Fiscal Studies (2000) vol. 21, no. 4, pp. 427–468

Evaluation Methods for Non-Experimental Data

RICHARD BLUNDELL and MONICA COSTA DIAS[*]

Abstract

This paper presents a review of non-experimental methods for the evaluation of social programmes. We consider matching and selection methods and analyse each for cross-section, repeated cross-section and longitudinal data. The methods are assessed drawing on evidence from labour market programmes in the UK and in the US.

JEL classification: J38, H3, C2.

I. AN OVERVIEW OF THE EVALUATION PROBLEM

The evaluation problem of concern here is the measurement of the impact of a policy reform or intervention — for example, a childcare subsidy or a targeted training programme — on a set of well-defined outcome variables. For the former example intervention, the outcome variables might include the child's exam results or the mother's labour market participation, while, for the latter, they could include individual employment durations, earnings and/or unemployment durations. Usually, individuals are identified by some observable type — for example, gender, age, education, location or marital status. The evaluation problem, therefore, is to measure the impact of the programme on

[*] University College London and Institute for Fiscal Studies.

The authors thank Costas Meghir, Barbara Sianesi, John Van Reenen, the editor and a referee for helpful comments. This review was originally prepared for the Department for Education and Employment and is also part of the programme of research of the Economic and Social Research Council (ESRC) Centre for the Microeconomic Analysis of Fiscal Policy at the Institute for Fiscal Studies. Co-funding from the Leverhulme Trust is gratefully acknowledged. The second author acknowledges 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.

each type of individual. It can be regarded as a missing-data problem since, at a moment in time, each person 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, there would be no evaluation problem. Thus constructing the counterfactual is the central issue that evaluation methods address. There are many references in the literature that document the development of the analysis of the evaluation problem. In the labour market area, from which we draw heavily in this review, the original papers that use longitudinal data are those by Ashenfelter (1978), Ashenfelter and Card (1985) and Heckman and Robb (1985 and 1986).

Evaluation methods in empirical economics fall into five broad and related categories. Implicitly, each provides an alternative approach to constructing the counterfactual. The first is the pure randomised social experiment. In many ways, this is the most convincing method of evaluation since there is a control (or comparison) group which is a randomised subset of the eligible population. The literature on the advantages of experimental data was developed in papers by Bassi (1983 and 1984) and Hausman and Wise (1985) which 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 study of the Canadian Self-Sufficiency Project reported in Card and Robins (1998), the labour supply responses of approximately 6,000 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, experiments have their own drawbacks. First, they are rare in economics and typically expensive to implement. Second, they are not amenable to extrapolation. That is, they cannot easily be used in the *ex ante* analysis of policy reform proposals. Finally, they require the control group to be completely unaffected by the reform, typically ruling out spillover, substitution 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 with those obtained from experimental data can help assess appropriate methods where experimental data are 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.

A second popular method of evaluation is the so-called natural experiment. This approach typically 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.

Under certain conditions, this approach can be used to recover the average effect of the programme on those individuals who 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 *common time effects across groups* and *no composition changes within each group*.[1] Together, these assumptions make choosing a comparison group extremely difficult. For example, in their heavily cited evaluation study of the impact of Earned Income Tax Credit reforms on the employment of single mothers in the US, Eissa and Liebman (1996) use single women without children as a control group. However, this comparison can be criticised for not capturing differential macro effects. In particular, this control group is already working to a very high level of participation in the US labour market (around 95 per cent) 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.

A third approach is the matching method. This has a long history in non-experimental statistical evaluation (see the references in Heckman, Ichimura and Todd (1997)). The aim of matching is simple. It is to select sufficient observable factors that any two individuals with the same values of these factors will display no systematic differences in their reactions to the policy reform. Consequently, if each individual undergoing the reform can be matched with an individual with the same matching variables who has not undergone the reform, the impact of the reform on individuals of that type can be measured. As in the choice of control group in a natural experiment, it is a matter of faith as to whether the appropriate matching variables have been chosen. If they have 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 fourth approach is the selection model. Developed by Heckman (1979), it was fully integrated into the evaluation literature in Heckman and Robb (1985 and 1986). This approach relies on an exclusion restriction, which requires a variable that determines participation in the programme but not the outcome of the programme itself. In contrast to matching, which can be considered as

[1] See Blundell, Duncan and Meghir (1998) for a precise description of these conditions.

'selection on the observables', the Heckman approach accounts for selection on the unobservables. A comparison of these two approaches turns out to be extremely informative in understanding the advantages and drawbacks of these methods.

The final approach is the structural simulation model. This approach is closely related to the selection model and has long been at the centre of tax reform evaluation where behaviour can often be reasonably modelled by some rational choice framework (see Blundell and MaCurdy (1999) for a review). It has the advantage of separating preferences from constraints and can therefore be used to simulate new policy reforms that change the constraint while leaving preferences unaffected. Moreover, this approach can feed into some overall general equilibrium evaluation. However, these models require a believable behavioural model for individuals, something the experimental and quasi-experimental approaches ignore by design.

Appropriate evaluation methods therefore depend on several overall criterion: (i) the nature of the programme — that is, whether it is local or national, small-scale or 'global'; (ii) the nature of the question to be answered — that is, the overall impact, the effect of treatment on the treated or the extrapolation to a new policy reform; and (iii) the nature of the data available. With regard to the nature of the data, there are a number of issues. Does the dataset contain information for individuals before and after their programme participation? Are similar questionnaires administered to potential comparison groups or are we to use other survey data to construct comparisons? In some studies, comparison groups are chosen in the same location and asked to respond to the same questionnaire as those in the programme. In other studies, a comparison group has to be drawn from individuals who are much less likely to be similar. This turns out to be critical in the implementation of matching methods, which we discuss in detail below.

This paper is organised as follows. Our aim is to discuss evaluation methods when experimental data are not available. We outline the measurement problem in Section II and consider the types of data and their implication for the choice of evaluation method in Section III. Section IV is the main focus of this paper as it presents a detailed comparison of alternative methods of evaluation for non-experimental data. In Section V, we illustrate these methods drawing on recent applications in the evaluation literature. Section VI concludes.

II. WHAT ARE WE TRYING TO MEASURE?

An important decision to be made when evaluating the impact of a programme is whether to assume homogeneous or heterogeneous treatment effects. Typically, we do not expect all individuals to respond to a policy intervention in exactly the same way. That is, there will be heterogeneity in the impact across individuals. Consequently, there are two possible questions that evaluation methods attempt to answer. The first is the measurement of the impact of the programme on individuals of a particular type as if they were assigned to such a programme randomly from the population of *all* people of that type. The second is the impact on individuals of a particular type among those who *were* assigned to the programme.

Under the assumption of homogeneous treatment effects, these two measures are identical, but this is not so when treatment effects can vary. In this case, the latter measure is often referred to as the 'effect of treatment on the treated'.

1. Homogeneous Treatment Effects

To make things more precise, suppose there is a policy reform or intervention 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*, and on a dummy variable, *d*, such that $d_i = 1$ if individual *i* has participated in the programme and $d_i = 0$ otherwise. For ease of exposition, we will assume that the programme takes place in period *k*, so that, in each period *t*,

$$(1) \qquad \begin{array}{ll} Y_{it} = X_{it}\beta + d_i\alpha + U_{it} & \text{if} \quad t > k \\ Y_{it} = X_{it}\beta + U_{it} & \text{if} \quad t \leq k, \end{array}$$

where α measures the homogeneous impact of treatment for individual *i*.[2] The set of parameters β in (1) define the relationship between the exogenous variables *X* and the dependent variable *Y*, and U_{it} is the error term of mean zero, which is assumed to be uncorrelated with *X*.

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_i — and the error term in the outcome equation — U_{it}. This happens because an individual's participation decision is probably based on personal 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, *U*, and the participation variable, *d*, is expected. In such case, the standard econometric approach, which would regress *Y* on a set of regressors including *d*, is not valid.

We assume that the participation decision can be parametrised in the following way. For each individual, there is an index, *IN*, depending on a set of

[2] In most of what follows, we will assume a linear specification of the outcome equation. However, this is relaxed when dealing with non-parametric estimators, as in the case of the general matching estimator described in Section IV(4).

variables Z and parameters γ, for which enrolment occurs when this index rises above zero. That is,

$$IN_i = Z_i\gamma + V_i, \tag{2}$$

where V_i is the error term and

$$\begin{aligned} &d_i = 1 \quad \text{if} \quad IN_i > 0 \\ &d_i = 0 \quad \text{otherwise.} \end{aligned} \tag{3}$$

2. Heterogeneous Treatment Effects

However, it seems reasonable to assume that the treatment impact varies across individuals. Naturally, these differentiated effects should also influence the decision process and so are likely to be correlated with the treatment indicator, d_i.

Abstracting from other regressors, *X*, the outcome equation takes the form (when $t > k$)

$$Y_{it} = \beta + d_i\alpha_i + U_{it}, \tag{4}$$

where α_i is the treatment impact on individual *i*. Define $\bar{\alpha}$ as the population mean impact, ε_i as worker *i*'s deviation from the population mean and α_T as the mean impact of treatment on the treated. Thus

$$\begin{aligned} &\alpha_i = \bar{\alpha} + \varepsilon_i \\ &\alpha_T = \bar{\alpha} + E(\varepsilon_i \mid d_i = 1), \end{aligned} \tag{5}$$

where $E(\varepsilon_i \mid d_i = 1)$ stands for the mean deviation of the impact among participants. The outcome regression equation may now be rewritten in the following way:

$$Y_{it} = \beta + d_i\bar{\alpha} + [U_{it} + d_i\varepsilon_i] = \beta + d_i\bar{\alpha} + [U_{it} + d_i(\alpha_i - \bar{\alpha})]. \tag{6}$$

Obviously, the additional problem with this heterogeneous specification of treatment effects concerns the form of the error term, $U_{it} + d_i(\alpha_i - \bar{\alpha})$. This can be seen to differ across observations according to the treatment status of each individual — as d_i assumes the values 0 and 1. The identification of the parameter $\bar{\alpha}$ is more difficult in the case of non-zero correlation with the treatment indicator. Notice that if $E(\varepsilon_i d_i) \neq 0$, we should have $E(\varepsilon_i \mid d_i) \neq 0$,[3] and thus

$$(7) \qquad E(Y_{it} \mid d_i) = \beta + d_i[\bar{\alpha} + E(\varepsilon_i \mid d_i)] + E(U_{it} \mid d_i).$$

In this case, the ordinary least squares (OLS) estimator identifies

$$(8) \qquad E(\hat{\alpha}) = \bar{\alpha} + E(\varepsilon_i \mid d_i = 1) + E(U_{it} \mid d_i = 1) - E(U_{it} \mid d_i = 0).$$

Consequently, even if U_{it} is uncorrelated with d_i, so that $E(U_{it} \mid d_i = 1) = E(U_{it} \mid d_i = 0) = 0$, an identification problem remains. It is clear from (8) that, without further assumptions or information, only the impact of treatment on the treated, $\alpha_T = \bar{\alpha} + E(\varepsilon_i \mid d_i = 1)$, is identifiable. This is because, even if the error term, *U*, is uncorrelated with the decision process, the individual-specific component of the treatment effect, *ε*, is most likely not to be. We expect individuals to decide taking into account their own specific conditions, in which case $E(\varepsilon_i \mid d_i = 1) \neq 0$ and the identification of $\bar{\alpha}$ becomes more difficult.

