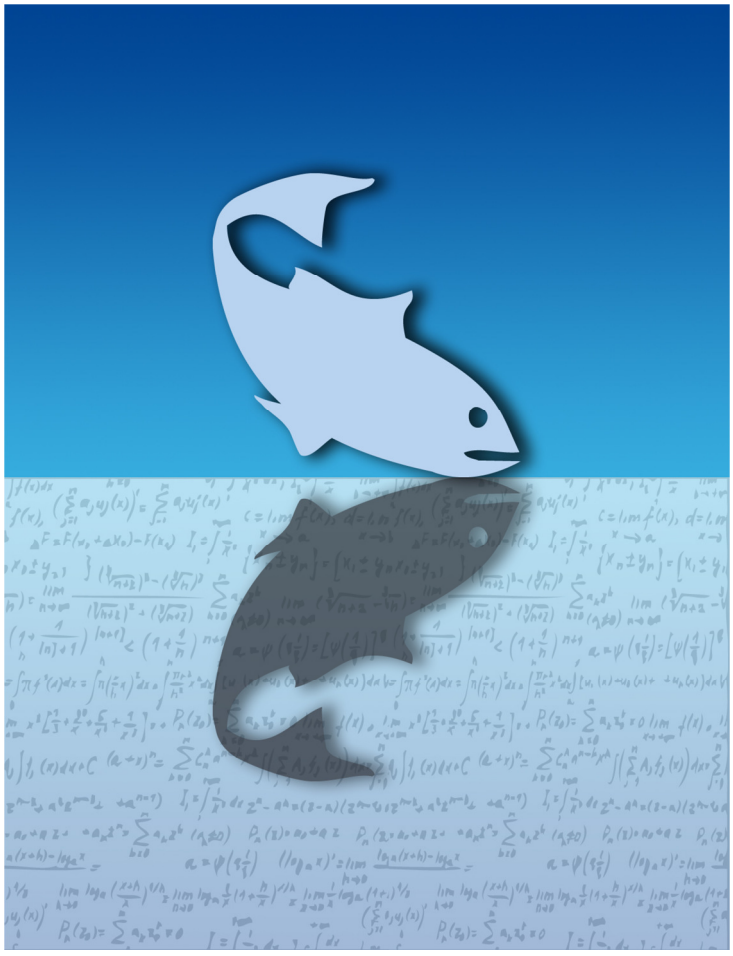
European Commission

JRC SCIENTIFIC AND POLICY REPORTS

A note on the impact evaluation of public policies: the counterfactual analysis

Massimo Loi and Margarida Rodrigues

2012



Report EUR 25519 EN

Joint Research Centre

European Commission
Joint Research Centre
Institute for the Protection and Security of the Citizen

Contact information
Forename Surname
Address: Joint Research Centre, Via Enrico Fermi 2749, TP 361, 21027 Ispra (VA), Italy
E-mail: margarida.rodrigues@jrc.ec.europa.eu
Tel.: +39 0332 78 5633
Fax: +39 0332 78 5733

http://ipsc.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC74778

EUR 25519 EN

ISBN 978-92-79-26425-2

ISSN 1831-9424

doi:10.2788/50327

Luxembourg: Publications Office of the European Union, 2012

Printed in Italy

Contents

1. Introduction

Public policies or interventions are implemented with the expectation of improving the situation of the individuals affected by them, but the extent to which they do so can only be assessed by undertaking an adequate policy evaluation exercise.

Consider the example of publicly-provided training programmes offered to unemployed individuals, aiming at updating and increasing the participants' skills and hopefully contributing significantly to their probability of finding a job. Measuring the effect of these programmes on an individual's future employability may seem, at a first glance, a simple exercise, but in fact it is not a trivial task. Should the outcomes of the participants be compared to their pre-programme situations? Or instead, should the outcomes of the participants be compared with those of the non-participants? What if none of these alternatives are correct? What is the best approach to measure the effect of the programme on the outcomes of participants? This report aims to present the policy evaluation framework, to explain why the two approaches just proposed are usually wrong and to describe four counterfactual evaluation methods currently used to measure policy effects.

The need to quantitatively identify the effects of a policy is nowadays indisputable, as it allows the measurement of its real effects and a comparison with the expected ones. Furthermore, the policy evaluation exercise gives essential and irreplaceable evidence for future policies in the same area. As such, it is an important ex-ante policy impact assessment tool, provided that past interventions and contextual situations are similar to the ones under consideration. As a result, in the recent decades the policy evaluation literature has gained increasing importance and new methodologies have been developed to identify the causal policy effects.

The aim of policy evaluation is to measure the causal effect of a policy on outcomes of interest, on which it is expected to have an impact. The policy's causal effect is defined as the difference between the outcome of the units affected by the policy (the actual situation) and the outcome that these *same* units would experience had the policy not been implemented. The fundamental evaluation problem is that we cannot observe simultaneously the *same* unit in the two scenarios, i.e. the scenario in which the policy is not implemented – the counterfactual – is an elusive one to produce or simulate.

What is needed is an adequate control group that is as similar as possible to the affected one, so that the policy effect can be identified by the comparison between the outcomes of these two groups. Finding such a control group is not an easy task. An analyst may be tempted by one of the two following naïve approaches: i) comparing the outcome of interest of the affected units before and after the intervention; and ii) comparing units affected by the intervention with those not affected. However, neither approach will identify the policy causal effect. In the former case,

the outcome of interest might have been affected by factors, other than the policy, which changed over time. Therefore it is possible that the policy had an effect on the outcome even if the outcome did not change or, on the contrary, that the effect was null even if the outcome did change over time. The latter approach is also in general not suitable because affected and unaffected units are typically different even in the absence of the policy. This is particularly the case when the policy is targeted towards a specific population or when the units can select themselves into participation.

