

COST A23
Training Course May 27-28th, CREST, Paris
Empirical policy evaluation methods

Richard Blundell and Monica Costa-Dias
UCL and IFS

May 2008

Overview

Empirical policy evaluation methods

- **(i) social experiments methods,**
 - and some nonparametric ideas from binary choice and selection models
- **(ii) natural experiment methods,**
 - difference-in-differences
- **(iii) matching methods,**
 - mixing matching and diff-in-diff
- **(iv) instrumental methods,**
 - local average (LATE) and marginal (MTE) effects
- **(v) discontinuity design methods,**
 - fuzzy and sharp designs
- **(vi) control function methods**
 - back to the selection model....

The 'Treatment Effect' Model

Consider the following model of outcomes from treatment. Suppose we wish to measure the impact of treatment on an outcome, y . Denote by d the treatment indicator or assignment rule. It is a dummy variable assuming the value 1 if the agent has been treated and 0 otherwise.

The **potential outcomes** for individual i at any time t are denoted by y_{it}^1 and y_{it}^0 . They are specified as

$$\begin{aligned} y_{it}^1 &= \beta + \alpha_i + u_{it} & \text{if } d_{it} = 1 \\ y_{it}^0 &= \beta + u_{it} & \text{if } d_{it} = 0 \end{aligned} \quad (1)$$

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

The **observable outcome** is then

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

so that

$$y_{it} = \beta + \alpha_i d_{it} + u_{it}. \quad (3)$$

Assignment into treatment determines the treatment status, d . We assume this assignment occurs at a fixed moment in time, say k , and depends on the information available at that time summarised by, Z_k , and unobservables, v_k .

Assignment to treatment is then assumed to be made on the basis of a selection rule

$$d_{it} = \begin{cases} 1 & \text{if } d_{ik}^* \geq 0 \text{ and } t > k, \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where d^* is a function of Z and v

$$d_{ik}^* = g(Z_{ik}, v_{ik}) \quad (5)$$

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

$$d_{it} = \mathbf{1}(Z_{ik}\gamma + v_{ik} \geq 0) \quad (6)$$

where γ is the vector of coefficients and $\mathbf{1}(\cdot)$ is the indicator function.

Which Treatment Parameter? 'Average Treatment on the Treated' and 'Average Treatment' Effects

In the *homogeneous linear model*, common in elementary econometrics, there is only one impact of a programme and it is one that would be common to participants and nonparticipants alike. In the *heterogeneous response model*, the treated and non-treated may benefit differently from programme participation. In this case, the treatment on the treated parameter will differ from the treatment on the untreated parameter or the average treatment effect. Indeed, we can define a whole distribution of the treatment effects. A central issue in understanding evaluation methods relates to the aspects of this distribution that can be recovered by the different approaches.

Three Average Treatment Effects of Interest

Estimation methods typically identify some **average impact of treatment over some sub-population**.

The three most commonly used parameters are:

- The population **average treatment effect** - which would be the outcome if individuals were assigned at random to treatment (ATE)
- The **average treatment effect on the treated** - individuals that were assigned to treatment (ATT)
- The **average treatment effect on non-participants** (ATNT).

When α_i is a constant α across all individuals then all these treatment effects are identical - the homogeneous treatment effect case.

Using the **heterogeneous effects** specification:

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

$$d_{it} = \mathbf{1}(g(Z_{ik}, v_{ik}) \geq 0)$$

or

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

we can express the three average parameters at time $t > k$ as follows

$$\alpha^{ATE} = E(\alpha_i) \tag{7}$$

$$\alpha^{ATT} = E(\alpha_i | d_{it} = 1) = E(\alpha_i | g(Z_{ik}, v_{ik}) \geq 0) \tag{8}$$

$$\alpha^{ATNT} = E(\alpha_i | d_{it} = 0) = E(\alpha_i | g(Z_{ik}, v_{ik}) < 0). \tag{9}$$

An increasing interest on the [distribution of treatment effects](#) has led to the study of additional treatment effects in the recent literature (Bjorklund and Moffitt, 1987, Imbens and Angrist, 1994, Heckman and Vytlacil, 1999).