III. EXPERIMENTAL AND NON-EXPERIMENTAL DATA

1. Experimental Data

As mentioned above, experimental data provide the correct missing counterfactual, eliminating the evaluation problem. The contribution of experimental data is to rule out self-selection (according to observables or unobservables) as a source of bias. In fact, as individuals are randomly assigned to the programme, a decision process such as the one described in Section II is ruled out.

Let us suppose, for example, that an experiment is conducted and that a random sample from a group of eligible individuals is chosen to participate in a programme; these are administered the treatment. Within that target group, assignment to treatment is completely independent of a possible outcome variable, which is to say that it is independent of the treatment effect. If no side-effects exist, the comparison group composed of the non-treated is statistically equivalent to the treated group in all respects except treatment status. In the case

[3] This is because, by iterated expectations,

$$E(\varepsilon_i d_i) = E_d[E(\varepsilon_i d_i \mid d_i)] = E_d[E(\varepsilon_i \mid d_i = 1)] = \text{Prob}(d_i = 1)E(\varepsilon_i \mid d_i = 1)$$

and, by construction,

$$E(\varepsilon_i) = \text{Prob}(d_i = 1)E(\varepsilon_i \mid d_i = 1) + \text{Prob}\,(d_i = 0)E(\varepsilon_i \mid d_i = 0) = 0,$$

which means that, in general, $E(\varepsilon_i \mid d_i) \neq 0$.

of homogeneous treatment effects, where the α_i are the same for all *i*, the impact of treatment can be easily measured by a simple subtraction of mean outcomes:

$$(9) \qquad \hat{\alpha} = \overline{Y}_t^{(1)} - \overline{Y}_t^{(0)}, \quad t > k,$$

where $\overline{Y}_t^{(1)}$ and $\overline{Y}_t^{(0)}$ are, respectively, the treated and non-treated mean outcomes at a time *t* after the programme.

However, some factors associated with the experimental design may invalidate this ideal setting. It is likely that some drop-out occurs, especially among the experimental controls. If this process is not random, it will alter the fundamental characteristic of experimental data. An idea of the importance of this non-random selection may be obtained by comparing the observable characteristics of both the control and treatment groups. This comparison ensures random assignment, at least with respect to the observables. If the non-treated are offered other treatment programmes, further differentiating factors are introduced and the comparison of means in (9) is unable to identify the treatment effect. Finally, other factors may change the behaviour of experiment participants, such as the experiment itself when selecting treated and non-treated. This also invalidates the consistency of such an estimator in an experimental framework.

2. Non-Experimental Data

Despite the above comments, non-experimental data are even more difficult to deal with and require special care. Imagine a dataset composed of a treatment group from a given programme and a comparison group drawn from the population at large. Even when the choice of the comparison group obeys the strict comparability rules based on observable information, which is frequently quite hard or even impossible to guarantee, we cannot be sure about the absence of differences in unobservables that are related to programme participation. This is the ‘econometric selection problem’, as commonly defined. In this case, using the estimator (9) results in a fundamental ‘non-identification problem’. Abstracting from other regressors in the outcome equation, for large samples the estimator identifies

$$(10) \qquad E(\hat{\alpha}) = \alpha + [E(U_{it} \mid d_i = 1) - E(U_{it} \mid d_i = 0)].$$

In the case where $E(U_{it}d_i) \neq 0$, unless the terms in the square brackets cancel out, this expectation will differ from *α*. Thus alternative estimators are needed. This motivates the methods we will focus on below in Section IV: instrumental variables, selection, difference-in-differences and matching methods.

3. An Example: The LaLonde Study

To highlight the distinction between experiments and non-experiments, we briefly consider the study by LaLonde (1986). This used an experiment dataset to compare between experimentally and non-experimentally determined results and between different types of non-experimental estimation methodologies. The programme the study is based on is called National Supported Work Demonstration (NSWD). This was operated in 10 sites across the US and was 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 nine to 18 months. Treatment and control groups totalled 6,616 individuals. Data on all participants were collected before, while and after treatment took place, and earnings were the chosen outcome measure.

To assess the reliability of the experimental design, pre-treatment earnings and other demographic variables for male treatments and controls are presented in Table 1 (see also LaLonde (1986)). It can be seen that there are no significant differences to be found between these two groups: they were statistically equivalent in terms of observables, at least at the start of the programme. In the absence of non-random drop-out 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.

Table 2 shows the earnings evolution for treatments and controls from a pre-programme year (1975), through the treatment period (1976–77), until the post-programme period (1978). It can be seen that the treatments' and controls' earnings were nearly the same before treatment, diverged substantially during the

TABLE 1

Comparison of Treatments and Controls: Characteristics for the NSWD Males

	Treatments	*Controls*
Age	24.49	23.99
Years of school	10.17	10.17
Proportion high-school drop-outs	0.79	0.80
Proportion married	0.14	0.13
Proportion black	0.76	0.75
Proportion Hispanic	0.12	0.14
Real earnings one year before treatment[a]	1,472	1,558
Real earnings two years before treatment[a]	2,860	3,030
Hours worked one year before treatment	278	274
Hours worked two years before treatment	458	469
Number of observations	2,083	2,193

[a] Annual earnings in US dollars.

TABLE 2

Annual Earnings of Male Treatments and Controls

	Treatments	*Controls*
1975	3,066	3,027
1976	4,035	2,121
1977	6,335	3,403
1978	5,976	5,090
Number of observations	297	425

programme and converged somewhat after it. The estimated impact one year after treatment is almost +$900.

Another interesting feature of the experimental data is the robustness to the choice of estimator. Table 3 (see Tables 5 and 6 of LaLonde (1986)) includes a set of estimates obtained using the control group and a number of other constructed comparison groups and based on different specifications that result in different estimation techniques. The choice of the 'non-experimentally

TABLE 3

Estimated Treatment Effects for the NSWD Male Participants using the Control Group and Comparison Groups from the PSID and the CPS-SSA

Comparison group	*Unadjusted difference of mean post-programme earnings*	*Adjusted difference of mean post-programme earnings*	*Unadjusted difference-in-differences*	*Adjusted difference-in-differences*	*Two-step estimator*
Controls	886	798	847	856	889
PSID 1	–15,578	–8,067	425	–749	–667
PSID 2	–4,020	–3,482	484	–650	—
PSID 3	697	–509	242	–1,325	—
CPS-SSA 1	–8,870	–4,416	1,714	195	213
CPS-SSA 2	–4,095	–1,675	226	–488	—
CPS-SSA 3	–1,300	224	–1,637	–1,388	—

Definitions:
PSID 1 — all male household heads continuously in the period studied (1975–78) who were less than 55 years old and did not classify themselves as retired in 1975.
PSID 2 — all men in PSID 1 not working when surveyed in the spring of 1976.
PSID 3 — all men in PSID 1 not working when surveyed in either the spring of 1975 or the spring of 1976.
CPS-SSA 1 — all males based on Westat's criterion except those over 55 years old.[4]
CPS-SSA 2 — all males in CPS-SSA 1 who were not working when surveyed in March 1976.
CPS-SSA 3 — all males in CPS-SSA 1 who were unemployed in 1976 and whose income in 1975 was below the poverty level.

[4] Westat's criterion selects individuals who were in the labour force in March 1976 with nominal income less than $20,000 and household income less than $30,000.

determined' comparison group is quite important, given the goal of reproducing the experimental setting as closely as possible. The aim is therefore to construct optimally a group of non-participants that closely reproduces what the participants would have been without the programme — which the group of controls is assumed to represent (experimental data). Given the observed characteristics, the comparison groups were drawn either from the Panel Study of Income Dynamics (those designated by PSID) or from the Current Population Survey, Social Security Administration (those designated by CPS-SSA).

Using comparisons from non-experimental control samples not only appears to change the results significantly but also raises the problem of dependence on the adopted specification for the earnings function and participation decision. We now turn to a general discussion of non-experimental methods in homogeneous and heterogeneous treatment effect models.

IV. 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. Datasets 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.

Two estimators will be considered when only a single cross-section is available — namely, the instrumental variables (IV) and the two-step Heckman selection estimators. The IV method uses at least one variable that is related to the participation decision but otherwise unrelated to the outcome. It provides the required randomness in the assignment rule since the instrument is assumed to be in no way related to the outcome except through participation. Thus the relationship between the instrument and the outcome for different participation groups identifies the impact of treatment avoiding selection problems. The Heckman selection estimator is a two-step method that uses an explicit model of the selection process to control for the part of the participation decision that is correlated with the error term in the outcome equation.

If the available data are in a longitudinal or repeated cross-section format, difference-in-differences (diff-in-diffs) can provide a more robust estimate of the impact of the treatment. We will outline the conditions necessary for diff-in-diffs to estimate the impact parameter of interest reliably. In particular, we will also suggest an extension to overcome the common trends assumption. This assumption, which is crucial for the consistency of the estimator, states that the treatment and comparison groups are affected in the same way by macro shocks. This, of course, is often difficult to justify for comparison groups chosen from non-experimental data.

An alternative approach is the method of matching, which can be adopted with either cross-section or longitudinal data, although typically detailed individual information is required from before and after the programme for both the participant group and the non-participant comparison group. It will be shown that, with sufficiently detailed data, a simple 'propensity score' method of matching can often produce quite reasonable results. Matching deals with the selection process by constructing a comparison group of individuals with observable characteristics similar to those of the treated. One way of doing this is to model the probability of participation, estimate its value for each individual (called the propensity score) and match individuals with similar propensity scores. As will be explained below, a non-parametric propensity score approach to matching that combines this method with diff-in-diffs has the potential to improve the quality of non-experimental evaluation results significantly.

For each estimator, we will discuss its ability to identify the treatment impact in a homogeneous and a heterogeneous environment, as well as other specific advantages and disadvantages. The cross-section methodologies are introduced in the first two subsections: first the IV estimator is presented and then the Heckman selection estimator (Heckman, 1979). Subsection 3 discusses the diff-in-diffs approach and potential extensions when the common macro trends restriction does not hold. In subsection 4, we present the standard matching method and extensions to more refined techniques, such as the use of propensity scores to match and the use of diff-in-diffs along with matching.

1. The Instrumental Variables (IV) Estimator

Consider, first, the 'homogeneous treatment effect' case. The IV method requires the existence of at least one regressor exclusive to the decision rule, Z^*, satisfying the following three conditions: first, Z^* determines programme participation — that is, it has a non-zero coefficient in the decision rule; second, we can find a transformation, g, such that $g(Z^*)$ is uncorrelated with the error, U, given the exogenous variables, X; finally, Z^* is not completely (or almost) determined by X. The variable(s) Z^* is/are called the instrument(s), and it 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. Under the above conditions, the standard IV procedure may be applied, replacing the treatment indicator by $g(Z^*)$ and running a regression. An alternative is to use both Z^* and X to predict d, building a new variable, $\hat{d}$, which is used in the regression instead of d.