The counterfactual analysis methodologies aim at identifying an adequate control group and, as a consequence, the counterfactual outcome and the policy effect. These methods became the standard approach to identify the causal policy effects in most institutions and international organizations in the last decades, with the World Bank playing a leading role[1]. However, the European Commission uses the counterfactual analysis somewhat parsimoniously in its evaluation and ex-ante policy impact assessment guidelines, which still rely on simple impact indicators (Martini, 2008) and on baseline scenarios that, in most cases, are not defined according to the counterfactual framework.

In this report we describe the policy evaluation framework and the different counterfactual analysis evaluation strategies: propensity score matching, regression discontinuity design, difference-in-differences and instrumental variables. For each method we present the main assumptions it relies on and the data requirements. These methodologies apply to any type of policy and, in general, to any type of intervention (for instance, a programme or treatment[2]). A selection of papers applying each approach in the context of labour market interventions is also included[3].

The rest of the report is organised as follows. In the next section we introduce the basic concepts of any policy evaluation framework. Section 3 presents the conditions for assessing the causal effect of policy interventions and the balancing methods available. Sections 4 to 8 describe each of the methods. Finally, section 9 concludes.

[1] See for instance: Khandker, S.R., Koolwal, G.B., & Samad, H.A. (2010). Handbook on impact evaluation. Quantitative methods and practices. Washington D.C.
More generally, in Europe the interest toward the techniques to evaluate the effect of public policies is quite recent and often liked to specific policies (e.g. it is quite emblematic the case of the Hartz reforms implemented in Germany between 2002 and 2005).

[2] The counterfactual approach was first developed to estimate the effect of medical and pharmaceutical treatments on specific target groups. Thus, most of the terminology related to this methodologies, as for instance the terms "treatment", "treated" and "control group", come from the medical field.

[3] Unless otherwise specified, the abstracts of the papers presented have been taken from Scopus (http://www.scopus.com/home.url).

2. Basic concepts

2.1 Policy intervention

A policy is an intervention targeted to a specific population with the purpose of inducing a change in a defined state and/or behaviour. This definition highlights the three constitutive elements of a policy intervention:

a) A target population: a well-defined set of units (e.g. persons, households, firms, geographic areas) upon which the intervention will operate at a particular time;
b) An intervention: an action, or a set of actions (≡ treatment), whose effect on the outcome the analyst wishes to assess relative to non-intervention. For sake of simplicity, this report considers only interventions that consist of a single treatment that can be represented by a binary variable (treatment vs. non-treatment or participation vs. non-participation). Members of the population that are exposed to the intervention are labelled as participants (≡ treated), while those who do not take part in the programme are labelled as non-participants (≡ non-treated);
c) An outcome variable: an observable and measurable characteristic of the units of the population on which the intervention may have an effect (≡ impact).

Example. The Active Labour Market Programs (ALMP) include three broad classes of interventions – training programs, subsidised employment programs, and job search assistance programs – that are used in many countries to help labour market participants to find and retain better jobs. The primary goal of an ALMP evaluation is to provide objective, scientifically-based evidence on the post-program impact of the programme. In most cases, an evaluation attempts to measure the extent to which participation raised employment and/or earning of participants at some point after the completion of the programme (Card et al., 2011).

2.2 Causality

Impact evaluation is essentially the study of cause-and-effect relationships. It aims to answer the key question: *"Does participation in the programme affect the outcome variable?"* In other words, to what extent can the variation observed in the outcome be attributed to the intervention, given that all other things are held constant? The answer to this question is obtained by subtracting the value of the outcome after exposure to the intervention from the value it would

have had in absence of the treatment (≡ net difference). In this context, causality refers to the net gain or loss observed in the outcome of the treated units that can be attributed to the intervention.

2.3 Randomised experiments vs. quasi-experiments

The most valid way to establish the effects of an intervention is a randomised field experiment, often called the "gold standard" research design for policy evaluation. With this design the units in the target population are randomly assigned to the control and to the intervention groups so that each unit has the same probability to be in either of these treatment statuses; outcomes are then observed for both groups, with differences being attributed to the intervention.

Example. In 2007, the French Labour Ministry organised a randomised experiment aiming to evaluate the delegation to private providers of placement services for young graduates that had spent at least six months in unemployment. Following the previous discussion, the target population of this intervention consists of young graduates that had spent at least six months in unemployment, the treatment is the exposition to private providers of placement services and the non-treatment is the French historical public placement agency (ANPE).

This experiment was realised on a large scale from August 2007 to June 2009 and involved 10.000 young graduates and 235 local public unemployment agencies scattered into 10 administrative regions. One of the main innovations of the study rested on a two-level randomisation. The first randomisation was at the area level. In a first stage, before the experiment started, each one of the 235 local employment agencies was randomly assigned the proportion *P* of jobseekers that were going to be assigned to treatment: either 0%, 25%, 50%, 75% or 100%. The second randomisation was within each treated area: eligible unemployed graduates were randomly selected, given the specified fraction (except of course for 100% areas, where everybody was assigned to treatment). Jobseekers assigned to treatment were offered the opportunity to be followed and guided by a caseworker in a private placement operator. For those who were assigned to the control group, nothing changed: they were still followed by the ANPE. The main results can be summarised as follows: the programme had indeed a strong impact on the employment situation of young job-seekers 8 months after the treatment (see Crépon et al., 2011).

