha}_{DID}(X) = \left[\bar{Y}_{t_1}^1(X) - \bar{Y}_{t_0}^1(X)\right] - \left[\bar{Y}_{t_1}^0(X) - \bar{Y}_{t_0}^0(X)\right] \tag{4.5}$$

where $\bar{Y}$ stands for the mean outcome among the specific group being considered. Under heterogeneous effects, the DID estimator recovers the TTE since

$$E(\hat{\alpha}_{DID}(X)) = E[\alpha_i(X) \mid D_i = 1] = \alpha_{TTE}(X)$$

[8] Notice that selection is allowed to occur on a temporary individual-specific effect that depends on the observables only, namely $g_t^0(X_i)$.

Where we have omitted a time subscript on $\alpha_{\mathsf{i}}(X_{\mathsf{i}})$ it refers to period t_1. In the homogeneous effect case, one may omit the covariates from equation (4.5) and average over the complete groups of treated and non-treated.

4.2.1. The DID Estimator: The Common Trends And Time Invariant Composition Assumptions

In contrast to the IV estimator, no exclusion restrictions are required under the DID methodology as there is no need for any regressor in the decision rule. Even the outcome equation may remain unspecified as long as the treatment impact enters additively. However, assumption (A4) together with the postulated specification (4.4) brings two main weaknesses to the DID approach. The first problem relates to the lack of control for *unobserved temporary individual-specific components* that influence the participation decision. If ε is not unrelated to D, DID is inconsistent and in fact approximates the following parameter,

$$E\left(\hat{\alpha}_{\mathsf{DID}}\left(X\right)\right) = \alpha_{\mathsf{TTE}}\left(X\right) + E(\varepsilon_{\mathsf{i}t_1} - \varepsilon_{\mathsf{i}t_0} \mid D_{\mathsf{i}} = 1) - E(\varepsilon_{\mathsf{i}t_1} - \varepsilon_{\mathsf{i}t_0} \mid D_{\mathsf{i}} = 0)$$

To illustrate the conditions such inconsistency might arise, suppose a training programme is being evaluated in which enrolment is more likely if a temporary dip in earnings occurs just before the programme takes place (so-called Ashenfelter's dip, see Heckman and Smith (1994)). A faster earnings growth is expected among the treated, even without programme participation. Thus, the DID estimator is likely to overestimate the impact of treatment. Moreover, if instead of longitudinal data one uses cross-section data, the problem is likely to worsen as it may extend to the fixed effect (ϕ_{i}) component: the before-after comparability of the groups under an unknown selection rule may be severely affected as the composition of the groups may change over time and be affected by the intervention, causing $E\left(\phi_{\mathsf{i}} \mid D_{\mathsf{i}}\right)$ to change artificially with t.

The second weakness occurs if the macro effect has a differential impact across the

two groups. This happens when the treatment and comparison groups have some (possibly unknown) characteristics that distinguish them and make them react differently to common macro shocks. This motivates the differential trend adjusted DID estimator that is presented below.

4.2.2. The DID Estimator: Adjusting For Differential Trends

Replace (A4) by

A4' Selection into treatment is independent of the temporary individual-specific effect, ε_{it}, under differential trends

$$E(U_{it} \mid D_i) = E(\phi_i \mid D_i) + k^D \theta_t$$

where the k^D acknowledges the differential macro effect across the two groups.

The DID estimator now identifies

$$E\left(\widehat{\alpha}_{DID}(X)\right) = \alpha_{TTE}(X) + (k^1 - k^0)[\theta_{t_1} - \theta_{t_0}] \tag{4.7}$$

which clearly only recovers the true TTE when $k^1 = k^0$.

Now suppose we take another time interval, say t_* to t_{**} (with $t_* < t_{**} < k$), over which a similar macro trend has occurred. Precisely, we require a period for which the macro trend matches the term $(k^1 - k^0)[\theta_{t_1} - \theta_{t_0}]$ in (4.7). 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 differentially adjusted estimator proposed by Bell, Blundell and Van Reenen (1999), which takes the form

$$\widehat{\alpha}_{TADID}(X) = \left\{(\widetilde{Y}^1_{t_1} - \widetilde{Y}^1_{t_0}) - (\widetilde{Y}^0_{t_1} - \widetilde{Y}^0_{t_0})\right\} - \left\{(\widetilde{Y}^1_{t_{**}} - \widetilde{Y}^1_{t_*}) - (\widetilde{Y}^0_{t_{**}} - \widetilde{Y}^0_{t_*})\right\} \tag{4.8}$$

will now consistently estimate α_{TTE}.