This is a very simple estimator but it suffers from two main drawbacks. The first concerns the instrument choice. In the 'treatment evaluation problem', it is not easy to think of a variable that satisfies all the three assumptions required to identify α. The difficulty lies, mainly, in the simultaneous requirements of ‘participation determination’ and ‘non-influence on the outcome of participation’. A commonly proposed solution, possible when longitudinal or past data are available, is to consider lagged values of some determinant variables. However, they are likely to be strongly correlated with future values, included in the outcome regression, and hence this is unlikely to solve the problem.

The second issue becomes clear when trying to evaluate the impact of training in a heterogeneous framework. To understand why, recall that, from (6), the error term is given by

$$U_{it} + d_i \varepsilon_i = U_{it} + d_i(\alpha_i - \bar{\alpha}). \tag{11}$$

It is now evident that, even if Z_i^* is uncorrelated with U_{it}, the same is not true with respect to $U_{it} + d_i(\alpha_i - \bar{\alpha})$ because Z_i^* determines d_i by assumption. The violation of this fundamental hypothesis invalidates the application of the IV methodology in a heterogeneous framework.

2. The Heckman Selection Estimator

This method is more robust than the IV estimator but also more demanding on assumptions about the structure of the model. As above, the simpler ‘homogeneous treatment effect’ case will be considered first. The main assumption required to guarantee reliable estimates of the treatment effect is the existence of at least one additional regressor in the decision rule. This regressor is required to have a non-zero coefficient in the decision rule equation and to be independent of the error term, *V*. Moreover, knowledge of or ability to estimate consistently the joint density of the distribution of the errors U_{it} and V_i — $h_i(U_{it}, V_i)$, say — is required. The rationale of this estimator is to control directly for the part of the error term in the outcome equation that is correlated with the participation dummy variable. The procedure uses two steps. In the first, the part of the error term U_{it} that is correlated with d_i is estimated. It is then included in the outcome equation and the effect of the programme is estimated in a second step. Of course, by construction, what remains of the error term in the outcome equation is not correlated with the participation decision.

Take, for example, the popular special case where U_{it} and V_i are assumed to follow a joint normal distribution. Adopting the standardisation $\sigma_V = 1$, we may now write the conditional outcome expectation as

$$(12)\qquad \begin{aligned} E(Y_{it} \mid d_i = 1) &= \beta + \alpha + \rho \frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} \\ E(Y_{it} \mid d_i = 0) &= \beta - \rho \frac{\phi(Z_i\gamma)}{1-\Phi(Z_i\gamma)}, \end{aligned}$$

where the last term on the right-hand side of each equation represents the expected value of the error term conditional on the participation variable, d_i. This is precisely what is missing from (1) when assignment to treatment is non-random, as described in subsection II(1). This new regressor deals with the part of the error term that is correlated with the decision process. By including it in the outcome equation, we are able to separate the true impact of treatment from the selection process, which accounts for the differences between participants and non-participants. Thus it is possible to estimate α, the Heckman selection estimator for the selection model, by replacing γ with $\hat{\gamma}$ (obtained from regressing *IN* on *Z*) and running a least squares regression on (12).[5]

(a) The Heckman Selection Estimator: Choice-Based Samples

One advantage of the two-step procedure in the 'homogeneous treatment effect' case relates to its robustness to choice-based sampling. This is the kind of non-randomness obtained when drawing the comparison group (non-treated) from the population. Usually, the sample proportion of treated (p_t^*) differs from the population one (p_t). The treated are likely to be over-represented in the sample, resulting in a non-zero expectation of the outcome error term:

$$(13)\qquad p_t^* E(U_{it} \mid d_i = 1) + (1 - p_t^*) E(U_{it} \mid d_i = 0) \neq 0.$$

Robustness is achieved by controlling for the part of U_{it} that is correlated with d_i. In fact, since the remaining error is orthogonal to d_i, it is unaffected by this type of stratification.

(b) The Heckman Selection Estimator: Heterogeneous Treatment Effects

Now suppose that the treatment impact differs across agents. The outcome equation becomes

$$(14)\qquad Y_{it} = \beta + \alpha_T d_i + \{U_{it} + d_i[\varepsilon_i - E(\varepsilon_i \mid d_i = 1)]\} = \beta + \alpha_T d_i + \xi_{it}.$$

[5] For a more detailed description of this estimator, see the Appendix.

The two-step procedure requires knowledge of the joint density of U_{it}, V_i and ε_i. Continuing to assume a joint normal distribution ($\sigma_V = 1$),

$$(15)\quad \begin{aligned} E(\xi_{it} \mid d_i = 1) &= \text{Corr}(U_{it} + \varepsilon_i, V_i)\text{Var}(U_{it} + \varepsilon_i)^{1/2} \frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} = \rho_{(U,V,\varepsilon)} \frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} \\ E(\xi_{it} \mid d_i = 0) &= \text{Corr}(U_{it}, V_i)\text{Var}(U_{it})^{1/2} \frac{-\phi(Z_i\gamma)}{1-\Phi(Z_i\gamma)} = \rho_{(U,V)} \frac{-\phi(Z_i\gamma)}{1-\Phi(Z_i\gamma)}. \end{aligned}$$

Hence the outcome regression equation is

$$(16)\quad Y_{it} = \beta + d_i \left[\alpha_T + \rho_{(U,V,\varepsilon)} \frac{\phi(Z_i\hat{\gamma})}{\Phi(Z_i\hat{\gamma})} \right] + (1-d_i)\rho_{(U,V)} \frac{-\phi(Z_i\hat{\gamma})}{1-\Phi(Z_i\hat{\gamma})} + \delta_{it},$$

which consequently identifies α_T.

However, this method is unable to identify $\bar{\alpha}$, the effect of training if individuals were randomly assigned to treatment. In fact, if $\bar{\alpha}$ is the parameter of interest, the appropriate equation is

$$(17)\quad Y_{it} = \beta + \bar{\alpha} d_i + (U_{it} + d_i \varepsilon_i) = \beta + \bar{\alpha} d_i + \eta_{it}.$$

Notice that the error term for the treated no longer has a zero expectation. Formally,

$$(18)\quad \begin{aligned} E(\eta_{it} \mid d_i = 1) &= E(U_{it} + d_i\varepsilon_i \mid d_i = 1) = E(\varepsilon_i \mid d_i = 1) + \rho_{(U,V,\varepsilon)} \frac{\phi(Z_i\gamma)}{\Phi(Z_i\gamma)} \\ E(\eta_{it} \mid d_i = 0) &= E(U_{it} \mid d_i = 0) = \rho_{(U,V)} \frac{-\phi(Z_i\gamma)}{1-\Phi(Z_i\gamma)}. \end{aligned}$$

Therefore the outcome equation is given by

$$(19)\quad \begin{aligned} Y_{it} &= \beta + d_i \left[\bar{\alpha} + E(\varepsilon_i \mid d_i = 1) + \rho_{(U,V,\varepsilon)} \frac{\phi(Z_i\hat{\gamma})}{\Phi(Z_i\hat{\gamma})} \right] \\ &\quad + (1-d_i)\rho_{(U,V)} \frac{-\phi(Z_i\hat{\gamma})}{1-\Phi(Z_i\hat{\gamma})} + \delta_{it}, \end{aligned}$$

which is exactly the same equation as the one obtained when trying to estimate α_T. That is, only the treatment-on-the-treated impact is identifiable.

3. The Difference-in-Differences (Diff-in-Diffs) Estimator

If longitudinal or repeated cross-section information is available, it is possible to estimate the treatment effect consistently without having to impose such restrictive conditions. To apply the diff-in-diffs estimator, at least one pre-programme set and one post-programme set of observations are required. Let t_0 and t_1 denote the pre- and post-programme periods for which data are available.

The diff-in-diffs estimator measures the excess outcome growth for the treated compared with the non-treated. Formally, abstracting from other regressors besides the treatment indicator,

$$(20)\qquad \hat{\alpha}_{DID} = (\overline{Y}_{t_1}^{T} - \overline{Y}_{t_0}^{T}) - (\overline{Y}_{t_1}^{C} - \overline{Y}_{t_0}^{C}),$$

where $\overline{Y}^T$ and $\overline{Y}^C$ are the mean outcomes for the treatment and comparison (non-treatment) groups, respectively.

(a) The Diff-in-Diffs Estimator: Heterogeneous Treatment Effects

Where the impact of treatment is heterogeneous, provided the above conditions are verified, the diff-in-diffs estimator recovers the impact of the treatment on the treated:

$$(21)\qquad \begin{aligned} E(\hat{\alpha}_{DID}) = {} & [\beta + \alpha_T + E(U_{t_1} \mid d=1) - \beta - E(U_{t_0} \mid d=1)] \\ & -[\beta + E(U_{t_1} \mid d=0) - \beta - E(U_{t_0} \mid d=0)] = \alpha_T. \end{aligned}$$

That is, the effect of treatment on the treated is identifiable, but not the population impact. Intuitively, this happens because the unobserved component of the treatment impact enters in the model as a temporary individual-specific effect that determines participation.

(b) The Diff-in-Diffs Estimator: The Common Trends and Time-Invariant Composition Assumptions

In contrast to the IV and Heckman selection estimators, no exclusion restrictions appear to be required for the diff-in-diffs estimator. In fact, there is no need for any regressor in the decision rule. Even the outcome equation does not have to be specified as long as the treatment impact enters additively. However, strong restrictions on common trends and error composition are implicit, which we now describe.

Consider the following decomposition of the unobservables, U_{it}:

$$(22)\qquad U_{it} = \phi_i + \theta_t + \mu_{it},$$

where ϕ_i is an individual-specific effect, constant over time, θ_t is a common macroeconomic effect, the same for all agents, and μ_{it} is a temporary individual-specific effect. Notice that, if the expectation of U_{it} conditional on the treatment status depends on the temporary individual-specific effect, μ_{it}, diff-in-diffs is inconsistent. This estimator is, however, able to control for the other two error-term components as they cancel out on subtraction. As is straightforward to verify, a separability condition between individual and temporal effects has to be assumed:

(23) $$E(U_{it} \mid d_i) = E(\phi_i \mid d_i) + \theta_t.$$

Even simpler than this estimator, a simple difference method could be applied if the only unobservable term is ϕ_i, the constant individual-specific effect. The estimator

(24) $$\hat{\alpha}_D = (\overline{Y}_{t_1}^T - \overline{Y}_{t_0}^T)$$

would suffice to identify α consistently.

There are two main weaknesses of the diff-in-diffs approach. The first relates to the lack of control for unobserved temporary individual-specific components that influence the participation decision. In fact, the following can be written:

(25) $$E(\hat{\alpha}) = \alpha + E(\mu_{t_1} - \mu_{t_0} \mid d = 1) - E(\mu_{t_1} - \mu_{t_0} \mid d = 0).$$

To illustrate the conditions under which such inconsistency might arise, suppose we are interested in evaluating a training programme in which enrolment is more prone to happen 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 to occur among the treated, even without programme participation. Thus the diff-in-diffs estimator is likely to overestimate the impact of treatment. Also, if only repeated cross-section data are available, it may be difficult to control for the before–after comparability of the groups under this type of selection into the programme. That is, if individuals select into the programme according to some unknown rule, and repeated cross-section data are being used, the assumption that $E(\phi_i \mid d_i)$ is constant over time for each group may be too strong because the composition of the groups may change over time and be affected by the intervention.