Example. Job Corps is the US largest vocationally-focused education and training programme for disadvantaged youths. Applicants must meet 11 criteria to be eligible for the programme: (1) be aged 16 to 24; (2) be a legal US resident; (3) be economically disadvantaged; (4) live in an environment characterised by a disruptive home life, high crime rates, or limited job opportunities; (5) need additional education, training, job skills; (6) be free of serious behavioural problems; (7) have a clean health history; (8) have an adequate child care plan (for those with children); (9) have registered with the Selective Service Board (if applicable); (10) have parental consent (for minors); and (11) be judged to have the capacity and aspirations to participate in Job Corps.

The heart of Job Corps is the services provided at training centres where participants receive intensive vocational training, academic education and a wide range of other services, including counselling, social skills training, and health education. Furthermore, a unique feature of Job Corps is that most participants reside at the centre while training.

Schochet et al. (2008) estimated the effectiveness of this programme using an experimental approach. This study is based on a representative US sample of eligible programme applicants. With a few exceptions, all eligible youths that applied to Job Corps between November 1994 and December 1995 were randomly assigned to either a programme or control group. Programme group members (9.409 youths) were allowed to enrol in Job Corps; control group members (5.977 youths) were not for three years after random assignment. Using this approach, the authors find that Job Corps participation increases educational attainment, reduces criminal activity and increases earnings for several post-programme years. However, the authors conclude that these benefits are not sustainable (i.e. the programme's costs exceed programme the resulting benefits), except for older workers.

A major obstacle to randomised experiments is that they are usually costly, time-consuming and (especially in Europe) considered to be non-ethical. Concerning this last point, randomisation is often seen as arbitrarily and capriciously depriving control groups from the possibility to be exposed to a treatment that in principle should bring some benefit. When the randomisation design is not feasible, there are alternative designs that an evaluator can choose. These approaches, called quasi-experimental or observational studies, compare target units receiving the intervention with a control group of selected, non-randomly assigned targets or potential targets that do not receive the intervention. If the latter resemble the intervention group on relevant characteristics, or can be adjusted to resemble it, then the programme effects can be assessed with a reasonable degree of confidence (Rossi et al., 2004: 233-300).

2.4 The fundamental evaluation problem

The main question of the evaluation problem is whether the outcome variable for a unit in the target population is affected by the participation to the programme. For instance, the main outcome of an ALMP for a participant to the programme could be the increased probability of employment or higher earnings after a certain number of months from the treatment. In this context, we would like to know the value of the participant's outcome in the actual situation and the value of the outcome if (s)he had not participated in the programme. The fundamental evaluation problem arises because we never observe the same person in both states (i.e. participation and non-participation) at the same time (Hujer and Caliendo, 2000). Therefore, inference about the impact of a policy on the outcome of a unit in the target group involves speculation about how this unit would have performed in the absence of the treatment. The standard framework to formalise this problem is the **counterfactual approach** (also called the "potential outcome approach" or "Roy-Rubin model"). *A counterfactual is a potential outcome, or the state of the affairs that would have happened in the absence of the cause* (Shadish et al., 2002). Thus, for a treated unit, a counterfactual is the potential outcome under the non-treatment state; conversely, for a non-participant unit, the counterfactual is the potential outcome under the treatment state. The key assumption of the counterfactual framework is that each unit in the target population has a potential outcome under each treatment state, even though each individual can be observed in only one treatment state at any point in time.

2.5 Formalising the counterfactual approach

The researcher observes the set of variables (Y, X, D) for each unit in the population: (Y_i, X_i, D_i), for $i = 1, \ldots, N$. D is a dummy variable indicating whether the treatment has been received ($D = 1$) or not ($D = 0$). Y is the outcome variable, i.e. the variable that is expected to be affected by the treatment. X is a set of observable characteristics of the unit and, eventually, of higher levels such as households or local/regional characteristics[4].

Let $(Y_1, Y_0)_i$ be the two *potential* outcomes on the *i-t*h population unit of being treated or not treated, respectively. If a specific member of the target population receives the treatment then Y_1 is observable (≡ factual), while Y_0 is irreversibly non-observable and corresponds to what we would have observed if this unit had not received the intervention (≡ counterfactual). Similarly, if a specific member of the target population is not exposed to the treatment, it is possible to observe only Y_0 (≡ factual); in this case Y_1 is the outcome of that specific unit in the case it had been treated (≡ counterfactual).

The identity that relates Y_i, the real outcome that is observed on unit *i* of the target population, to the potential outcomes of the same unit is the following:

[4] *X* defines the set of observable characteristics and *x* a single observable variable. These variables are usually called explanatory variables or covariates. Examples of observable characteristics in labour market applied research are age, gender, level of education, employment status, income, and urban/rural area, among others.

$$[1] \qquad Y_i = Y_{1i}D_i + Y_{0i}(1 - D_i) = \begin{cases} Y_{1i} & if\ D_i = 1 \\ Y_{0i} & if\ D_i = 0 \end{cases}$$

where D_i is the treatment status of the i-th population unit: $D_i = 1$ if this unit received the intervention and 0 otherwise. Basically, the identity of equation [1] indicates which of the two outcomes will be observed in the data (Y_{1i} or Y_{0i}) depending on the treatment condition ($D_i = 1$ or $D_i = 0$). The key message of this equation is that to infer a causal relationship between D_i (the cause) and Y_i (the outcome) the analyst cannot directly link Y_{1i} and D_i under the condition $D_i = 1$; instead the analyst must check the outcome of Y_{0i} under the condition $D_i = 0$, and compare Y_{0i} with Y_{1i} (Guo and Fraser, 2010).