Two particularly important parameters are the local average treatment effect (LATE) and the marginal treatment effect (MTE). To introduce them we need to assume that d^* is a non-trivial function of Z , meaning that it changes with Z . Now suppose there exist two distinct values of Z , say Z^* and Z^{**} , for which only a subgroup of participants under Z^{**} will also participate if having experienced Z^* .

The average impact of treatment on individuals that move from non-participants to participants when Z changes from Z^* to Z^{**} is the LATE parameter

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

where $d_i(Z)$ is a dichotomous random variable representing the treatment status.

The [MTE](#) measures the change in aggregate outcome due to an infinitesimal change in the participation rate,

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

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

All these parameters will be identical under homogeneous treatment effects. Under heterogeneous treatment effects, however, a non-random process of selection into treatment may lead to differences between them.

Whether the impact of treatment is homogeneous or heterogeneous, *selection bias* may be present.

Assumptions about the **Assignment Rule** are the key to the choice of estimator in evaluation studies.

The **social experiment** is closest to the ‘theory’ free method of a clinical trial, relying on the availability of a randomized assignment rule. The **control function** approach is closest to the structural econometric approach, directly modelling the assignment rule in order to control for selection in observational data. **Natural experiments** mimic the randomized assignment of the experimental setting but do so with non-experimental data and some ‘natural’ randomisation. **Matching** attempts to reproduce the treatment group among the non-treated, re-establishing the experimental conditions in a non-experimental setting, but relies on observable variables to account for selection bias. **Instrumental variables** is a closer to the structural method, relying on exclusion restrictions to achieve identification. **Discontinuity design** methods are closest in spirit to the natural experiment as they exploit discreteness in the rules used to assign individuals to receive a treatment.

(i) The Social Experiment Approach - Random Assignment to Treatment

Suppose it was possible to run a social experiment. In this case, assignment to treatment would be random, and thus independent from the outcome or the treatment effect. This ensures that the treated and the non-treated groups are equal in all aspects apart from the treatment status.

In the heterogeneous impact specification individuals are allowed to respond differently to treatment, $y_i = \beta + \alpha_i d_i + u_i$.

The following are the randomization assumptions:

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

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

Random assignment ensures d_i is independent of u_i **and** α_i . These conditions are enough to identify the average returns in the experimental population using OLS, the ATE.

An Example: The LaLonde Study

Table 1: Comparison of treatments and controls characteristics:
NSW males

Variable	Treatments	Controls
Age	24.49	23.99
Years of school	10.17	10.17
Proportion high school dropouts	0.79	0.80
Proportion married	0.14	0.13
Proportion black	0.76	0.75
Proportion hispanic	0.12	0.14
Real earnings 1 year before treatment	1472	1558
Real earnings 2 year before treatment	2860	3030
Hours worked 1 year before treatment	278	274
Hours worked 2 year before treatment	458	469
Number of observations	2083	2193

AER, 1986: used a randomised trial to examine the impact of training on wages.

Table 2: Annual earnings of male treatments and controls

Year	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

- Why is it that randomised trials are less common in social science than in medical science?
- What are the likely drawbacks of experiments in evaluating economic policies?

The selection problem and the assignment rule

Collecting all the unobserved heterogeneity terms together we can rewrite the outcome equation

$$\begin{aligned} y_i &= \beta + \alpha^{ATE} d_i + (u_i + d_i (\alpha_i - \alpha^{ATE})) \\ &= \beta + \alpha^{ATE} d_i + e_i. \end{aligned} \tag{10}$$

where α^{ATE} is the ATE parameter. Non-random selection occurs if the unobservable term e in (10) is correlated with d .

This implies that e is either correlated with the regressors determining assignment, Z , or correlated with the unobservable component in the selection or assignment equation, v . Consequently there are two forms of non-random selection: *selection on the observables* and *selection on the unobservables*. Different estimators use different assumptions about assignment to identify the impact of treatment.

As a result of selection, the relationship between y and d is not directly observable from the data since participants and non-participants are not comparable.

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

$$y_i = \beta' x_i + \alpha d_i + u_i$$

where α is the impact of treatment on any individual since this is constant across the population in this case.