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 diff-in-diffs estimator that is presented below.

(c) The Diff-in-Diffs Estimator: Adjusting for Differential Trends

Suppose that the comparison group and target group actually satisfy

$$(26) \qquad E(U_{it} \mid d_i) = E(\phi_i \mid d_i) + k_g \theta_t ,$$

where the k_g acknowledges the differential macro effect across the two groups. Now it can be seen that the diff-in-diffs estimator identifies

$$(27) \qquad E(\hat{\alpha}_{DID}) = \alpha + (k_T - k_C)(\theta_{t_1} - \theta_{t_0}),$$

where T and C refer to the treatment and control groups, respectively. This clearly only recovers the true effect of the programme when $k_T = k_C$.

Now suppose we take another time interval t_* to t_{**}, over which a similar macro trend has occurred. Precisely, we require a period for which the macro trend matches the term $(k_T - k_C)(\theta_{t_1} - \theta_{t_0})$ in (27). It is likely that the most recent cycle is the most appropriate, earlier cycles possibly having 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

$$(28) \qquad \hat{\alpha}_{TADID} = [(\overline{Y}^T_{t_1} - \overline{Y}^T_{t_0}) - (\overline{Y}^C_{t_1} - \overline{Y}^C_{t_0})] - [(\overline{Y}^T_{t_{**}} - \overline{Y}^T_{t_*}) - (\overline{Y}^C_{t_{**}} - \overline{Y}^C_{t_*})],$$

will now consistently estimate α.

4. The Matching Estimator

The matching method is a non-parametric approach to the problem of identifying the treatment impact on outcomes. It is more general in the sense that no particular specification has to be assumed. Moreover, it can be combined with other methods, producing more accurate estimates and allowing for less restrictive assumptions. However, it too rests on strong assumptions and particularly heavy data requirements.

The main purpose of matching is to re-establish the conditions of an experiment when no such data are available. As discussed earlier, with total random assignment within one group, one could compare the treated and the non-treated directly, without having to impose any structure on the problem. With the matching method, the construction of a correct sample counterpart for the missing information on the treated outcomes had they not been treated consists in pairing each programme participant with members of a comparison group (non-treated). Under the matching assumption, the only remaining difference between the two groups is programme participation.

(a) The Matching Estimator: General Method

To illustrate the matching solution in a more formal way, consider a general specification of the outcome function,

$$\begin{aligned} Y^T &= g^T(X) + U^T \\ Y^C &= g^C(X) + U^C, \end{aligned} \tag{29}$$

where Y^T and Y^C are the outcomes of the treated and the non-treated (comparison group), which can be written as a function of the set of observables, *X*, plus the unobservable term, U^T or U^C. Note that we allow for different outcome functions according to the participation decision.

As above, the most common goal of evaluation is to identify the impact of the treatment on the treated:

$$\alpha_T = E(Y^T - Y^C \mid X, d = 1). \tag{30}$$

The solution advanced by matching is based on a fundamental assumption of conditional independence between non-treated outcomes and programme participation:

$$Y^C \perp d \mid X. \tag{31}$$

This assumption states that the outcomes of the non-treated are independent of the participation status, *d*, once one controls for the observable variables, *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.[6] For each treated observation, Y^T, we can look for a non-treated (set of) observation(s), Y^C, with the same *X*-realisation. With the matching assumption, this Y^C constitutes the required counterfactual. Actually, this is a process of rebuilding an experimental dataset which, in general, places strong requirements on data collection.

Additionally, matching also assumes that $0 < \text{Prob}(d = 1 \mid X) < 1$ in order to guarantee that all treated agents have a counterpart in the non-treated population, and that anyone constitutes a possible participant. However, this does not ensure

[6] Rosenbaum and Rubin, 1985; Rubin, 1979.

that the same happens within any sample, and it is, in fact, a strong assumption when programmes are directed to tightly specified groups.

Let *S* be the set of all possible values the vector of explanatory variables, *X*, may assume. It is called the 'support of *X*'. Let S^* be the common support of *X*, or the space of *X* that is simultaneously observed among participants and non-participants for the specific dataset being used. Assuming the above described conditions, a subset of comparable observations is formed from the original sample, and with those a consistent estimator for the treatment impact on the treated, α_T, is the empirical counterpart of

$$\frac{\int_{S*} E(Y^T - Y^C \mid X, d=1) \mathrm{d}F(X \mid d=1)}{\int_{S*} \mathrm{d}F(X \mid d=1)}. \tag{32}$$

The numerator of the above expression represents the expected gain from the programme among the subset of participants who are sampled and for whom one can find a comparable non-participant (that is, over S^*). To obtain a measure of the impact of the treatment on the treated, individual gains must be integrated over the distribution of observables among participants and re-scaled by the measure of the common support, S^*. The fraction therefore represents the expected value of the programme effect in the common support of *X*, S^*. It is simply the mean difference in outcomes over the common support, appropriately weighted by the distribution of participants. If the second assumption is fulfilled and the two populations are large enough, the common support is the entire support of both.

As should now be clear, the matching method avoids specifying a particular form for the outcome equation, decision process or either unobservable term. We simply need to ensure that, given the right observables, *X*, the observations of non-participants are statistically what the observations of the treated would be had they not participated. Under a slightly different perspective, it might be said that we are decomposing the treatment effect in the following way:

$$\begin{aligned} E(Y^T - Y^C \mid X, d=1) = [&E(Y^T \mid X, d=1) - E(Y^C \mid X, d=0)] \\ &- [E(Y^C \mid X, d=1) - E(Y^C \mid X, d=0)], \end{aligned} \tag{33}$$

the latter right-hand-side term being the bias conditional on *X*, which is assumed to be zero. The technique is to replace the unobserved outcomes of the participants had they not been treated with the outcomes of non-participants with the same *X*-characteristics.

(b) The Matching Estimator: The Role of the Participation Decision

Up to now, we have been differentiating individuals based on participation. However, the structural difference should rely on the participation decision. The participation decision, though, is not observable among non-participants. These form a mixture of those who, if offered the programme, would have decided to participate and those who would have decided not to. All the participants, however, were willing to be treated when the programme was offered to them. In such case, the outcome equations would be

$$(34) \qquad \begin{aligned} Y^T &= g^T(X) + U^T \\ Y^C &= g^C(X) + [d^D U_1^C + (1-d^D)U_0^C], \end{aligned}$$

where d^D is a dummy variable standing for participation decision and U_1^C and U_0^C are the outcome error terms for non-participants who would and would not be willing to participate, respectively.

The parameter of interest — the mean impact of treatment — is

$$(35) \qquad E(Y^T - Y^C \mid X, d=1) = g^T(X) - g^C(X) + E(U^T - U_1^C \mid X, d^D = 1).$$

Therefore there are two possibilities underlying matching assumptions: $\text{Prob}(d^D = 1 \mid X) = 1 \quad \forall X \in S^*$ or $E(U_0^C \mid X) = E(U_1^C \mid X) \quad \forall X \in S^*$. The first hypothesis states that *X* completely determines the participation decision: anyone characterised by a value of *X* on the common support, S^*, would be willing to participate if offered the programme. This is the desired outcome if one is willing to reconstruct an experimental setting, since it states that *X* is enough to build up a comparison group with the desired similarities to the treatment group. The second assumption states that, at least as far as the unobservables are concerned, the two comparison groups defined by the participation decision are equal. This means that participation decisions are being based on observables alone and the matching assumption (31) follows.

Under this formulation, matching is always preferable to random sampling if it increases $\text{Prob}(d^D = 1)$ among comparisons and/or if it brings $E(U_1^C)$ and $E(U_0^C)$ closer in the support, S^*. Any of these conditions causes the comparison group to become more similar to the treatment group in the sense that at least a part of the difference is being controlled for by the observables. This is the advantage of applying matching under such circumstances.

(c) The Matching Estimator: The Use of the Propensity Score

It is clear that when a wide range of variables *X* is in use, matching can be very difficult due to the high dimensionality of the problem. 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*: $P(X_i) = \text{Prob}(d_i = 1 \mid X_i)$, which is the propensity score. Its use is usually motivated by Rosenbaum and Rubin's (1983 and 1984) result. It is shown that, under the (matching) assumptions

$$(36) \qquad (Y^T, Y^C) \perp d \mid X \quad \text{and} \quad 0 < \text{Prob}(d = 1 \mid X) < 1,$$

the conditional independence remains valid if controlling for *P*(*X*) instead of *X*:

$$(37) \qquad (Y^T, Y^C) \perp d \mid P(X).$$

More recently, a study by Hahn (1998) shows that the propensity score is ancillary for the estimation of the average effect of treatment on the population. However, it is also shown that knowledge of the propensity score may improve the efficiency of the estimates of the average effect of treatment on the treated. Its value for the estimation of this latter parameter lies in the 'dimension reduction' feature.

When using the propensity score, the comparison group for each treated individual is chosen with a pre-defined criterion (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 with 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, which 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

$$(38) \qquad \hat{\alpha}_{MM} = \sum_{i \in T} \left(Y_i - \sum_{j \in C} W_{ij} Y_j \right) w_i,$$

where W_{ij} is the weight placed on comparison observation *j* for individual *i* and w_i accounts for the reweighting that reconstructs the outcome distribution for the treated sample. For example, in the nearest neighbour matching case, the estimator becomes

$$(39) \qquad \hat{\alpha}_{MM} = \sum_{i \in T} (Y_i - Y_j) \frac{1}{N_T},$$

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

(d) The Matching Estimator: Parametric Approach

Specific functional forms assumed for the g-functions in (29) can be used to estimate the impact of treatment on the treated over the whole support of X, reflecting the 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.

When a linear specification is assumed with common coefficients for treatments and controls, so that

$$(40) \qquad \begin{aligned} Y^T &= X\beta + \alpha_T d + U \\ Y^C &= X\beta + U, \end{aligned}$$

not even the common support requirement is needed to estimate the impact of treatment on the treated — a simple OLS regression using all information on the treated and non-treated will consistently identify α_T.

(e) The Matching Estimator: Drawbacks

It is likely that matching does not succeed in finding a non-treated observation with similar propensity score for all the participants. That is, for some observations, we might be unable to find the right counterfactual, which means that the common support is just a subset of the complete treated support. If the impact of treatment is homogeneous, at least within the treatment group, no additional problems appear besides the loss of information. Note, however, that the setting is general enough to include the heterogeneous case. If the impact of training is heterogeneous within the treatment group itself and the counterfactual is more difficult to obtain for some subgroup(s) of the participants, it may be impossible to identify α_T. In other words, if the matching process leads to a considerable loss of observations, the estimator is limited by the loss of information and is only consistent for the common support. In the 'heterogeneous response' case, if the expected impact of participation differs across the treated, it is possible that the estimated impact does not represent the mean outcome of the programme.