To illustrate this approach, lets consider the case where treatments and controls belong to different cohorts. Suppose treatments are drawn from a younger cohort,

making them more responsive to macroeconomic cycles. If the outcome of interest is affected by the macro conditions, we expect the time-specific effect to differ between treatments and controls. But if similar cyclical conditions were observed in the past and the response of the two groups has been kept unchanged, it is possible to find a past period characterised by the same differential, $\theta_{t_1} - \theta_{t_0}$.

4.3. The Matching Estimator

The third method we present is the matching approach. Like the DID, matching does not require an exclusion restriction or a particular specification of the participation decision or the outcomes equation. It also does not require the additive specification of the error term as postulated for the DID estimator. Its additional generality comes from being a non-parametric method, which also makes it quite versatile in the sense that it can easily be combined with other methods to produce more accurate estimates. The cost is paid with data: matching requires abundant good quality data to be at all meaningful.

The main purpose of matching is to re-establish the conditions of an experiment when no randomised control group is available. As we have noted, total random assignment allows for a direct comparison of the treated and non-treated, without particular structure requirements. The matching method aims to construct *the* correct sample counterpart for the missing information on the treated outcomes had they not been treated by pairing each participant with members of non-treated group. Under the matching assumption, the only remaining difference between the two groups is programme participation.

The solution advanced by matching is based on the following assumption,

A5: *Conditional independence assumption (CIA):* conditional on the set of observables X, the non-treated outcomes are independent of the participation status,

$$Y^0 \perp D \mid X$$

That is, given X the non-treated outcomes are what the treated outcomes would have been had they not been treated or, in other words, *selection occurs only on observables*. For each treated observation (Y^1) we can look for a non-treated (set of) observation(s) (Y^0) with the same X-realisation. With the matching assumption, this Y^0 constitutes the required counterfactual. Thus, matching is a process of re-building an experimental data set.

The second assumption guarantees that the required counterfactual actually exists

A6: All treated agents have a counterpart on the non-treated population and anyone constitutes a possible participant:

$$0 < P(D = 1 \mid X) < 1$$

However, assumption (A6) does not ensure that the same happens within any sample, and is, in fact, a strong assumption when programmes are directed to tightly specified groups.

Call S^* to the common support of X. Assuming (A5) and (A6), a subset of comparable observations is formed from the original sample and a consistent estimator for TTE is produced using the empirical counterpart of

$$\frac{\int_{S^*} E(Y^1 - Y^0 \mid X,\, D = 1)\, dF(X \mid D = 1)}{\int_{S^*} dF(X \mid D = 1)} \qquad \text{at a time } t > k \tag{4.9}$$

where the numerator represents the expected gain from the programme among the subset of sampled participants for whom one can find a comparable non-participant (that is, over S^*). To obtain a measure of the TTE, individual gains must be integrated over the distribution of observables among participants and re-scaled by the measure of the common support, S^*. Therefore, equation (4.9) represents the expected value of the programme effect over S^*.[9] If (A5) is fulfilled and the two populations are large enough, the common support is the entire support of both.

[9] It is simply the mean difference in outcomes over the common support, appropriately weighted by the distribution of participants.

The challenge of matching is to ensure that the 'correct' set of observables X is being used so that the observations of non-participants are what the observations of treated would be had they not participated, forming the right counterfactual and satisfying CIA. In practical terms, however, the more detailed the information is, the harder it is to find a similar control and the more restricted the common support becomes. That is, the appropriate trade-off between the quantity of information at use and the share of the support covered may be difficult to achieve. If, however, the right amount of information is used, matching deals well with potential bias. This is made clear by decomposing the treatment effect in the following way

$$E(Y^1 - Y^0 \mid X, D = 1) = \{E(Y^1 \mid X, D = 1) - E(Y^0 \mid X, D = 0)\} - \\ - \{E(Y^0 \mid X, D = 1) - E(Y^0 \mid X, D = 0)\} \tag{4.10}$$

where the latter term is the bias conditional on X. Conditional on X, the only reason the true parameter, $\alpha_{TTE}(X)$, might not be identified is selection on the unobservables.