The OLS estimator will then identify

$$E \left[\hat{\alpha}^{OLS} \right] = \alpha + E[u_i | d_i = 1] - E[u_i | d_i = 0]$$

which is in general different from α if d and u are related.

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

$$E \left[\hat{\alpha}^{OLS} \right] = \bar{\alpha} + E [\alpha_i - \bar{\alpha} | d_i = 1] + E [u_i | d_i = 1] - E [u_i | d_i = 0]$$

Note that the first term $\bar{\alpha} + E [\alpha_i - \bar{\alpha} | d_i = 1]$ is the ATT.

Thus, if d and u are not related, as long as $E [d_i (\alpha_i - \bar{\alpha})] \neq 0$, OLS will recover the ATT not the ATE. $E [d_i (\alpha_i - \bar{\alpha})] \neq 0$ will, in general, imply that the idiosyncratic gains to treatment α_i are used in the participation decision itself.

In the absence of randomised assignment we need to consider the modelling of assignment d , and the relationship between d and the unobservable terms α_i and u_i .

Binary Assignment Models

Let $d_i = 1$, if an assignment to treatment is made

and $d_i = 0$, otherwise

for an individual $i = 1, 2, \dots, N$. We wish to model the probability that $d_i = 1$ given a $k \times 1$ vector of explanatory characteristics $z'_i = (z_{1i}, z_{2i}, \dots, z_{ki})$.

Write this conditional probability as:

$$\Pr[d_i = 1 | z_i] = F(z'_i \gamma)$$

This is a **single linear index** specification.

Parametric if F is known: Probit - *normal*; Logit - *logistic*.

Semi-parametric if F is unknown. We need to know F and γ and this will give us a complete guide to the assignment behaviour.

Semi- and Non-parametric Estimation

(i) Semiparametric

Note that $E(d|z) = \Pr[d = 1|z]$, so that we can write:

$$E(d_i|z_i) = F(z_i'\gamma)$$

keep 'parameters' in the linear index but relax the parametric form for F .

(ii) Nonparametric

$$E(d_i|z_i) = F(g(z_i))$$

both F and g are nonparametric.

Typically (i) has been followed in research.

Notice that the function $F^*(a + bz_i'\gamma)$ cannot be separately identified from $F(z_i'\gamma)$.

Therefore γ is only identified up to location and scale.

From the binary probability model

$$E(d_i|x_i) = F(z_i'\gamma)$$

so that

$$d_i = F(z_i'\gamma) + \varepsilon_i \text{ with } E(\varepsilon_i|z_i) = 0.$$

So, if we knew γ , then we could find F from a conditional mean regression. A regression where the function F was allowed to be fully flexible - 'nonparametric'.

The standard nonparametric approach is to use kernel regression. Write model

$$d_i = F(z_i'\gamma) + \varepsilon_i \text{ as}$$

$$d_i = m(X_i) + \varepsilon_i$$

the kernel regression estimate of m is given by

$$\hat{m}(x, h) = \frac{\sum_{i=1}^n d_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

where $K(\cdot)$ is a kernel function (typically a function proportional to a symmetric unimodal density function centered at zero) and h is a bandwidth or a window width parameter. Observations whose X_i are close to x receives more weight.

Note the similarity of the kernel regression estimator and the cell mean estimator.

Writing $w_{ni}(x) = K\left(\frac{X_i - x}{h}\right) / \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)$, we see that

$$\hat{m}(x, h) = \sum_{i=1}^n w_{ni}^K(x) d_i \text{ where } \sum_{i=1}^n w_{ni}^K(x) = 1.$$

Semiparametric Estimation of the Assignment Rule

Note again that

$$E(d_i | z_i) = F(z_i'\gamma)$$

so that

$$d_i = F(z_i'\gamma) + \varepsilon_i \text{ with } E(\varepsilon_i | z_i) = 0.$$

A semiparametric least squares estimator can be derived. Choose γ to minimise

$$S(\gamma) = \frac{1}{N} \sum \pi(x_i) (d_i - F(z_i'\gamma))^2$$

replacing F_h with a kernel regression at each step with bandwidth h , . simply a function of the scalar $z_i'\gamma$ for some given value of γ . $\pi(x_i)$ is a trimming function that downweights observations near the boundary of the support of $z_i'\gamma$. Ichimura (1993) shows that the estimator of γ is \sqrt{N} consistent and asymptotically normal.