Another potential problem with matching is the (heavy) requirements on data. To guarantee that assumption (31) is verified, it is important to obtain the relevant information to distinguish potential participants from others, which is not always easy. On the other hand, 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 correct balance between the quantity of information to use and the share of the support covered may be difficult to achieve.

(f) A Bias Decomposition

The bias term can be decomposed into three distinct parts:

$$(41) \qquad Bias = E(Y^C \mid X, d=1) - E(Y^C \mid X, d=0) = B_1 + B_2 + B_3,$$

where B_1 represents the bias component due to non-overlapping support of *X*, B_2 is the error part due to misweighting on the common support of *X* as the resulting empirical distributions of treated and non-treated are not the same even when restricted to the same support, and B_3 is the true econometric selection bias resulting from 'selection on unobservables'. Through the process of choosing and reweighting observations, matching corrects for the first two sources of bias, and the third term is assumed to be zero.

(g) Matching and Diff-in-Diffs

The assumption of conditional independence between the error term in the outcome equation and the training status (depicted by (31)) is quite strong if it is possible that individuals decide according to their forecast outcome. However, if matching is combined with diff-in-diffs, there is scope for an unobserved determinant of participation as long as it can be represented by separable individual- and/or time-specific components of the error term. To clarify the exposition, let us now assume the following model specification:

$$(42) \qquad \begin{aligned} Y_{it}^T &= g_t^T + \phi_i + \theta_t^T + \mu_{it}^T \\ Y_{it}^C &= g_t^C + \phi_i + \theta_t^C + \mu_{it}^C, \end{aligned}$$

which differs from (29) by the composition assumed for the error term and by explicitly acknowledging that the function g may change over time.[7]

If performing matching on the set of observables X within this setting, the conditional independence assumption (31) can now be replaced by

$$(43) \qquad Y_{t_1}^C - Y_{t_0}^C \perp d \mid X,$$

where t_0 and t_1 stand for the before- and after-programme time periods. Given (42), assumption (43) is equivalent to

$$(44) \qquad (g_{t_1}^C - g_{t_0}^C) + (\theta_{t_1}^C - \theta_{t_0}^C) \perp d \mid X.$$

The main matching hypothesis is now stated in terms of the before–after evolution instead of levels. If both terms of the sum in (44) are conditionally independent of the participation decision, then (44) is verified. 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 treated. This happens both on the observable component of the model and on the unobservable time trend.

The effect of the treatment on the treated can now be estimated over the common support of X, S^*, using an extension to (38):

$$(45) \qquad \hat{\alpha}_{MMDID}^{LD} = \sum_{i \in T} \left[(Y_{it_1} - Y_{it_0}) - \sum_{j \in C} W_{ij} (Y_{jt_1} - Y_{jt_0}) \right] w_i,$$

where LD denotes ‘longitudinal data’ and MMDID denotes ‘method of matching with difference-in-differences’.

Quite obviously, this estimator requires longitudinal data to be applied. It is, however, 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 being treated: to find the comparable treated before the programme and the controls before and after the programme. If the same assumptions apply, one can estimate the effect of treatment on the treated using the following estimator:

$$(46) \qquad \hat{\alpha}_{MMDID}^{RCS} = \sum_{i \in T_1} \left[\left(Y_{it_1} - \sum_{j \in T_0} W_{ijt_0}^T Y_{it_0} \right) - \left(\sum_{j \in C_1} W_{ijt_1}^C Y_{jt_1} - \sum_{j \in C_1} W_{ijt_0}^C Y_{jt_0} \right) \right] w_i,$$

[7] Of course, this latter point is only important when comparing different periods, as done within the diff-in-diffs methodology.

where RCS denotes 'repeated cross-section', T_0, T_1, C_0 and C_1 stand for the treatment and control groups, before and after the programme, respectively, and W_{ijt}^{G} represent the weights attributed to individual *j* in group *G* (where *G* = *C* or *T*) and at time *t* when comparing with treated individual *i*.[8]

V. SOME EMPIRICAL STUDIES

In this section, we draw on two studies, one from the UK and one from the US, to illustrate some of the non-experimental techniques presented in this review. In the influential study by LaLonde (1986), it was concluded that none of the used econometric methodologies estimate accurately the treatment impact when only non-experimental data are available.[9] However, there are two potential issues with the LaLonde study (see Heckman, Ichimura and Todd (1997)). The first concerns the questionnaires: controls and comparisons answered different questions, based on different definitions. Second, the comparison group was not guaranteed to operate in the same labour market as the treatments, and hence different macro effects may influence each group's behaviour.

The studies presented below illustrate that the methods we have described for non-experimental data can provide good evaluation information if carefully handled. Both illustrations concern labour market programmes, the first taking place in the UK and the second in the US.

1. Diff-in-Diffs and Differential Trends: The New Deal Evaluation in the UK

The New Deal for Young People is a recent initiative of the UK government to help young unemployed people make their way into or back to work. The programme is targeted at the 19- to 24-year-old long-term unemployed. Participation is compulsory, so that every eligible individual is due to participate under the threat of losing entitlement to benefits. The criteria for eligibility are simple: every individual aged 19–24 by the time of completion of the sixth month on jobseeker's allowance (JSA) is immediately assigned to the programme and starts receiving treatment. Given the stated rules, the programme can be classified as one of global implementation, being administered to literally everyone in the UK meeting the eligibility criteria. Indirect effects are therefore expected. The nature of these effects will be discussed below.

Treatment is composed of three steps. On assignment to the programme, the individual starts an intensive job-search assistance period, called the Gateway, which lasts for up to four months. The second stage is composed of a six-month spell in subsidised employment or up to 12 months in full-time education or

[8] For a more detailed discussion with an application of the combined matching and diff-in-diffs estimator, see Blundell, Costa Dias, Meghir and Van Reenen (2000).

[9] Tables 5 and 6 of LaLonde (1986) reveal that better estimates are attained when using two-step estimators.

training. The former involves a payment of a subsidy to the employer while the employee receives the offered wage. For the latter, the individual receives an amount equivalent to the JSA payment and may be eligible for special grants in order to cover exceptional expenses. Once the option period is over, individuals who remain unemployed enrol in a new period of intense job search, the Follow-Through, which takes up to 13 weeks.

The programme was launched in the whole UK by April 1998. There was, however, a previous three-month experimental period (January 1998 to March 1998) when the programme was tried in 12 regions, called Pathfinders. The goal was to perform a three-month experiment with the Pathfinders, having as counterfactual the rest of the UK or some regions that would match the Pathfinders more closely. Clearly, identification of the treatment effect under these conditions requires stronger assumptions than when the experiment is run within regions using random assignment. As will be discussed, the problem relates to the fact that the counterfactual must be drawn either from a different labour market or from a group with different characteristics operating in the same labour market. Different types of hypotheses will be studied below.

The analysis that follows is based on the study by Blundell, Costa Dias, Meghir and Van Reenen (2000). It uses the publicly available 5 per cent sample of the whole population claiming JSA in the UK since 1982 (JUVOS). This database includes a small set of demographic variables and the start and exit dates from the claimant count, making it possible to reconstruct the unemployment history of the individuals. The outflow from the claimant count is the outcome of interest, the choice having been determined by the availability and quality of the data used (outflows by destination are also covered in Blundell, Costa Dias, Meghir and Van Reenen (2000), but since the necessary information is only available since late 1996, we have chosen to focus here on the outflows to all destinations taken together). Also, since the programme is very recent, it is still not possible to make a long-run analysis of the effect of participation. Given this, we will use two measures in trying to evaluate the effect of the programme: outflows from the claimant count within, respectively, two and four months of completion of the sixth month on unemployment subsidy.

The rest of this section goes as follows. We start by briefly discussing the nature of the experiment. The second subsection addresses the problem of choosing and assessing a control group. We then present the estimates of the effect of the programme, and finally we discuss these results and their potential problems, mainly related to the nature of the programme.

(a) The Experimental Period

We will present a detailed analysis of the experimental period of the New Deal for Young People. The experiment was undertaken during the first three months of 1998 in a selected set of 12 regions in the UK. Every individual attending the local employment offices and meeting the eligibility criteria was assigned to the programme and started receiving treatment. Outside the Pathfinder areas, however, the New Deal was only released three months later, by April 1998.

To clarify things, it must be recognised that what has been done is not a true experiment. The main reason relates to the lack of random assignment. The regions were chosen and the programme was globally implemented in the selected places. Also, the information collected within the programme only included participants. We do not make use of data collected by the programme administration in non-Pathfinder regions. This latter issue, however, raises no relevant problem for the analysis being performed since we are using a truly random sample of all individuals claiming JSA, and for all of them the same type of information is available.

(b) Defining and Assessing the Potential Comparison Groups

The analysis will be performed based on three possible comparison groups. They are defined as follows. Comparison Group 1 is composed of individuals living in non-Pathfinder areas, aged 19–24 and completing their sixth month on JSA during the first quarter of 1998. To construct Comparison Group 2, we have used information on the labour market to determine which regions are ‘closest’ to the Pathfinder areas in the following sense. The variable selected to choose the regions was the time taken to leave unemployment by agents aged 19–24. The procedure used monthly data on the median number of days claiming JSA, by region, and for each Pathfinder area selected the two non-Pathfinder regions that best reproduced its time-series pattern before the programme took place. Systematic differences in levels were not the main concern, since they can be controlled for using diff-in-diffs methodologies. Instead, the variability in the difference between the two curves was minimised, attempting to make the difference as constant over time as possible. Thus Comparison Group 2 comprises the subset of non-Pathfinder local labour markets with a time pattern that most closely resembles the ones observed for Pathfinder areas. Comparison Group 3 is taken to be the set of individuals living in Pathfinder areas, aged 25 to 30 and completing their sixth month on JSA during the first quarter of 1998. The treatment group is, of course, composed of individuals living in Pathfinder areas, aged 19–24 and completing their sixth month on JSA during the first quarter of 1998.