Note, however, if one integrates over the common support S^*, two additional causes of bias can occur: non-overlapping support of X and misweighting over the common support. Through the process of choosing and re-weighting observations, matching corrects for the latter two sources of bias and selection on the unobservables is assumed to be zero.

4.3.1. The Matching Estimator: The Use Of Propensity Score

As with all non-parametric methods, the dimensionality of the problem as measured by X may seriously limit the use of matching. A more feasible alternative is to match on a function of X. Usually, this is carried out on the propensity to participate given the set of characteristics X: $P(X_i) = P(D_i = 1 \mid X_i)$ the *propensity score*. Its use is usually motivated by Rosenbaum and Rubin's result (1983, 1984), which shows that

the CIA remains valid if controlling for $P(X)$ instead of X:

$$Y^0 \perp D \mid P(X)$$

More recently, a study by Hahn (1998) shows that $P(X)$ is ancillary for the estimation of ATE. However, it is also shown that knowledge of $P(X)$ may improve the efficiency of the estimates of TTE, its value lying on the "dimension reduction" feature.

When using $P(X)$, the comparison group for each treated individual is chosen with a pre-defined criteria (established by a pre-defined measure) of proximity. Having defined the neighbourhood for each treated observation, the next issue is that of choosing the appropriate weights to associate the selected set of non-treated observations for each participant one. Several possibilities are commonly used, from a unity weight to the nearest observation and zero to the others, to equal weights to all, or kernel weights, that account for the relative proximity of the non-participants' observations to the treated ones in terms of $P(X)$.

In general the form of the matching estimator is given by

$$\widehat{\alpha}_{M} = \sum_{i \in T} \left\{ Y_i - \sum_{j \in C} W_{ij} Y_j \right\} w_i \tag{4.11}$$

where T and C represent the treatment and comparison groups respectively, W_{ij} is the weight placed on comparison observation j for individual i and w_i accounts for the re-weighting that reconstructs the outcome distribution for the treated sample.[10]

4.3.2. The Matching Estimator: Parametric Approach

Specific functional forms assumed for the g-functions in (2.1) can be used to estimate the impact of treatment on the treated over the whole support of X, reflecting the

[10] For example, in the nearest neighbour matching case the estimator becomes

$$\widehat{\alpha}_{M\,M} = \sum_{i \in T} \{Y_i - Y_j\} \frac{1}{N_T}$$

where, among the non-treated, j is the nearest neighbour to i in terms of $P(X)$. In general, kernel weights are used for W_{ij} to account for the closeness of Y_j to Y_i.

trade-off between the structure one is willing to impose in the model and the amount of information that can be extracted from the data. To estimate the impact of treatment under a parametric set-up, one needs to estimate the relationship between the observables and the outcome for the treatment and comparison groups and predict the respective outcomes for the population of interest. A comparison between the two sets of predictions supplies an estimate of the impact of the programme. In this case, one can easily guarantee that outcomes being compared come from populations sharing exactly the same characteristics.[11]

4.4. Matching and DID

The CIA is quite strong once it is acknowledged that individuals may decide according to their forecast outcome. However, by combining matching with DID there is scope for an unobserved determinant of participation as long as it lies on separable individual and/or time-specific components of the error term. To clarify the exposition, lets take model (4.4).[12] If performing matching on the set of observables X within this setting, the CIA can now be replaced by,

$$(\varepsilon_{\mathtt{t}_1} - \varepsilon_{\mathtt{t}_0}) \perp D \mid X$$

where $t_0 < k < t_1$. Since DID effectively controls for the other components of the outcomes under non-treatment, only the temporary individual-specific shock requires additional control. The main matching hypothesis is now stated in terms of the before-after evolution instead of levels. It means that controls have evolved from a pre- to a post-programme period in the same way treatments would have done had they not been

[11] If, for instance, a linear specification is assumed with common coefficients for treatments and controls, so that

$$Y = X\beta + \alpha_{\mathtt{TTE}} D + U$$

then no common support requirement is needed to estimate $\alpha_{\mathtt{TTE}}$ - a simple OLS regression using all information on treated and non-treated will consistently identify it.