Recall that the problem with non-random assignment is that the parameter identified by the OLS estimator will now be biased

$$E \left[\hat{\alpha}^{OLS} \right] = \bar{\alpha} + E [\alpha_i - \bar{\alpha} | d_i = 1] + E [u_i | d_i = 1] - E [u_i | d_i = 0]$$

We have a general way of modelling assignment d but we need to go further to consider the bias terms $E [\alpha_i - \bar{\alpha} | d_i = 1]$, $E [u_i | d_i = 1]$ and $E [u_i | d_i = 0]$.

The analysis of these terms occurs frequently in the analysis of selection bias in selection models.

This '**Selection Model**' is a mixture of discrete and continuous processes. Our general model is of the form

$$y_i = x_i' \beta + \alpha_i d_i + u_i.$$

Selection models typically consider the modelling of y_{it} when $d_i = 1$. We could write that as

$$\begin{aligned} y_i &= x_i' \beta + \alpha_i + u_i \text{ for } d_i = 1 \\ &= x_i' \beta + u_{1i} \text{ for } d_i = 1 \end{aligned}$$

Selection is according to the assignment rule

$$d_i = \begin{cases} 1 & \text{if } d_i^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $d_i^* = z_i' \gamma + v_i \geq 0$ as in the binary response model. We will assume x is included in z .

To analyse non-random assignment bias, or selection bias, consider the bias in OLS from estimation using the selected sample

$$\begin{aligned} E(u_{1i}|d_i^* > 0) &= E(u_{1i}|z_i'\gamma + v_i) \\ &= E(u_{1i}|v_i > -z_i'\gamma) \\ &\neq 0, \text{ if } u_1 \text{ and } v \text{ are correlated.} \end{aligned}$$

Suppose to begin with we assume (u_1, v) are jointly normally distributed with mean zero and constant covariance matrix,

$$\begin{pmatrix} u_1 \\ v \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{1v} \\ \sigma_{v1} & 1 \end{pmatrix} \right).$$

We can write the orthogonal decomposition of u_1 given v as

$$u = \sigma_{1v}v + \varepsilon$$

where ε is distributed independently of v and has a marginal normal distribution.

Substituting we have

$$\begin{aligned} E(u_{1i}|y_{2i}^* > 0) &= E(\sigma_{1v}v_i + \varepsilon_i|v_i > -z_i'\gamma) \\ &= \sigma_{1v}E(v_i|v_i > -z_i'\gamma) + E(\varepsilon_i|v_i > -z_i'\gamma) \\ &= \sigma_{1v}E(v_i|v_i > -z_i'\gamma) \end{aligned}$$

The conditional mean for the truncated normal

$$\begin{aligned} E(w|w > c) &= \int_c^\infty wf(w|w > c)dw \\ &= \frac{\sigma}{1 - \Phi\left(\frac{c}{\sigma}\right)} \left[-\phi\left(\frac{w}{\sigma}\right) \right]_c^\infty = \sigma \frac{\phi\left(\frac{c}{\sigma}\right)}{1 - \Phi\left(\frac{c}{\sigma}\right)} \end{aligned}$$

Noting that $\sigma_{22} \equiv 1$, we have

$$\begin{aligned}
 E(u_{1i}|d_i^* > 0) &= \sigma_{1v}E(v_i|v_i > -z_i'\gamma) \\
 &= \sigma_{1v}\frac{\phi(-z_i'\gamma)}{1 - \Phi(-z_i'\gamma)} \\
 &= \sigma_{1v}\frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)} \\
 &= \sigma_{12}\lambda(z_i'\gamma).
 \end{aligned}$$

In general provided we have this linear index specification

$$E(u_{1i}|d_i^* > 0) = g(z_i'\gamma)$$

So that selection is simply a function of the single index in the selection equation $z_i'\gamma$, even when joint normality can not be assumed.