In what follows, we will compare the characteristics of the different groups before the programme is released. We begin by analysing the time to leave the claimant count, the variable chosen to select the regions used in Comparison Group 2. This variable is not exactly what will be used as a measure of the

FIGURE 1

**Median Number of Days Claiming JSA:
Comparing 19- to 24-Year-Olds and 25- to 30-Year-Olds Living in Pathfinder and Non-Pathfinder Areas**

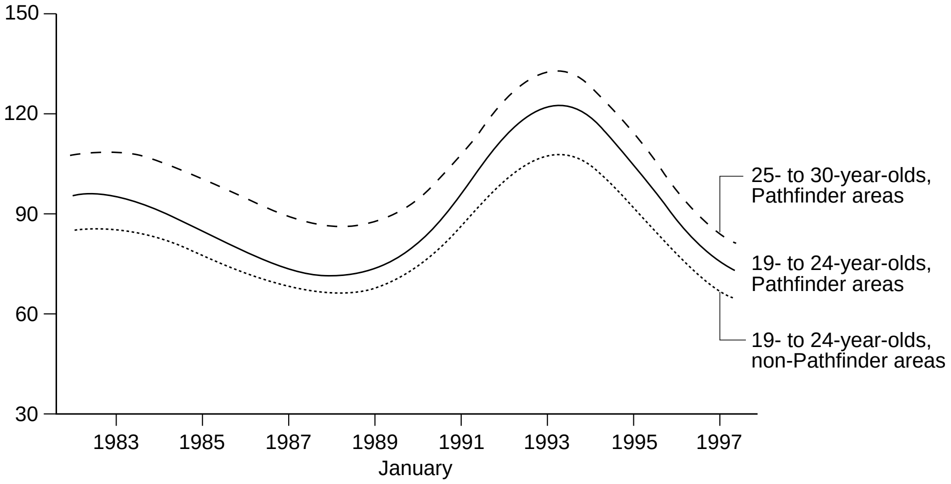
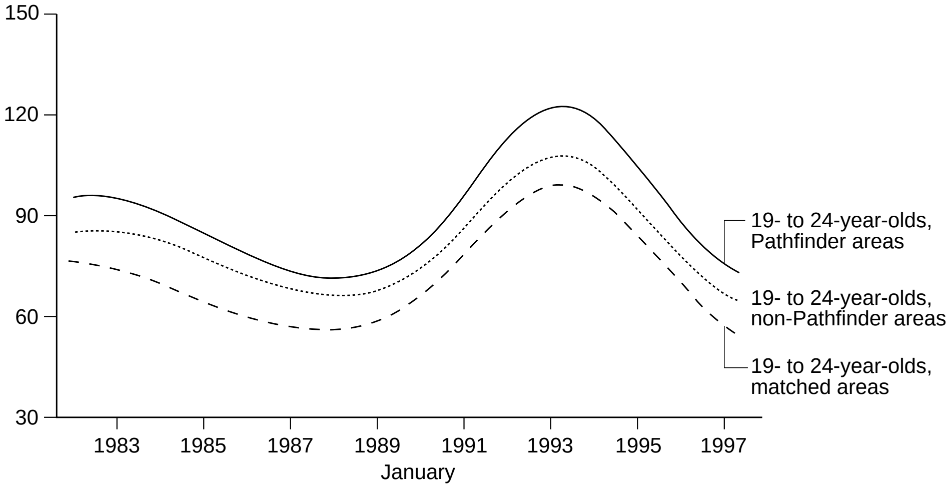


FIGURE 2

**Median Number of Days Claiming JSA:
Comparing 19- to 24-Year-Olds Living in Pathfinder, Non-Pathfinder and Matched Non-Pathfinder Areas**



outcome, since it includes everybody entering unemployment, not just those remaining unemployed for more than six months. However, it may provide a good characterisation of the labour market.

Figures 1 to 3 illustrate the performance of the different comparison groups against the performance of the treatment group in terms of the time to leave unemployment. Figure 1 includes Comparison Groups 1 and 3 along with the treatment group. The younger groups take less time to leave the claimant count, and the Pathfinder areas seem to behave historically worse than the rest of the UK as the unemployed there take longer to leave the claimant count. However, since constant differences do not affect the estimates of the treatment effect, we are more interested in analysing the variability of the differences over time. The three curves exhibit some parallelism but maybe not as much as would be desirable. The difference between the treatment group and Comparison Group 1 curves seems to be more volatile than the difference between the curves corresponding to the treatment group and Comparison Group 3 (the variances of the differences are 3.2 and 2.6, respectively). This seems to indicate that labour markets for different age-groups in the same region are more similar than labour markets in different regions for the same age-group.

Figure 2 presents Comparison Groups 1 and 2 against the treatment group. The matching procedure seems to have created a better comparison group. In fact, the variance of the difference between treatment group and Comparison Group 2 is about 2.6, lower than the 3.2 found when using Comparison Group 1.

FIGURE 3

Median Number of Days Claiming JSA: Comparing 25- to 30-Year-Olds Living in Pathfinder, Non-Pathfinder and Matched Non-Pathfinder Areas

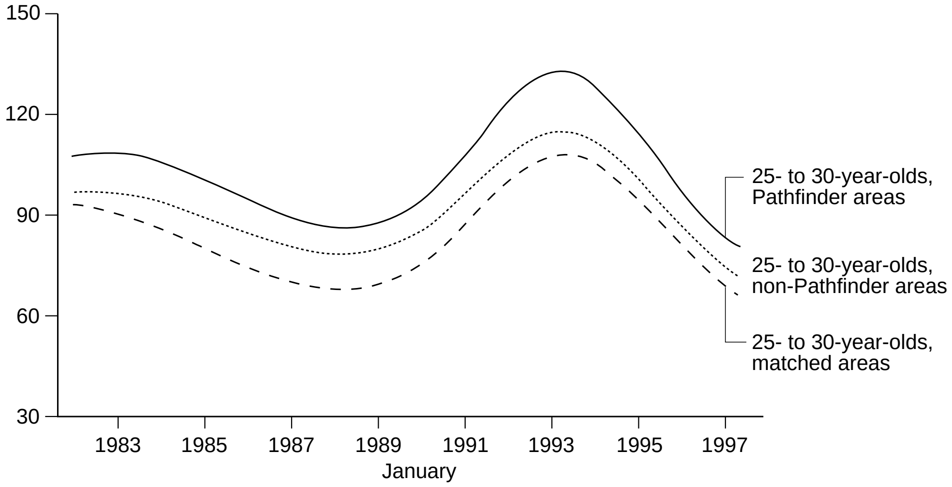


TABLE 4

Comparing the Characteristics of the Treatment and Comparison Groups

Marital status: proportion married

	Entry quarter							
	1990:I	*1991:I*	*1992:I*	*1993:I*	*1994:I*	*1995:I*	*1996:I*	*1997:I*
Treatment group	0.040	0.036	0.052	0.036	0.040	0.038	0.027	0.028
Comp. Group 1	0.043	0.045*	0.046	0.040	0.036	0.031	0.028	0.026
Comp. Group 2	0.039	0.041	0.043	0.040	0.036	0.027*	0.027	0.026
Comp. Group 3	0.353**	0.318**	0.359**	0.293**	0.290**	0.242**	0.239**	0.204**

Unemployed less than six months over the last two years

	Entry quarter							
	1990:I	*1991:I*	*1992:I*	*1993:I*	*1994:I*	*1995:I*	*1996:I*	*1997:I*
Treatment group	0.749	0.781	0.865	0.727	0.688	0.645	0.663	0.692
Comp. Group 1	0.762	0.786	0.883**	0.737	0.683	0.651	0.673	0.674
Comp. Group 2	0.773*	0.812**	0.896**	0.745	0.708	0.661	0.685	0.684
Comp. Group 3	0.520**	0.621**	0.853	0.586**	0.523**	0.450**	0.432**	0.444**

Unemployed less than 12 months over the last two years

	Entry quarter							
	1990:I	*1991:I*	*1992:I*	*1993:I*	*1994:I*	*1995:I*	*1996:I*	*1997:I*
Treatment group	0.885	0.901	1.000	0.864	0.832	0.803	0.831	0.827
Comp. Group 1	0.899	0.911	0.999	0.883**	0.833	0.801	0.810**	0.822
Comp. Group 2	0.908**	0.928**	1.000	0.887**	0.842	0.807	0.813	0.830
Comp. Group 3	0.732**	0.803**	1.000	0.768**	0.704**	0.652**	0.650**	0.619**

No unemployment spells within the last two years

	Entry quarter							
	1990:I	*1991:I*	*1992:I*	*1993:I*	*1994:I*	*1995:I*	*1996:I*	*1997:I*
Treatment group	0.350	0.367	0.537	0.442	0.464	0.438	0.425	0.445
Comp. Group 1	0.361	0.398**	0.528	0.441	0.437**	0.418	0.419	0.418*
Comp. Group 2	0.361	0.409**	0.549	0.445	0.457	0.417	0.431	0.430
Comp. Group 3	0.220**	0.300**	0.483**	0.332**	0.254**	0.235**	0.219**	0.212**

Number of observations

	Entry quarter							
	1990:I	*1991:I*	*1992:I*	*1993:I*	*1994:I*	*1995:I*	*1996:I*	*1997:I*
Treatment group	1,727	1,762	1,815	1,752	1,628	1,623	1,512	1,424
Comp. Group 1	11,102	11,869	11,951	12,029	11,014	11,585	9,721	8,402
Comp. Group 2	2,349	2,631	2,834	2,875	2,709	2,555	2,401	2,054
Comp. Group 3	781	881	1,036	1,140	1,089	1,028	1,013	949

Key: Treatment group: men aged 19–24 living in Pathfinder areas
Comp. Group 1: men aged 19–24 living in non-Pathfinder areas
Comp. Group 2: men aged 19–24 living in matched non-Pathfinder areas
Comp. Group 3: men aged 25–30 living in Pathfinder areas

*Estimates for treatment and respective comparison group are statistically different at 10 per cent.

**Estimates for treatment and respective comparison group are statistically different at 5 per cent.

Figure 3 presents the same kind of comparisons for the 25- to 30-year-olds, using the regions matched for the younger group. Similar comments apply.

Table 4 compares the treatment group with the three selected comparison groups in a range of other characteristics before the programme was launched.

In general, there are no significant differences between the treatment group and Comparison Groups 1 and 2. Comparison Group 3, however, exhibits a different pattern in literally all the presented dimensions. As mentioned above, differences that are constant over time do not affect the consistency of the diff-in-diffs estimates, and these differences certainly show some systematic pattern.

(c) The Effect of the Programme

To assess the effect of the treatment, we have chosen to use two possible outcome variables: the outflow from the claimant count within two months of completing the sixth month on unemployment subsidy (the start of the treatment) and the outflow from the claimant count within four months of the start of the treatment. Table 5 presents some of the estimated effects for both measures when comparing the treatment group with the three comparison groups considered.

The first column of estimates presented in Table 5 uses a single differences method. It is assumed that the instrumental variables used to define the comparison groups (either the living area or the age) are correlated with the treatment indicator but uncorrelated with the outcome. The results obtained are as follows. Using Comparison Group 1, we estimate that the probability of leaving the claimant count within two months of completion of the sixth month on unemployment benefit is 4.4 per cent higher for treated individuals. This estimate is not significant, however, but after four months the estimated effect rises to almost 12 per cent and achieves statistical significance. If these are the true parameters, it means that the New Deal is effectively helping people out of unemployment quite significantly.

We have reproduced these estimates under weaker assumptions. The second estimator in the table is the diff-in-diffs, using the first quarter of 1997 as the ‘before-programme’ period. This procedure assumes that treatments and comparisons are equally affected by the same macro shocks, but they are allowed to have group-specific characteristics that are constant over time. Given that the comparison groups are drawn either from different regions or from a different age-group, this assumption may be rather strong. The results obtained are significantly higher than the ones obtained with the single differences procedure: the estimated effects increase by over 3 percentage points for both outcome variables using Comparison Group 1.