Example. We might assume that a worker i with low education (i.e. ISCED 1 or 2) has also a low income. Here the treatment variable is $D_i = 1$ if the worker has a low level of education; the income $Y_{1i} < p$ if the worker has low income, where p is a cutoff defining a low income, and $Y_{1i} > p$ otherwise. To make a causal statement that being poorly educated ($D_i = 1$) causes low income $Y_{1i} < p$, the researcher must examine the outcome under the status of not being poorly educated. That is, the task is to determine the worker's salary Y_{0i} under the condition $D_i = 0$, and ask the question, "What would have happened had the worker had a medium-to-high level of education?" If the answer to the question is $Y_{0i} > p$, than the researcher can have confidence that $D_i = 1$ causes $Y_{1i} < p$. The most critical issue is that Y_{0i} is not observed (adapted from Guo and Fraser, 2010).

The estimation of a plausible value for the counterfactual Y_{0i} is the central object of the methods presented in this document.

2.6 Average treatment effect

Given that it is typically impossible to calculate individual-unit causal effects (and it is also less interesting from the policy point of view), the literature focuses its attention on the estimation of aggregated causal effects, usually alternative average causal effects. The two most commonly used are: the population average treatment effect (ATE) and the average effect on units in the target population that were assigned to treatment (ATT). With $E(.)$ denoting the expectation operator, these two average parameters can be expressed as follows:

$$[2] \qquad \text{ATE} = E(Y_{1i} - Y_{0i}) = E(Y_{1i}) - E(Y_{0i})$$

$$[3] \qquad \text{ATT} = E(Y_{1i} - Y_{0i}|D_i = 1) = E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)$$

Identity [2], the ATE, represents the average effect that would result from having all population members taking part in the programme. The ATE is the parameter of interest when the

programme under consideration has universal applicability, in the sense that all the units in the population are exposed to the treatment. The ATT, identity [3], measures the average treatment effect for the units actually exposed to the intervention and is the parameter of major interest for policy evaluation. It is important to note that the first term of the identity defining the ATT ($E(Y_{1i}|D_i = 1)$, the average effect on the treated) is observable (≡ factual outcome), while the second term ($E(Y_{0i}|D_i = 1)$, the average effect on the treated in the case they had not been treated) is not (≡ counterfactual outcome). Therefore the ATT cannot be directly identified.

This report concerns solely the estimation of the ATT (and not of the ATE) as it is the most interesting parameter for a policy maker.

2.7 Selection bias

As the outcome of the counterfactual is not observable, one could take instead the outcome of non-participants as an approximation to the outcome that participants would have had without treatment. This would be a correct approach if (and only if) participants and non-participants have similar characteristics, i.e. if they are comparable *a priori*, had the treatment not been implemented. In general, however, participants and non-participants differ in crucial characteristics that are related both with the participation status and the outcome[5]. This problem is known as "selection bias": a good example is the case where highly-skilled individuals have a higher probability of entering a training programme and also have a higher probability of finding a job (Caliendo and Kopeinig, 2008).

Example. Suppose we have data on wages and personal characteristics of workers that include whether or not a worker is a union member. A naïve way of estimating the effect of unionisation on wages is to take the difference between the average wage of non-unionised workers ($D_i = 0$) and the average wage of unionized workers ($D_i = 1$). The problem with this approach relies on the fact that it treats the unionisation decision as exogenous, i.e. that is completely unrelated with other individual characteristics of the workers. In fact, there are many factors affecting a worker's decision to join the union (Guo and Fraser, 2010) and therefore the selection bias should be taken into account.

[5] Random assignment to the treatment status solves the selection problem because random assignment makes the treatment status ($D_i = 1$ or $D_i = 0$) independent of potential outcomes (see, for example, Angrist and Pischke, 2009: 15-22). Randomising eliminates all systematic pre-existing group differences, because only chance determines which units are assigned to a given group. Consequently, each experimental group has the same expected values for all characteristics, observables or non-observables. Randomisation of a given sample may produce groups that differ by chance, however. These differences are random errors, not biases. The laws of probability ensure that the larger the number the units in the target population, the smaller pre-existing group differences are likely to be.

Let us consider the observed difference between the average outcome of the treated units and the average outcome of the non-treated units:

$$[4] \qquad E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0)$$

and subtract and add to it the counterfactual outcome for the treated units ($E(Y_{0i}|D_i = 1)$):

$$[5] \qquad E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) =$$

$$= \underbrace{[E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)]}_{} + \underbrace{[E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)]}_{}$$

selection bias, i.e. the difference between treated and non-treated that would have been observed even if the policy had not taken place and depends on pre-existing differences between the two groups

Identity [5] tells us that the observed difference is equal to the average treatment effect on the treated units if and only if there is no selection bias. Thus, whether or not the observed difference in means between treated and non-treated units (≡ difference in mean factual outcomes) corresponds to the average treatment effect on the treated depends on having or not having selection bias.

In the literature, estimating the counterfactual corresponds to overcoming the selection bias problem, therefore, all the policy evaluation methods that have been developed so far aim at solving it.

3. Conditions for assessing the causal effect of public policies

This section presents the two main assumptions that a researcher has to make to detect the causal effect of a public policy.

3.1 Stable unit treatment value assumption (SUTVA)

The first assumption embodies two conditions: (a) there is only one form of treatment and one form of non-treatment for each unit (≡ the treatment should not be "blurry"), and (b) there is no interference among units (≡ no interaction between units), in the sense that the outcome experienced by unit *i* is not influenced by the treatment state nor the outcome of any other member of the population.

3.2 Common support

This assumption also embodies two conditions, the first being that both treated and non-treated units are observed. The second assumption states that for each treated unit there is a comparable non-treated unit, i.e. there is a non-treated unit with similar levels of the observable characteristics, X, and with a similar probability of being treated. Formally:

$$[6] \qquad \Pr(D_i = 1|X_i) < 1,$$

that is, the probability of being exposed to the intervention for the unit i given $X_i = x$ (i.e. gender = female) is below 1 meaning that there is at least one unit j with the same characteristics (i.e. gender = female) that is not exposed to the treatment ($D_j = 0$).