[12] An extension to consider differential trends can be considered similarly to what have been discussed before.

treated.

The effect of the treatment on the treated can now be estimated over the common support of X, S^*, using an extension to (4.11),

$$\widehat{\alpha}^{LD}_{MDID} = \sum_{i \in T} \left\{ [Y_{it_1} - Y_{it_0}] - \sum_{j \in C} W_{ij} [Y_{jt_1} - Y_{jt_0}] \right\} w_i$$

where the notation is similar to what has been used before.

Quite obviously, this estimator requires longitudinal data to be applied. However, it is possible to extend it for the repeated cross-sections data case. If only repeated cross-sections are available, one must perform matching three times for each treated individual after treatment: to find the comparable treated before the programme and the controls before and after the programme. If the same assumptions apply, the TTE is identified by,

$$\widehat{\alpha}^{RCS}_{MDID} = \sum_{i \in T_1} \left\{ \left[Y_{it_1} - \sum_{j \in T_0} W^{T}_{ijt_0} Y_{jt_0} \right] - \left[\sum_{j \in C_1} W^{C}_{ijt_1} Y_{jt_1} - \sum_{j \in C_1} W^{C}_{ijt_0} Y_{jt_0} \right] \right\} w_i$$

where T_0, T_1, C_0 and C_1 stand for the treatment and comparison groups before and after the programme, respectively, and W^{G}_{ijt} represent the weights attributed to individual j in group D (where $G = C$ or T) and time t when comparing with treated individual i (for a more detailed discussion with application of the combined matching and DID estimator, see Blundell, Costa Dias, Meghir and Van Reenen, 2001).

5. Interpreting the Evidence

In this section we briefly draw on some recent studies to illustrate some of the non-experimental techniques presented in this review. The studies presented below show that the methods we have described should be carefully applied and even more cautiously interpreted (see also Blundell and Costa-Dias (2000)).

5.1. The LaLonde Study and the NSWD Evaluation

LaLonde (1986) aimed at assessing the reliability of the non-experimental techniques by comparing the results produced by these methods as commonly applied and the true parameters obtained using experimental data. This study used the National Supported Work Demonstration (NSWD), a programme operated in 10 sites across USA and designed to help disadvantaged workers, in particular women in receipt of AFDC (Aid for Families with Dependent Children), ex-drug addicts, ex-criminal offenders and high-school drop-outs. Qualified applicants were randomly assigned to treatment, which comprised a guaranteed job for 9 to 18 months. Treatment and control groups summed up to 6,616 individuals. Data on all participants were collected before, during and after treatment takes place, and earnings were the chosen outcome measure.

To assess the reliability of the experimental design, Lalonde presents pre-treatment earnings and other demographic variables for treatments and controls (males). As far as can be inferred from the observables, treatments and controls are not different before the treatment takes place. In the absence of non-random drop-outs and with no alternative treatment offered and no changes in behaviour induced by the experiment, the controls constitute the perfect counterfactual to estimate the treatment impact. An analysis of the earnings evolution for treated and controls from a pre-programme year, 1975, through the treatment periods, 1976-77, until the post-programme period, 1978, it can be seen that the treated group and control group earnings were nearly the same before treatment. They then diverged substantially during the programme and somehow converged after it. The estimated impact one year after treatment is almost $+\$900$.

To evaluate the quality of the non-experimental techniques, Lalonde applied a set of different methods using both the control group and a number of other, non-experimentally determined, comparison groups. The aim being to reproduce what the participants would have been in the absence of the programme. The comparison groups were drawn

from either the Panel Study of Income Dynamics (PSID) or from the Current Population Survey - Social Security Administration (CPS-SSA). Tables 5 and 6 of LaLonde reveal the robustness of the experimental results to the choice of estimator. They show that using comparisons from non-experimental samples significantly changes the results and raises the problem of dependence on the adopted specification for the earnings function and participation decision.