We also considered the possibility of group-specific time effects. This suggests the use of a trend-adjusted diff-in-diffs estimator. This method does

TABLE 5

Treatment Effects for People Joining the Programme during the First Quarter of 1998

Comparison Group 1: 19- to 24-year-olds living in non-Pathfinder areas

	Single differences	*Diff-in-diffs*	*Trend-adjusted diff-in-diffs*	*Linear matching diff-in-diffs*	*Linear matching trend-adjusted diff-in-diffs*
No. of observations	1,627	3,716	8,556	3,716	8,556
Effect after two	0.044	0.082**	0.072	0.076*	0.062
months of treatment	(0.031)	(0.041)	(0.056)	(0.041)	(0.056)
Effect after four	0.119**	0.152**	0.144**	0.147**	0.135**
months of treatment	(0.033)	(0.044)	(0.061)	(0.044)	(0.061)

Comparison Group 2: 19- to 24-year-olds living in matched non-Pathfinder areas

	Single differences	*Diff-in-diffs*	*Trend-adjusted diff-in-diffs*	*Linear matching diff-in-diffs*	*Linear matching trend-adjusted diff-in-diffs*
No. of observations	683	1,590	3,350	1,590	3,350
Effect after two	0.011	0.109**	0.073	0.191**	0.060
months of treatment	(0.036)	(0.049)	(0.066)	(0.052)	(0.066)
Effect after four	0.070*	0.098**	0.180**	0.173**	0.164**
months of treatment	(0.039)	(0.049)	(0.071)	(0.052)	(0.072)

Comparison Group 3: 25- to 30-year-olds living in Pathfinder areas

	Single differences	*Diff-in-diffs*	*Trend-adjusted diff-in-diffs*	*Linear matching diff-in-diffs*	*Linear matching trend-adjusted diff-in-diffs*
No. of observations	469	1,096	2,137	1,096	2,137
Effect after two	0.060	0.031	0.031	0.031	0.022
months of treatment	(0.042)	(0.055)	(0.079)	(0.056)	(0.080)
Effect after four	0.154**	0.144**	0.117	0.137**	0.113
months of treatment	(0.046)	(0.061)	(0.089)	(0.062)	(0.089)

*Significant at 10 per cent level.
**Significant at 5 per cent level.
Notes: Standard errors are given in parentheses below the estimate. Trend-adjusted estimates used the 1989:I–1990:I period.

indeed allow for distinct time trends across groups but requires the group-specific macro shocks to exhibit cyclical behaviour, repeating themselves over the cycles. Under this assumption, the best choice for the comparison period is the comparable part of the previous cycle. We have used the 1989:I–1990:I period. The estimates remain at similar levels for Comparison Group 1, being pushed down by around 1 percentage point, but the effect after two months loses statistical significance.

Finally, a linear matching procedure has been applied. It guarantees that groups with similar observable characteristics are being compared. We have combined it with the diff-in-diffs and with the trend-adjusted diff-in-diffs methods. The necessary assumptions on the group-specific effects are being relaxed to the hypothesis that their temporary part is independent of participation, given that we control for a set of observables. However, the use of linear matching along with diff-in-diffs and trend-adjusted diff-in-diffs changes the results for Comparison Group 1 only marginally.

When using Comparison Group 2 — a subset of Comparison Group 1 using the most similar regions — the estimates increase, in general, by between 2 and 4 percentage points. This does not happen, however, with the single differences, where the estimates actually fall.

Constructing the counterfactual from the older group (Comparison Group 3) weakens the results, especially when considering the effect of two months of treatment: these estimates are generally lower when using this comparison group and none of them is significant at conventional levels. The effect after four months of treatment is occasionally estimated with less precision but at levels very similar to the ones obtained when using Comparison Group 1. Given the size of the sample for these comparisons, some loss of statistical significance is to be expected.

Overall, the estimates give the same indication, independently of the chosen comparison group or estimation technique: there is a positive and significant impact of the programme in taking people out of the benefit account. However, these results are not free from criticism. There are a number of reasons why they may not be robust. It could be that the programme itself pushes participants out of the claimant count by placing them in options that they are expected to accept. There may be self-selection on unobservables that are not controlled for by the matching and differencing methods. A third potential criticism of the results relates to substitution. Suppose that the labour supplied by participants if at work is substitutable for the labour supplied by workers similar to but older than the ones we are comparing participants to. If participants are being made more effective at job-searching and are being offered subsidised jobs, it is likely that they will take some of the jobs that would have been taken by their older counterparts. However, without very strong assumptions, it is generally not possible to distinguish the substitution effects from macro shocks. Finally, the global nature of the programme may also give rise to wage effects, especially if the target group is relatively large. These issues are discussed more fully in Blundell, Costa Dias, Meghir and Van Reenen (2000).

2. The Method of Matching: The JTPA Evaluation in the US

A recent study by Heckman, Ichimura and Todd (1997) evaluates matching methods under different assumptions on the richness of available data. Information gathered under the Job Training Partnership Act (JTPA) was used to compare the performance of matching models with experimental procedures. The JTPA is the main US government training programme for disadvantaged workers. It provides on-the-job training, job-search assistance and classroom training to youths and adults. Eligibility is determined by family income being near or below the poverty level for six months prior to application or by participation in federal, state or local welfare and food stamp programmes. Detailed longitudinal data were collected under an experimental setting for a group of treatments and randomised-out controls, as well as for a potential comparison group of eligible non-participants (see Devine and Heckman (1996), Kemple, Dolittle and Wallace (1993) and Orr et al. (1994)). All the groups were resident in the same narrowly defined geographic regions and were administered the same questionnaire. The richness of information also allowed the construction of close comparison groups from other surveys. As in the LaLonde (1986) study, earnings are the outcome measure. With such data, a formal analysis of estimated bias was possible and thus the relative advantages of matching were clearly stated.

Let us start by focusing on the results concerned with the comparability of supports. Heckman, Ichimura and Todd (1997) draw the densities of $P(X)$ (probability of programme participation) for controls and eligible non-participants. It is clear from this study that the common support defined by the propensity to participate is very restricted. This means that the potential non-experimental comparison group, composed of the eligible non-participants, does not reproduce the characteristics of the treated as represented by the experimental comparison group, composed of the controls. Therefore a significant source of bias when dealing with non-experimental data should come from not controlling for non-overlapping support. It is also clear that if the common support is a relatively small subset of the whole support for the treatment group, then the entire group is unlikely to be represented. Of course, the fact that non-experimental evaluations use only a small part of the treated support in trying to avoid the 'non-overlapping support' type of bias implies that the parameter being estimated is not the same as when an experiment dataset is available.

An empirical decomposition of the evaluation bias as measured by the average monthly earnings is presented in Table 6 (see also Table 2 in Heckman, Ichimura and Todd (1997)). As already mentioned, the total evaluation bias is given by $B = E(Y^C \mid X, d=1) - E(Y^C \mid X, d=0)$. Recall that the total bias may be decomposed into three parts: the bias due to non-overlapping support of X

(B_1), the bias due to misweighting on the common support of X (B_2) and the bias resulting from selection on unobservables (B_3). In this study, the first term is estimated with the controls' reported earnings, while for the second term three options are used: eligible non-participants, a group based on the Survey on Income and Program Participation (SIPP) and a group of no-shows which include controls and persons assigned to treatment that dropped out before

TABLE 6

Bias Decomposition of Simple Difference in Post-Programme Mean Earnings Estimator

Experimental controls and eligible non-participants

	Mean difference $\hat{B}$	*Non-overlap* $\hat{B}_1$	*Density weighting* $\hat{B}_2$	*Selection bias* $\hat{B}_3$	*Average bias* $\hat{\bar{B}}_{\text{common}}$	$\hat{\bar{B}}_{\text{common}}$ *as percentage of treatment impact*
Adult males	–342	218	–584	23	38	87%
Adult females	33	80	–78	31	38	129%
Male youth	20	142	–131	9	14	23%
Female youth	42	74	–67	35	49	7,239%

Experimental controls and SIPP eligibles

	Mean difference $\hat{B}$	*Non-overlap* $\hat{B}_1$	*Density weighting* $\hat{B}_2$	*Selection bias* $\hat{B}_3$	*Average bias* $\hat{\bar{B}}_{\text{common}}$	$\hat{\bar{B}}_{\text{common}}$ *as percentage of treatment impact*
Adult males	–145	151	–417	121	192	440%
Adult females	47	97	–172	122	198	676%
Male youth	–188	65	–263	9	21	36%
Female youth	–88	83	–168	–3	–13	1,969%

Experimental controls and no-shows

	Mean difference $\hat{B}$	*Non-overlap* $\hat{B}_1$	*Density weighting* $\hat{B}_2$	*Selection bias* $\hat{B}_3$	*Average bias* $\hat{\bar{B}}_{\text{common}}$	$\hat{\bar{B}}_{\text{common}}$ *as percentage of treatment impact*
Adult males	29	–13	3	38	42	97%
Adult females	9	1	–9	18	20	68%
Male youth	84	14	–21	91	99	171%
Female youth	18	3	–31	46	51	7,441%

receiving any service. The estimated biases result from a simple difference estimator of treatment impact.

It is clear that types 1 and 2 bias account for the majority of the error in any case. None the less, selection bias as correctly defined is a significant error when compared with the treatment impact and is even greater when evaluating the bias on the common support. Another relevant point concerns the usage of different datasets to construct the comparison group. The SIPP data panel includes information detailed enough to evaluate eligibility, but the precise location of respondents is unknown and the survey questions are not exactly the same. As a result, selection bias for estimates using this information is typically higher in both absolute and relative terms.

The results obtained when using no-shows as a comparison group are quite interesting. Those people are likely to be very similar to the treated. In fact, if non-enrolment were random with respect to outcomes, they would be just like the experimental group. Most probably, this is not the case, but the same matching methods as the ones used with eligible non-participants can be applied here to control for the differences. The third panel of Table 6 shows that the bias is substantially lower when using this group instead of eligible non-participants (except for male youth) but it is more heavily weighted toward the selection bias component, B_3.

The comparison between diff-in-diffs and single difference matching estimators using the group of eligible non-participants is reported in Table 7. The outcome measures are the quarterly earnings for quarters 1 to 6 after treatment. The values presented are estimates of the selection bias on common support, B_{S_c}, from four different matching estimators — respectively, simple and regression-adjusted single differences and difference-in-differences. The regression-adjusted estimator lies between fully non-parametric and parametric approaches and is likely to improve the results when compared with completely non-parametric estimators. It is based on a particular specification for the no-treatment outcomes, linear say: $Y^C = X\beta + U^C$. To estimate the treatment impact, we should firstly estimate β and then remove $X\hat{\beta}$ from each Y^T and Y^C observation. With such values, we perform matching on *X* or *P*(*X*) as desired and estimate the impact by a simple mean difference. When using the diff-in-diffs estimator, the removal operation is required for each pre- and post-treatment observation.