If the common support holds partially, that is just for a subset of values of X (i.e. only for females or only for males), the researcher has to restrict the analysis only to that subset.

4. Balancing methods

This section introduces four conventional methods that help reducing the selection bias (i.e. balancing the data): propensity score matching, regression discontinuity design, difference-in-differences, and instrumental variables.

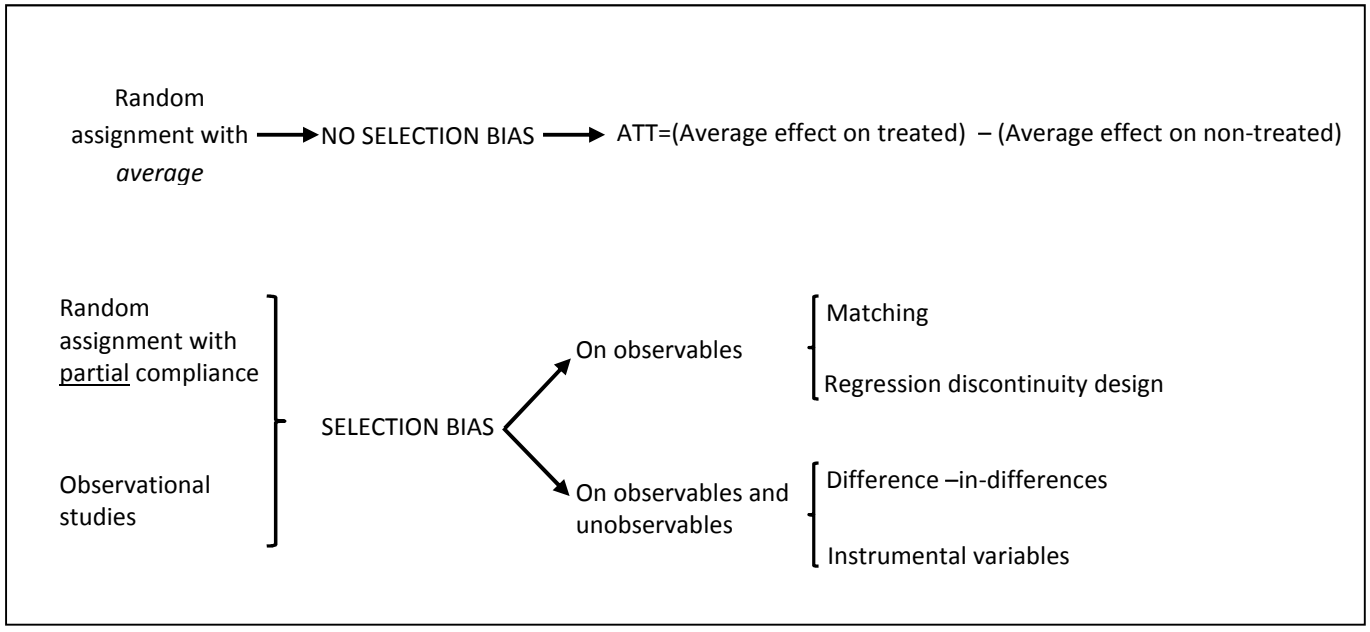


Figure 1 - Balancing methods

The first two methods – the matching and the regression discontinuity design – can be applied when, in addition to the SUTVA and to the common support assumptions, the analyst knows and observes in the data all variables that influence the exposure to the treatment and the potential outcomes. This additional assumption – known as *selection on observables*[6] – is difficult to defend in fields like education and the labour market where unobservable characteristics (like ability, motivation, and intelligence) more likely dictate individual behaviours. In these cases, when the analyst has the suspicion that the selection into treatment is driven by observable and unobservable factors, two other methods are available: the difference-in-differences or the instrumental variables approach (Figure 1).

For sake of simplicity these methods are presented one by one, but they can be combined (indeed, they usually are) to tackle the evaluation problem under study. Unfortunately there is no established best practice on this: the evaluator has to decide whether and how to combine these methods.

5. Propensity score matching (PSM)[7]

The method of matching has achieved popularity more recently as a tool of evaluation. It assumes that selection can be explained purely in terms of observable characteristics. Applying this method is, in principle, simple. For every unit in the treatment group a matching unit (≡ twin) is found among the non-treatment group. The choice of the match is dictated by observable characteristics. What is required is to match each unit exposed to the treatment with one or more non-treated units sharing similar observable characteristics. The degree of similarities between different units is measured on the basis of the probability of being exposed to the intervention given a set of observable characteristics not affected by the programme, the so called propensity score. The idea is to find, from a large group of non-participants, units that are observationally similar to participants in terms of characteristics not affected by the intervention. The mean effect of treatment can then be calculated as the average difference in outcomes between the treated and non-treated units after matching.

The aim of this section is to introduce the steps an analyst has to follow in order to implement propensity score matching (subsection 5.1), to present some recent applications of this method (subsection 5.2), and to highlight the main hypotheses, data requirements and pro and cons of this method (subsection 5.3).

---

[6] A variety of terms have emerged to describe this assumption: *unconfoundedness* (Rosenbaum and Rubin,1983), *selection on observables* (Barnow et al., 1980), *conditional independence* (Lechner, 1999), and *exogeneity* (Imbens, 2004).

[7] From Caliendo M., Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1): 31-72.

5.1 Implementation steps

Implementing PSM can be summarised in the 5-step process shown in Figure 2.