The model can now be rewritten:

$$y_i^* = x_i'\beta + \sigma_{1v}\lambda(z_i'\gamma) + \varepsilon_i$$

with $E(\varepsilon|x, z) = 0$.

Require $\lambda(z_i'\gamma)$ to vary independently of $x_i'\beta$, which related to nonparametric identification. In the joint normal case the λ is a nonlinear function given by $\frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)}$ and is **not** perfectly collinear with $x_i'\beta$ even if exactly the same variables are in x and z . However, even in the joint normal case $\frac{\phi(z_i'\gamma)}{\Phi(z_i'\gamma)}$ can be approximately linear over large ranges of $z_i'\gamma$.

In general, identification of parameters when there is non-random assignment requires an exclusion restriction.

Many alternative approaches have been developed to remove the biases that result from the non-random selection terms $E[\alpha_i - \bar{\alpha}|d_{it} = 1]$, $E[u_{it}|d_{it} = 1]$ and $E[u_{it}|d_{it} = 0]$.

(ii) Natural Experiments: the difference-in-differences (DID) estimator

The difference-in-differences approach uses a natural experiment to mimic the randomisation in a social experiment, some naturally occurring event which creates a policy shift for one group and not another. This may refer to a change of law in one jurisdiction but not another, it may refer to some natural disaster which changes a policy of interest in one area but not another, or it may refer to the eligibility of a certain group to a change of policy for which a similar group is ineligible. The difference between the two groups before and after the policy change is contrasted - thereby creating a difference-in-differences estimator of the policy impact.

The DID estimator is applied to situations where there is longitudinal data, or repeated cross section data (where samples are drawn from the same population before and after the intervention being studied.)

We start by considering the evaluation problem when longitudinal data is available. We assume a change in policy occurs at time $t = k$ and each individual is observed before and after the policy change, at times $t = t_0 < k$ and $t = t_1 > k$, respectively. For simplicity of notation, we denote by d_i (without the time subscript) the treatment group to which individual i belongs to. This is identified by the treatment status at $t = t_1$:

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

The DID estimator uses a **common trend assumption** to rewrite the outcome equation as

$$y_{it} = \beta + \alpha_i d_{it} + u_{it} \tag{11}$$

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

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

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

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

The DID estimator essentially uses a decomposition of the error term to rewrite the outcome equation as follows

$$y_{it} = \beta + \alpha_i d_{it} + (\phi_i + m_t + \epsilon_{it}) \quad (12)$$

where u is decomposed into three terms: an unobservable fixed effect, ϕ_i , a common macro shock, m_t , and an idiosyncratic transitory shock, ϵ_{it} .

DID1: The common trend decomposition ($\phi_i + m_t + \epsilon_{it}$)

DID2: Selection into treatment is independent of the temporary individual-specific effect, ϵ_{it} , so that

$$E(\epsilon_{it} | d) = 0$$

so that

$$E(u_{it} | d) = E(\phi_i | d) + m_t.$$

Assume we observe individuals in two periods, before and after the policy change, designated by $t = t_0$ and $t = t_1$, respectively. Denote by d the treatment group, where d can be either 1 (composed of treated individuals at $t = t_1$) or 0 (control group or non-treated individuals at $t = t_1$).

Under assumption (DID) we can write

$$E[\bar{y}_t^d] = \begin{cases} \beta + E[\alpha_i|d=1] + E[\phi_i|d=1] + m_{t_1} & \text{if } d=1 \text{ and } t=t_1 \\ \beta + E[\phi_i|d] + m_t & \text{otherwise} \end{cases}$$

where \bar{y}_t^d is the average outcome over group d at time t .

By sequential differences

$$\begin{aligned} \alpha^{ATT} &= E(\alpha_i|d=1) & (13) \\ &= [E(y_{it}|d=1, t=t_1) - E(y_{it}|d=1, t=t_0)] \\ &\quad - [E(y_{it}|d=0, t=t_1) - E(y_{it}|d=0, t=t_0)] \end{aligned}$$

This is precisely the DID identification strategy. The sample analog of equation (13) is the DID estimator:

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

$$E[\hat{\alpha}^{DID}] = E[\alpha_i|d=1] = \alpha^{ATT}$$

This means that an alternative way of obtaining $\hat{\alpha}^{DID}$ is to take the first differences of (11) to obtain

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

where ϵ represents the transitory idiosyncratic shocks.