5.2. A Critique of the Lalonde Study

LaLonde's results have been criticised on the basis that the chosen non-experimental comparison groups do not satisfy the necessary requests to successfully identify the correct parameter (see Heckman, Ichimura and Todd, 1997, and Heckman, Ichimura, Smith and Todd, 1998). It is argued that to ask for identification of the true parameter from the data used in LaLonde (1986) to construct the counterfactual is to make unfair requests on data that has not been selected to truly represent what the treated would have been without treatment. Three main reasons are pointed out: First, comparisons are not drawn from the same local labour markets; Second, data on treated and comparisons were collected from different questionnaires and does not, therefore, measure the same characteristics; and Third, data are not rich enough to clearly distinguish between individuals.

A recent study by Smith and Todd (2000) is based on precisely the same data that was used by LaLonde. A careful evaluation of the bias present in non-experimental studies is performed by using a variety of methodologies and experimenting with the data. They use LaLonde's outcome variable, earnings, and directly compare non-experimental comparisons with experimental controls to obtain a measure of the bias.

This study suggests that matching may substantially improve the results when only cross-section data is available, in which case a careful choice of the matching variables is central for the quality of the estimates. When using longitudinal data, however, other

demands on information are somewhat relaxed. The quality of the estimates improves significantly and to the same order of magnitude independently of the technique or amount of information used. It seems as if much of the key information in a cross-section analysis pertains to variables that stay relatively constant over time and that cancel out on the sequential differencing that characterises DID.

5.3. A Simulation Study

To further investigate the accuracy of non-experimental methods, Heckman and Smith (1998, see also Heckman, LaLonde and Smith, 1999) ran a fully controlled experiment based on simulated data. Data were created with an individual's model of participation and earnings and subsequently used to illustrate how biased the different methods are under different underlying hypotheses. Such approach requires a structural model of individual's decisions to be established *a priori*, and the results depend on the particular specification assumed. The model considered allows for heterogeneous responses, α_i, but these are assumed to be independent from all other error components in the model. Perfect certainty is also assumed except for one case where the individual-specific gains from training, $\alpha_i - E(\alpha)$, are not known at the moment of enrolling into treatment. The model reproduces the widely discussed Ashenfelter's dip.

Given a particular choice of the parameters, the model was used to simulate the behaviour of 1000 individuals over 10 periods (from $k-5$ to $k+4$, where k is the treatment period) 100 times under different assumptions. The results of this simulation show the bias by type of estimator for unmatched and matched samples, where matching is based on earnings two periods before treatment ($k-2$). In each case, three possible estimators are considered: the simple cross-sectional differences (CS), DID using periods $k-3$ and $k+3$ and IV. The magnitude of the bias is computed under four possible underlying hypotheses about the nature of the effect (homogeneous vs heterogeneous), the amount of information available to the agent at the enrolling period and the magnitude of the

variability of α.

Under the homogeneity assumption the CS estimator is severely downward biased. DID controls for the fixed effect and significantly reduces the bias. It is, however, negatively affected by matching mainly due to the period matching takes place: controls are selected to reproduce treatments at $k-2$ but they start differentiating immediately at $k-1$ by recovering earlier in time from the characteristic dip in earnings. Finally, as expected, IV performs well and is consistent in this case. Relaxing the homogeneity assumption but allowing agents only know about $E(\alpha)$ before taking treatment, the same conclusions can be drawn. Finally, in the heterogeneous / perfect foresight case, CS and DID perform better given that selection now occurs largely on α_i. DID, in particular, shows remarkable small bias. In contrast IV performs much worse now, a feature that is not unexpected since IV is not consistent for TTE under these conditions. It does, however, consistently estimate LATE since the model satisfy Imbens and Angrist (1994) monotonicity assumption. In the unmatched case, the LATE parameter is estimated to deviate 25% from the TTE. With increased variability on α_i the performance improves for CS and DID as participation decisions are more heavily based of α_i, but LATE does not seem to get closer to TTE.

5.4. Matching and Difference in Differences: An Area Based Evaluation of the British 'New Deal' Program

The 'New Deal for the Young Unemployed' is an initiative to provide work incentives to individuals aged 18 to 24 and claiming Job Seekers Allowance (JSA) for 6 months (see Bell, Blundell and Van Reenen, 1999). The program was first introduced in January 1998, following the election of a new government in Britain in the previous year. It combines initial job search assistance followed by various subsidized options including wage subsidies to employers, temporary government jobs and full time education and training. Prior to the New Deal, young people in the UK could, in principle, claim unemployment benefits indefinitely. Now, after 6 months of unemployment, young people

enter the New Deal 'Gateway', which is the first period of job search assistance. The program is mandatory, including the subsidized options part, which at least introduces an interval in the claiming spell.