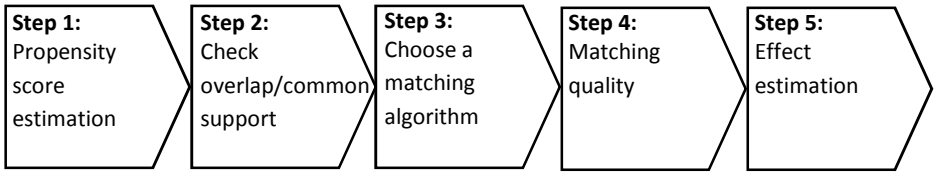


Figure 2 - PSM: implementation steps

Step 1: Propensity score estimation. The propensity score is the probability of a unit in the target group (treated and control units) to be treated given its observed characteristics X_i; formally: $\Pr(D_i = 1|X_i)$. The propensity score is a balancing score in the sense that, as demonstrated by Rosenbaum and Rubin (1983), if two units have similar propensity scores, than they are also similar with respect the set of covariates X used for its estimation.

Step 2: Check overlap and common support. Comparing the incomparable must be avoided, i.e. only the subset of the comparison group that is comparable to the treatment group should be used in the analysis.

Hence, an important step is to check if there is at least one treated unit and one non-treated unit for each value of the propensity score. Several methods are suggested in the literature, but the most straightforward one is a visual analysis of the density distribution of the propensity score in the two groups. Another possible method is based on comparing the minima and maxima of the propensity score in the treated and in the non-treated group. Both approaches require deleting all the observations whose propensity score is smaller than the minimum and larger than the maximum in the opposite group.

Step 3: Choose a matching algorithm. The next step consists of matching treated and non-treated units that have similar propensity scores using an appropriate algorithm. Propensity score matching algorithms differ not only in the way they measure the degree of similarity between treated and non-treated units (i.e. the way they find twins between these two groups) but also with respect to the weight they assign to the matched units. The aim of this report is not to discuss the technical details of each estimator; rather to present the general ideas of each algorithm.

a) Nearest-neighbour matching. The treated unit is matched with the unit in the comparison group that presents the closest estimated propensity score. Two variants are possible:

matching *with replacement* (an untreated unit can be used more than once as a match) and matching *without replacement* (an untreated unit can be used only once as a match). A problem related to matching without replacement is that estimates depend on the order in which observations get matched. Hence, when using this approach it should be ensured that the ordering is done randomly.

b) Calliper and radius matching. Nearest-neighbour matching faces the risk of bad matches, if the closest neighbour is not sufficiently similar. This can be avoided by imposing the condition that, in order to be matched, the propensity score of treated and non-treated units should not differ, for example, by more than 5%. This tolerance level (5%) is called the *calliper*. The *radius matching* approach is a variant of calliper matching – the basic idea of this variant is to use not only the nearest neighbour within each calliper but all the units that are within the calliper.

c) Kernel matching. The two matching algorithms discussed above have in common that only some observations from the comparison group are used to construct the counterfactual outcome of a treated unit. *Kernel matching* uses weighted averages of all individuals in the control group to construct the counterfactual outcome. Weights depend on the distance between each individual from the control group and the unit exposed to the treatment for which the counterfactual is estimated. The kernel function assigns higher weight to observations close in terms of propensity score to a treated individual and lower weight to more distant observations.

The choice between different matching algorithms implies a trade-off between bias and variance reduction. For instance, nearest-neighbour matching only uses the participant and its closest neighbour. Therefore it minimises the bias but might also involve an efficiency loss, since a large number of close neighbours is disregarded. Kernel-based matching on the other hand uses more (all) non-participants for each participant, thereby reducing the variance but possibly increasing the bias. Finally, using the same non-treated unit more than once (nearest neighbour matching with replacement) can possibly improve matching quality, but it increases the variance (Caliendo et al, 2005).

Step 4: Matching quality. The quality of the matching procedure is evaluated on the basis of its capability in balancing the control and the treatment groups with respect to the covariates used for the propensity score estimation. There are several procedures for this. The basic idea of all approaches is to compare the distribution of these covariates in the two groups before and after matching on the propensity score. If there are significant differences after matching, than matching on the propensity score was not (completely) successful in making the groups comparable and remedial measures have to be taken.

Rosenbaum and Rubin (1985) highlight that a good matching procedure should reduce the standardised bias for each of the covariates used in the estimation of the propensity scores. Thus,

this approach requires comparing the standardised bias for each covariate *x* before and after matching. The standardised bias before matching is given by:

$$[7] \qquad \text{Standardised bias}_{\text{ before}} = \frac{\bar{x}_1 - \bar{x}_0}{\sqrt{0.5(V_1(x) - V_0(x))}}$$

The standardised bias after matching is given by:

$$[8] \qquad \text{Standardised bias}_{\text{ after}} = \frac{\bar{x}_{1M} - \bar{x}_{0M}}{\sqrt{0.5(V_{1M}(x) - V_{0M}(x))}}$$

where $\bar{x}_1$ (V_1) is the mean (variance) in the treatment group before matching and $\bar{x}_0$ (V_0) the analogue for the control group. $\bar{x}_{1M}$ (V_{1M}) and $\bar{x}_{0M}$ (V_{0M}) are the corresponding values for the matched samples. Even though this method does not provide any clear indication for the success of the matching procedure, most empirical studies consider as sufficient a standardised bias below 3% or 5% after matching.

A similar approach uses a two-sample *t*-test to check if there are significant differences in covariate means for both groups. After matching the covariates should be balanced in both the treatment and the non-treatment group therefore no significant difference should be found.