Under the DID assumptions, the above regression equation can be consistently estimated using OLS.

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

It follows that repeated cross-sectional data would be enough to identify ATT for as long as treatment and control groups can be separated before the policy change, in period $t = t_0$.

With cross-section data need a further assumption to rule out systematic composition changes in each group over time.

Such information is sufficient for the average fixed effect per group to cancel out in the before after differences.

Weaknesses of DID

(a) Selection on idiosyncratic temporary shocks: Ashenfelter's dip

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

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

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

(b) Differential macro trends

The identification of ATT using DID relies on the assumption the treatment and controls experience the same macro shocks. If this is not the case, the DID approach will yield a biased and inconsistent estimate of ATT. For example, differential trends might arise in the evaluation of training programs if treated and controls operate in different labour markets.

(c) Compositional changes over time

Although DID does not require longitudinal data to identify the true ATT parameter, it does require that the same group treatment and control to be followed over time. In particular, the composition of the groups with respect to the fixed effects term must remain unchanged to ensure before-after comparability.

Examples and discussion:

Two examples:

1. The EITC tax reform as a natural experiment: Eissa, N. and Liebman, J. (1996), "Labor Supply Response to the Earned Income Tax Credit", *Quarterly Journal of Economics*, CXI, 605-637

- tax-credit program for single parents in the US
- reformed many times

2. The New Deal for Young People in the UK: Blundell, R., Costa Dias, M., Meghir, C. and Van Reenen, J. (2004), "Evaluating the Employment Impact of a Mandatory Job Search Assistance Program", *Journal of the European Economics Association*, 2(4), 596-606

- using the pilot control area comparisons as a natural experiment.

Differential trend adjusted DID estimator

The possibility of differential trends motivates the ‘differential trend adjusted DID estimator’.

Suppose we suspect that the common trend assumption of DID does not hold but can assume that selection into treatment is independent of the temporary individual-specific effect, under differential trends

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

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

The DID estimator now identifies

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

which does not recover the true ATT unless $k^1 = k^0$.

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

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

This means that there is a point in history where the relative conditions of the two groups being compared, treatments and controls, evolves similarly to what they do in the pre-post reform period, (t_0, t_1) .

This is the differentially adjusted ATT estimator proposed by Bell, Blundell and Van Reenen (1999)

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

Non-linear Differences in Differences

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

$$y_{it} = \mathbf{1}(\beta + \alpha_i d_{it} + u_{it} > 0) \quad (14)$$

where $\mathbf{1}(\cdot)$ is usual indicator function. Suppose F is the binary probability model.

The trend can be identified from the comparison of non-treated before and after treatment:

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

Then

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

(iii) The matching estimator (M)

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

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

To start we need to include some observable regressors in the outcome equation in a very general way. The covariates X explain part of the 'residual' term u and part of the idiosyncratic gains from treatment:

$$\begin{aligned} y_i^1 &= \beta + u(X_i) + \alpha(X_i) + [(u_i - u(X_i)) + (\alpha_i - \alpha(X_i))] \\ y_i^0 &= \beta + u(X_i) + (u_i - u(X_i)) \end{aligned} \quad (15)$$

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

The choice of the appropriate matching variables, X , is a delicate issue. To the extent that the goal of evaluation methods is to control for selection, the correct information is that available to the individual at the time of deciding about participation. What remains unexplained is random with respect to treatment status.

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

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

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

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

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

This means that, conditional on X , treated and non-treated individuals are comparable in what respect to the outcome y in the non-treatment case. Thus, there can be no selection on the unobservable term u_i .

Assumption M1 implies a conditional version of the randomization hypothesis (R1)

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

which, under the usual hypothesis of exogeneity of X yields $E [u_i]$. Again, nothing like the randomization hypothesis (R2) is required to identify the ATT, which means that selection on the unobservable gains can be accommodated by matching.