The estimates in Table 7 are based on kernel weights. Specifically, each treatment observation is matched with a weighted average of the outcomes for all individuals in the comparison group. Local linear weights are used because they

TABLE 7
Estimated Bias for Alternative Matching Methods: Experimental Controls and Eligible Non-Participants

Quarter	*Local linear matching*	*Regression-adjusted local linear matching*	*Diff-in-diffs local linear matching*	*Diff-in-diffs regression-adjusted local linear matching*
Adult males				
t=1	33	39	97	104
t=2	37	39	77	77
t=3	29	21	90	74
t=4	80	65	112	98
t=5	64	50	19	–5
t=6	37	17	4	–35
Average, 1–6	47	38	67	52
% of adjusted impact	77%	62%	109%	85%
Adult females				
t=1	45	55	65	74
t=2	48	55	53	60
t=3	26	31	10	14
t=4	36	35	12	7
t=5	48	48	29	23
t=6	23	16	–5	–18
Average, 1–6	38	40	27	27
% of adjusted impact	109%	114%	78%	76%
Male youth				
t=1	3	8	43	80
t=2	40	28	43	61
t=3	33	–8	92	70
t=4	44	4	9	–5
t=5	84	42	18	–11
t=6	28	–31	–23	–64
Average, 1–6	39	7	30	22
% of adjusted impact	108%	19%	84%	61%
Female youth				
t=1	31	–8	–7	–14
t=2	79	27	60	27
t=3	121	49	135	83
t=4	37	–28	45	4
t=5	65	8	45	–7
t=6	34	1	31	6
Average, 1–6	61	8	52	17
% of adjusted impact	248%	33%	209%	67%

enable a faster convergence rate at boundary points and adapt better to different data densities (for more details, see Heckman, Ichimura and Todd (1997) and Fan (1992)).[10]

The first two columns of Table 7 present the results using a simple difference matching estimator and the last two contain diff-in-diffs results. The last row of each panel shows the bias as a proportion of the estimated experimental impact on the common support of treatments and eligible non-participants. As predicted, the combination of non-parametric and parametric techniques performs better than fully non-parametric approaches. The diff-in-diffs estimator does better for some groups, but not all.[11] For all estimators presented, there is considerable variation in the estimated bias over time.

In spite of the considerable improvements relative to simpler estimates, the bias remains overly strong as a percentage of the adjusted impact of treatment. There is still considerable selection on unobservables that contaminates the non-experimental estimates.

VI. CONCLUSIONS

This paper has presented an overview of alternative evaluation methods, focusing on approaches that do not require experimental data. We have assessed a number of approaches, including the use of selection, difference-in-differences and propensity score matching. Drawing on studies from the UK and the US, we have reviewed the performance of alternative methods.

The appropriate choice of evaluation method has been shown to depend on a combination of the data available and the policy parameter of interest. Where non-experimental data are all that is available, a careful combination of matching and differencing can provide useful insights into the impact of some policy interventions. For example, in the study of training programmes, it has been found that, where data on local labour market characteristics and previous work experience are collected, an approach that combines propensity score matching with the difference-in-differences technique is quite robust. It allows matching

[10] The expression for the local linear weights is the following:

$$W_{N_C,N_T}(i,j)=\frac{G_{ij}\sum_{k\in I_C}G_{ik}(X_k-X_i)^2-G_{ij}(X_j-X_i)\sum_{k\in I_C}G_{ik}(X_k-X_i)}{\sum_{l\in I_C}G_{il}\sum_{k\in I_C}G_{ik}(X_k-X_i)^2-\left\{\sum_{k\in I_C}G_{ik}(X_k-X_i)\right\}^2},$$

where $W_{N_C,N_T}(i,j)$ is the weight for the comparison *j* when matching with the treated *i*, and the numbers of comparisons and treatments are N_C and N_T, respectively. G_{ij} is a kernel function, $G_{ik}=G\{(X_i-X_k)/a_{N_C}\}$, and a_{N_C} is the band width. Finally, I_C is the sample of comparisons.

[11] It is noteworthy that the identifying hypothesis underlying the diff-in-diffs estimator for symmetric differences around the enrolment date (independence between the post- and pre-treatment mean difference and treatment status) was the only one not being rejected in tests performed by Heckman, Ichimura and Todd (1997).

on pre-programme ‘shocks’ and, by collecting good local pre-programme labour market history data, allows the comparison group to be ‘placed’ in the same labour market.

The methods presented have been discussed in a comparable framework, and the respective assumptions required to estimate the parameter of interest have been laid out systematically. We hope that, by doing so, this review can provide a useful resource in deciding on an appropriate evaluation method and understanding its properties.

APPENDIX: THE HECKMAN SELECTION ESTIMATOR

The two-step selection estimator deals with the selection bias problem through direct control of the part of the error term that is correlated with the treatment status indicator. The procedure is as follows (see Heckman (1979)). Given the independence of Z and V, the probability of programme participation can be estimated using discrete choice analysis. With such information for each agent, and along with knowledge of the joint distribution of the error terms, one can compute the conditional expectation of U_{it},

$$\text{(A.1)}\quad \begin{aligned} E(U_{it} \mid d_i = 0, Z_i) &= \frac{\int_{-\infty}^{+\infty} t_1 \int_{-\infty}^{F^{-1}(\text{Prob}(d_i=0|Z_i))} h_i(t_1,t_2)\mathrm{d}t_2\mathrm{d}t_1}{\text{Prob}(d_i = 0 \mid Z_i)} \\ E(U_{it} \mid d_i = 1, Z_i) &= \frac{\int_{-\infty}^{+\infty} t_1 \int_{F^{-1}(\text{Prob}(d_i=0|Z_i))}^{+\infty} h_i(t_1,t_2)\mathrm{d}t_2\mathrm{d}t_1}{\text{Prob}(d_i = 1 \mid Z_i)}, \end{aligned}$$

where F is the cumulative distribution function of V. This information should be incorporated in the outcome regression equation, jointly with all the other covariates, as a selection bias control. The remaining unobservable will be totally independent of treatment status under the accepted hypothesis, and therefore the estimator is consistent.

BIBLIOGRAPHY

Ashenfelter, O. (1978), ‘Estimating the effect of training programs on earnings’, *Review of Economics and Statistics*, vol. 60, pp. 47–57.

— and Card, D. (1985), ‘Using the longitudinal structure of earnings to estimate the effect of training programs’, *Review of Economics and Statistics*, vol. 67, pp. 648–60.

Bassi, L. (1983), ‘The effect of CETA on the post-program earnings of participants’, *Journal of Human Resources*, vol. 18, pp. 539–56.

— (1984), ‘Estimating the effects of training programs with nonrandom selection’, *Review of Economics and Statistics*, vol. 66, pp. 36–43.

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*, vol. 6, pp. 339–60.

Blundell, R., Costa Dias, M., Meghir, C. and Van Reenen, J. (2000), ‘Evaluating the employment impact of mandatory job-search assistance: the UK New Deal Gateway’, unpublished manuscript, Institute for Fiscal Studies.

—, Dearden, L. and Meghir, C. (1996), *The Determinants and Effects of Work-Related Training in Britain*, London: Institute for Fiscal Studies.

—, Duncan, A. and Meghir, C. (1998), ‘Estimating labour supply responses using tax policy reforms’, *Econometrica*, vol. 66, pp. 827–61.

— and MaCurdy, T. (1999), ‘Labor supply: a review of alternative approaches’, in O. Ashenfelter and D. Card (eds), *Handbook of Labor Economics*, Elsevier North-Holland.

Burtless, G. (1985), ‘Are targeted wage subsidies harmful? Evidence from a wage voucher experiment’, *Industrial and Labor Relations Review*, vol. 39, pp. 105–14.

Card, D. and Robins, P. K. (1998), ‘Do financial incentives encourage welfare recipients to work?’, *Research in Labor Economics*, vol. 17, pp. 1–56.

Cochrane, W. and Rubin, D. (1973), ‘Controlling bias in observational studies’, *Sankyha*, vol. 35, pp. 417–46.

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*, vol. 15, pp. 111–70, Greenwich, CT: JAI Press.

Eissa, N. and Liebman, J. (1996), ‘Labor supply response to the Earned Income Tax Credit’, *Quarterly Journal of Economics*, vol. 111, pp. 605–37.

Fan, J. (1992), ‘Design adaptive nonparametric regression’, *Journal of the American Statistical Association*, vol. 87, pp. 998–1004.

Fisher, R. (1951), *The Design of Experiments*, sixth edition, London: Oliver and Boyd.

Hahn, J. (1998), ‘On the role of the propensity score in efficient semiparametric estimation of average treatment effects’, *Econometrica*, vol. 66, pp. 315–31.

Hausman, J. A. and Wise, D. A. (1985), *Social Experimentation*, Chicago: University of Chicago Press for National Bureau of Economic Research.

Heckman, J. (1979), ‘Sample selection bias as a specification error’, *Econometrica*, vol. 47, pp. 153–61.

— (1990), ‘Varieties of selection bias’, *American Economic Review*, vol. 80, pp. 313–18.

— (1992), ‘Randomization and social program’, in C. Manski and I. Garfinkle (eds), *Evaluating Welfare and Training Programs*, Cambridge, MA: Harvard University Press.

— (1996), ‘Randomization as an instrumental variable estimator’, *Review of Economics and Statistics*, vol. 56, pp. 336–41.

— (1997), ‘Instrumental variables: a study of the implicit assumptions underlying one widely used estimator for program evaluations’, *Journal of Human Resources*, forthcoming.

— and Hotz, V. J. (1989), ‘Choosing among alternative nonexperimental methods for estimating the impact of social programs’, *Journal of the American Statistical Association*, vol. 84, pp. 862–74.

—, Ichimura, H. and Todd, P. (1997), ‘Matching as an econometric evaluation estimator’, *Review of Economic Studies*, vol. 64, pp. 605–54.

— and Robb, R. (1985), ‘Alternative methods for evaluating the impact of interventions’, in *Longitudinal Analysis of Labour Market Data*, New York: Wiley.

— and — (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.

— and Smith, J. (1994), ‘Ashenfelter’s dip and the determinants of program participation’, University of Chicago, mimeo.

—, — and Clements, N. (1997), ‘Making the most out of program evaluations and social experiments: accounting for heterogeneity in program impacts’, *Review of Economic Studies*, vol. 64, pp. 487–536.

Kemple, J., Dolittle, F. and Wallace, J. (1993), *The National JTPA Study: Site Characteristics in Participation Patterns*, New York: Manpower Demonstration Research Corporation.

LaLonde, R. (1986), ‘Evaluating the econometric evaluations of training programs with experimental data’, *American Economic Review*, vol. 76, pp. 604–20.

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*, report to the US Department of Labor, 132, Bethesda, MD: Abt Associates.

Rosenbaum, P. and Rubin, D. B. (1983), ‘The central role of the propensity score in observational studies for causal effects’, *Biometrika*, vol. 70, pp. 41–55.

— and — (1984), ‘Reducing bias in observational studies using subclassification on the propensity score’, *Journal of the American Statistical Association*, vol. 79, pp. 516–24.

— and — (1985), ‘Constructing a control group using multivariate matched sampling methods that incorporate the propensity score’, *American Statistician*, pp. 39–58.

Rubin, D. B. (1978), ‘Bayesian inference for causal effects: the role of randomization’, *Annals of Statistics*, vol. 7, pp. 34–58.

— (1979), ‘Using multivariate matched sampling and regression adjustment to control bias in observational studies’, *Journal of the American Statistical Association*, vol. 74, pp. 318–29.