Additionally, Sianesi (2004) suggests to re-estimate the propensity score in the matched sample and compare the *pseudo-*R^2's before and after matching. After matching there should be no systematic differences in the distribution of the covariates between both groups. Therefore, the *pseudo-*R^2 after matching should be fairly low. The same can be done inspecting the *F-statistics* before and after matching. In fact, these statistics indicate the joint significance of all regressors used for the estimation of the propensity score.

Table 1. Pros and cons of the approaches commonly used to evaluate the quality of the matching procedure

	Pros	Cons
Standardised bias	Easy to compute	- To be performed for each covariate used for the propensity score estimation - No objective indication of the success of the matching procedure
t-test	Easy to compute	- To be performed for each covariate used for the propensity score estimation - The bias reduction is not clearly visible
Joint significance test	- Easy to compute - To be performed only once (jointly on all the covariates used for the propensity score estimation)	

Step 5: Effect estimation. After the match has been judged of acceptable quality, computing the effect becomes a quite easy task: it is enough to calculate the average of the difference between the outcome variable in the treated and non-treated groups.

Before running a *t*-test to check the statistical significance of the effect, however, one needs to compute standard errors. This is not straightforward. The problem is that the estimated variance of the treatment effect should also include the variance due to the estimation of the propensity score, the imputation of the common support, and possibly also the order in which treated individuals are matched. One way to deal with this problem is to use bootstrapping as suggested by Lechner (2002). Even though Imbens (2004) notes that there is in fact little formal evidence to justify bootstrapping, this method is widely applied.

5.2 Selected applications

Rinne, U., Schneider, M., Uhlendorff, A., (2011). Do the skilled and prime-aged unemployed benefit more from training? Effect heterogeneity of public training programmes in Germany. *Applied Economics*, 43 (25): 3465-3494

Abstract. This study analyzes the treatment effects of **public training programs for the unemployed** in Germany. Based on propensity score matching methods we extend the picture that has been sketched in previous studies by estimating treatment effects of medium-term programs for different sub-groups with respect to vocational education and age. Our results indicate that program participation has a positive impact on employment probabilities for all sub-groups. Participants also seem to find more often higher paid jobs than non-participants. However, we find only little evidence for the presence of heterogeneous treatment effects, and the magnitude of the differences is quite small. Our results are thus – at least in part – conflicting with the strategy to increasingly provide training to individuals with better employment prospects.

Huber, M., Lechner, M., Wunsch, C., Walter, T., (2011). Do German welfare-to-work programmes reduce welfare dependency and increase employment? *German Economic Review*, 12(2): 182-204.

During the last decade, many Western economies reformed their welfare systems with the aim of activating welfare recipients by increasing welfare-to-work programmes (WTWP) and job-search enforcement. We evaluate the short-term effects of three important German WTWP implemented after a major reform in January 2005 ('Hartz IV'), namely **short training, further training with a planned duration of up to three months and public workfare programmes**

('One-Euro-Jobs'). Our analysis is based on a combination of a large-scale survey and administrative data that is rich with respect to individual, household, agency level and regional information. We use this richness of the data to base the econometric evaluation on a selection-on-observables approach. We find that short-term training programmes, on average, increase their participants' employment perspectives. There is also considerable effect heterogeneity across different subgroups of participants that could be exploited to improve the allocation of welfare recipients to the specific programmes and thus increase overall programme effectiveness.

Nuria R., Benus, J., (2010). Evaluating Active Labor Market Programs in Romania. *Empirical Economics*, 38 (1): 65-84

Abstract. We evaluate the presence of effects from joining one of four **active labour market programs** in Romania in the late 1990s compared to the no-program state. Using rich survey data and propensity score matching, we find that three programs (training and retraining, small business assistance, and employment and relocation services) had success in improving participants' economic outcomes and were cost-beneficial from society's perspective. In contrast, public employment was found detrimental for the employment prospects of its participants. We also find that there is considerable heterogeneity, which suggests that targeting may improve the effectiveness of these programs.

Jespersen, S.T., Munch, J.R., Skipper, L., (2008). Costs and benefits of Danish active labour market programmes. *Labour Economics*, 15(5): 859-884.

Abstract. Since 1994, unemployed workers in the Danish labour market have participated in **active labour market programmes** on a large scale. This paper contributes with an assessment of costs and benefits of these programmes. Long-term treatment effects are estimated on a very detailed administrative dataset by propensity score matching. For the years 1995 - 2005 it is found that private job training programmes have substantial positive employment and earnings effects, but also public job training ends up with positive earnings effects. Classroom training does not significantly improve employment or earnings prospects in the long run. When the cost side is taken into account, private and public job training still come out with surplusses, while classroom training leads to a deficit.

Sianesi, B. (2008). Differential effects of active labour market programs for the unemployed. *Labour Economics*, 15(3): 392-421.

Abstract. The differential performance of six Swedish **active labour market programs** for the unemployed is investigated in terms of short- and long-term employment probability and un-employment-benefit dependency. Both relative to one another and compared to more intense job

search, the central finding is that the more similar to a regular job, the more effective a program is for its participants. Employment subsidies perform best by far, followed by trainee replacement and, by a long stretch, labour market training. Relief work and two types of work practice schemes appear by contrast to be mainly used to re-qualify for unemployment benefits.

Stenberg, A., Westerlund, O., (2008). Does comprehensive education work for the long-term unemployed? *Labour Economics*, 15(1): 54-67.

Abstract. In this paper we evaluate the **effects of comprehensive adult education on wage earnings of long-term unemployed**, an essentially unexplored issue. We use register data pertaining to a large sample of long-term unemployed in Sweden who enrolled in upper secondary comprehensive adult education. Estimates with propensity score matching indicate that more than one semester of study results in substantial increases in post program annual earnings for both males and females. According to our rough calculations, the social benefits of offering these individuals comprehensive education surpass the costs within five to seven years.