The implication of (M1) is that for each treated observation (y^1) we can look for a non-treated (set of) observation(s) (y^0) with the same X -realization and be certain that such y^0 constitutes the correct counterfactual. Thus, matching is explicitly a process of re-building an experimental data set. Its ability to do so, however, depends on the availability of the counterfactual. That is, we need to ensure that each treated observation can be reproduced among the non-treated. This is captured in the second matching assumption.

M2: All treated individuals have a counterpart on the non-treated population and anyone constitutes a possible participant:

$$0 < P(d_i = 1 | X_i) < 1$$

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

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

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

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

$$\hat{\alpha}^M = \sum_{i \in T} \left\{ y_i - \sum_{j \in C} \varpi_{ij} y_j \right\} \omega_i \quad (16)$$

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

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

$$\begin{aligned} E[u_i | d_i, X_i] &= E[u_i | X_i] \\ E[\alpha_i | d_i, X_i] &= E[\alpha_i | X_i] \end{aligned} \quad (17)$$

Under (17) the ATE over the common support S is

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

using the distribution of the X 's over the whole population, $F_X(X)$.

Propensity score matching

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

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

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

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

The **Nearest Neighbor Matching** assigns a weight 1 to the closest non-treated observation and 0 to all others. A widespread alternative is to use a certain number of the closest non-treated observations to match the treated, generally the 10 closest observations. This reduces the variability of the nearest neighbor estimator.

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

The complexity of propensity score matching requires bootstrapping to be used in computing the standard errors for the effect of treatment. The problem with the nearest neighbor technique is that bootstrapping is not guaranteed to deliver consistent estimates since choosing only 1 (or a fixed number of) match(es) per treated individual means that the quality of the match does not necessarily improve as the sample (of controls) gets bigger. The same is not true for kernel and local linear matching as with these estimator the sample of matched controls expands with the sample size (for a thoroughly discussion of bootstrapping see Horowitz, 2001). The general form of the matching estimator is not altered by the sort of weights one decides to apply. As before, it is given by $\hat{\alpha}^M$ in (16).

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

Weaknesses of matching

The main weaknesses of matching are data driven: its availability and our ability to select the right information. The common support assumption (M2) ensures that the missing counterfactual can be constructed from the population of non-treated. What (M2) does not ensure is that the same counterfactual exists in the sample. If some of the treated observations cannot be matched, the definition of the estimated parameter becomes unclear. It is the average impact over some subgroup of the treated, but such subgroup may be difficult to define.

Combining matching and DID (MDID)

We start by decomposing the unobservable term u into a fixed effect (ϕ), macro shock (ψ) and an idiosyncratic transitory shock (ϵ)

$$\begin{aligned} y_{it}^1 &= \beta + u(X_i) + \alpha(X_i) + [(\phi_i + m_t + \epsilon_{it} - u(X_i)) + (\alpha_i - \alpha(X_i))] \\ y_{it}^0 &= \beta + u(X_i) + (\phi_i + m_t + \epsilon_{it} - u(X_i)) \end{aligned} \quad (18)$$

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

DID1: Conditional on the set of observables X , the before-after difference in the unobservable u is independent of the participation status,

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

which, under specification (18) is the same as assuming

$$\epsilon_{it} \perp d_{it_1} \mid X_i$$

where $t_0 < k < t_1$.

Assumption (MDID1) is not enough to ensure identifiability of ATT. Just as in the matching case, we also need to impose a common support hypothesis. This will be the same as (M2) when longitudinal data is available. If we only dispose of repeated cross-section data, however, we will need to strengthen it to ensure that the treated group can be reproduced in all three control groups characterized by treatment status before and after the program.

Thus:

DID2: All treated individuals have a counterpart on the non-treated population before and after the treatment and anyone constitutes a possible participant,

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

where $P(d_{it_1} = 1 \mid X_i, t)$ is the probability that an individual observed at time t with characteristics X_i had been treated would the same observation correspond to time t_1 .

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

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

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

With repeated cross-section data, however, matching must be performed over the three control groups: treated and non-treated at t_0 and non-treated at t_1 . In this case, the matching-DID estimator would be

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

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