The Blundell, Costa-Dias, Meghir and Van Reenen (2001) study investigates the impact of the program on employment in the first 18 months of the scheme. In particular it exploits an important design feature by which the programme was rolled out in certain pilot areas prior to the national roll out. Since the programme is targeted at a specific age group a natural comparison group would be similar individuals with corresponding unemployment spells but who are slightly too old to be eligible. A before and after comparison can then be made using a regular DID estimator. This can be improved by a matching DID estimator as detailed in section 4.4. These estimators are all implemented in the study. The pilot area based design also means that matched individuals of the same age can be used as an alternative control group. These are not only likely to satisfy the quasi-experimental conditions more closely but also allow an analysis of the degree to which the DID comparisons within the treatment areas suffer from general equilibrium or market level biases and whether they suffer from serious substitution effects.

The evaluation approach therefore consists of exploring sources of differential eligibility and different assumptions about the relationship between the outcome and the participation decision to identify the effects of the New Deal. On the 'differential eligibility' side, two potential sources of identification are used. First, the fact that the program is age-specific implies that using slightly older people of similar unemployment duration is a natural comparison group. Second, the fact that 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). As mentioned already this provides an additional dimension to explore on the construction of the control groups, especially concerning substitution and equilibrium wage effects. Substitution occurs if

participants take (some of) the jobs that non-participants would have got in the absence of treatment. Equilibrium wage effects may occur when the program is wide enough to affect the wage pressure of eligible and ineligible individuals.

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

6. Conclusions

This paper has presented an overview of alternative empirical methods for the evaluation of policy interventions at the microeconomic level. It has focussed on social experiments, natural experiments, matching methods, and instrumental variable methods. The idea has been to describe the assumptions and data requirements of each approach and to assess the parameters of interest that they are able to estimate. The appropriate choice of evaluation method has been shown to depend on a combination of the data available and the policy parameter of interest. No one method dominates and all methods rest on heavy assumptions. Even social experiments rely on strong assumptions: they rule out spill over effects and are sensitive to non-random drop outs from the programme. Natural experiment methods, matching methods and instrumental variable methods all place tough requirements on the data and are fragile to untestable assumptions. Moreover, with heterogeneous response parameters, they each estimate different aspects of the programme impact. It is essential to have a clear understanding of the assumptions and data requirements involved in each method before undergoing

an evaluation.

References

[1] Abadie, A., Angrist, J. and Imbens, G. (1998). "Instrumental Variables Estimation of Quantile Treatment Effects", NBER working paper 229, forthcoming in Econometrica.

[2] Ashenfelter, O. (1978), "Estimating the Effect of Training Programs on Earnings", Review of Economics and Statistics, 60, 47-57.

[3] 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.

[4] Bassi L. (1983), "The Effect of CETA on the Post-Program Earnings of Participants", Journal of Human Resources, Fall 1983, 18, 539-556.

[5] Bassi, L. (1984), "Estimating the Effects of Training Programs with Nonrandom Selection", Review of Economics and Statistics, 66, 36-43.

[6] 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.

[7] Blundell, R., Costa Dias, M., Meghir, C. and Van Reenen, J. (2001), "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program", IFS working paper W01/20 (www.ifs.org.uk)

[8] Blundell, R. and M. Costa Dias (2000), "Evaluation Methods for Non-Experimental Data", Fiscal Studies, December 2000, vol. 21, no. 4, pp. 427-468.

[9] Blundell, R., Dearden, L. and Meghir, C. (1996), "The Determinants and Effects of Work-Related Training in Britain", London: Institute for Fiscal Studies.

[10] Blundell, R., Duncan, A. and Meghir, C. (1998), "Estimating Labour Supply Responses using Tax Policy Reforms", Econometrica, 66, 827-861.

[11] 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.

[12] Burtless, G. (1985), "Are Targeted Wage Subsidies Harmful? Evidence from a Wage Voucher Experiment", Industrial and Labor Relations Review, 39, 105-114.

[13] Card, D. and Robins, P. K. (1998), "Do Financial Incentives Encourage Welfare Recipients To Work?", Research in Labor Economics, 17, 1-56.

[14] Cochrane, W. and Rubin, D. (1973), "Controlling Bias in Observational Studies", Sankyha, 35, 417-446.

[15] 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.

[16] Eissa, N. and Liebman, J. (1996), "Labor Supply Response to the Earned Income Tax Credit", Quarterly Journal of Economics, CXI, 605-637.

[17] Fan, J. (1992), "Design Adaptive Nonparametric Regression", Journal of the American Statistical Association, 87, 998-1004.

[18] Fisher, R. (1951), The Design of Experiments, 6th edition, London: Oliver and Boyd.

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

[20] Hausman, J.A. and Wise, D.A. (1985), Social Experimentation, NBER, Chicago: University of Chicago Press.

[21] Heckman, J. (1979), "Sample Selection Bias as a Specification Error", Econometrica, 47, 153-61.

[22] Heckman, J. (1990), "Varieties of Selection Bias", American Economic Review, 80, 313-318.

[23] Heckman, J. (1992), "Randomization and Social program", *in* C. Manski and I. Garfinkle, eds., Evaluating Welfare and Training Programs, Cambridge, Mass: Harvard University Press.

[24] Heckman, J. (1996), "Randomization as an Instrumental Variable Estimator", Review of Economics and Statistics, 56,336-341.

[25] 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.

[26] Heckman, J. and Hotz, V.J. (1989), "Choosing among Alternatives Nonexperimental Methods for Estimating the Impact of Social programs", Journal of the American Statistical Association, 84, 862-874.

[27] Heckman, J., Ichimura, H. and Todd, P. (1997), "Matching as an Econometric Evaluation Estimator", Review of Economic Studies, 64, 605-654.

[28] Heckman, J., Ichimura, H., Smith, J. and Todd, P. (1997), "Characterising Selection Bias Using Experimental Data", Econometrica, 66(5), 1017-1098.

[29] 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.

[30] Heckman, J. and Robb, R. (1985), "Alternative Methods for Evaluating the Impact of Interventions", *in* Longitudinal Analysis of Labour Market Data, New York: Wiley.

[31] 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.

[32] Heckman, J. and Smith, J. (1994), "Ashenfelter's Dip and the Determinants of Program Participation", University of Chicago, mimeo.

[33] Heckman, J. and Smith, J. (1998), "The Sensitivity of Non-Experimental Evaluation Estimators: A Simulation Study", University of Chicago, mimeo.

[34] Heckman, J., Smith, J. and N. Clements, (1997), "Making the Most out of Program Evaluations and Social Experiments: Accounting for Heterogeneity in program Impacts", Review of Economic Studies, 64, 487-536.

[35] 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, Vol. 96, No. 8, pp 4730-4734.

[36] Imbens, G. and Angrist, J. (1994), "Identification and Estimation of Local Average Treatment Effects", Econometrica, 62(2), 467-75.

[37] Kemple, J., Dolittle, F. and Wallace, J. (1993), "The National JTPA Study: Site Characteristics in participation patterns", New York: Manpower Demonstration Research Corporation.

[38] LaLonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", American Economic Review, 76, 604-620.

[39] 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.

[40] Rosenbaum, P. and Rubin, D.B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", Biometrika, 70, 41-55.

[41] Rosenbaum, P. and Rubin, D.B. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score", Journal of the American Statistical Association, 79, 516-524.

[42] Rosenbaum, P. and Rubin, D.B. (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score", American Statistician, 39-38.

[43] Rubin, D.B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization", Annals of Statistics, 7, 34-58.

[44] Rubin, D.B. (1979), "Using Multivariate Matched Sampling and regression Adjustment to Control Bias in Observational Studies", Journal of the American Statistical Association, 74, 318-329.

[45] Sianesi, B. (2001), "An evaluation of the Swedish system of active labour market programmes in the 1990s ", IFS WP W02/01

[46] Smith, J. and Todd, P. (2000), "Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators?", unpublished manuscript.