Fitzenberger, B., Speckesser, S., (2007). Employment effects of the provision of specific professional skills and techniques in Germany. *Empirical Economics*, 32(2-3): 529-573.

Abstract. Based on unique administrative data, which has only recently become available, this paper estimates the employment effects of the most important type of **public sector sponsored training** in Germany, namely the provision of specific professional skills and techniques (SPST). Using the inflows into unemployment for the year 1993, the empirical analysis uses local linear matching based on the estimated propensity score to estimate the average treatment effect on the treated of SPST programs by elapsed duration of unemployment. The empirical results show a negative lock-in effect for the period right after the beginning of the program and significantly positive treatment effects on employment rates of about 10 percentage points and above a year after the beginning of the program. The general pattern of the estimated treatment effects is quite similar for the three time intervals of elapsed unemployment considered. The positive effects tend to persist almost completely until the end of our evaluation period. The positive effects are stronger in West Germany compared to East Germany.

Winterhager, H., Heinze, A., Spermann, A., (2006). Deregulating job placement in Europe: A microeconometric evaluation of an innovative voucher scheme in Germany. *Labour Economics*, 13(4): 505-517.

Abstract. **Job placement vouchers** can be regarded as a tool to spur competition between public and private job placement activities. The German government launched this instrument in order to end the public placement monopoly and to subsidize its private competitors. We exploit very

rich administrative data provided for the first time by the Federal Employment Agency and apply propensity score matching as a method to solve the fundamental evaluation problem and to estimate the effect of the vouchers. We find positive treatment effects on the employment probability after one year of 6.5 percentage points in Western Germany and give a measure for deadweight loss.

Nivorozhkin, A. (2005). An evaluation of government-sponsored vocational training programmes for the unemployed in urban Russia. *Cambridge Journal of Economics*, 29(6):1053-1072.

Abstract. This is the first study on the effects of active **labour market programs such as training** in Russia. We use the data from the official unemployment register combined with information from the follow-up survey in a large industrial city in the year 2000. The method of propensity score matching was applied to learn whether participation in the training programmes increased the monthly salaries of participants. The findings suggest that individuals tend to benefit from the participation in the training programmes. However, one year later, this effect disappeared.

Caliendo, M., Hujer, R., Thomsen, S., (2005). *The employment effects of job creation schemes in Germany. A microeconomic evaluation*. Discussion Paper n. 1512, Bonn, IZA. (http://repec.iza.org/dp1512.pdf)

Abstract. In this paper we evaluate the employment effects of **job creation schemes** on the participating individuals in Germany. Job creation schemes are a major element of active labour market policy in Germany and are targeted at long-term unemployed and other hard-to-place individuals. Access to very informative administrative data of the Federal Employment Agency justifies the application of a matching estimator and allows to account for individual (group-specific) and regional effect heterogeneity. We extend previous studies in four directions. First, we are able to evaluate the effects on regular (unsubsidised) employment. Second, we observe the outcome of participants and non-participants for nearly three years after programme start and can therefore analyse mid- and long-term effects. Third, we test the sensitivity of the results with respect to various decisions which have to be made during implementation of the matching estimator, e.g. choosing the matching algorithm or estimating the propensity score. Finally, we check if a possible occurrence of 'unobserved heterogeneity' distorts our interpretation. The overall results are rather discouraging, since the employment effects are negative or insignificant for most of the analysed groups. One notable exception are long-term unemployed individuals who benefit from participation. Hence, one policy implication is to address programmes to this problem group more tightly.

Ohkusa, Y., (2004). Programme evaluation of unemployment benefits in Japan. An average treatment effect approach. *Japan and the World Economy*, 16(1): 95-111.

Abstract. Empirical results show that **unemployment benefits** (UB) recipients significantly change to worse job conditions with respect to wages and firm size, but change to better job conditions with respect to occupation, position, industry, and residence. While the effects for occupation are not significant, UB recipients have a significant tendency to stay the same. In other words, results of other conditions imply that they reduce the reservation wage to get better conditions with respect to occupation, industry, and residence. This means strong inertia in these aspects.

Sianesi, B., (2004). An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s. *Review of Economics and Statistics*, 86(1): 133-155.

Abstract. We investigate the presence of short- and long-term effects from joining a Swedish **labor market program** vis-à-vis more intense job search in open unemployment. Overall, the impact of the program system is found to have been mixed. Joining a program has increased employment rates among participants, a result robust to a misclassification problem in the data. On the other hand it has also allowed participants to remain significantly longer on unemployment benefits and more generally in the unemployment system, this being particularly the case for those entitled individuals entering a program around the time of their unemployment benefits' exhaustion.

Gerfin, M., Lechner, M., Steiger, H., (2002). *Does subsidized temporary employment get the unemployed back to work? An econometric analysis of two different schemes*. Discussion Paper n. 606, Bonn, IZA.
(http://repec.iza.org/dp606.pdf)

Abstract. Subsidized employment is an important tool of active labour market policies to improve the chances of the unemployed to find permanent employment. Using informative individual administrative data we investigate the effects of two different schemes of subsidized temporary employment implemented in Switzerland. One scheme operates as a **non-profit employment programme (EP)**, whereas the other one is a **subsidy for temporary jobs (TEMP)** in firms operating in competitive markets. Using econometric matching methods we find that TEMP is considerably more successful in getting the unemployed back into work than EP. We also find that compared to nonparticipation both programmes are ineffective for unemployed who find job easily anyway as well as for those with short unemployment duration. For unemployed with potentially long unemployment duration and for actual long term unemployed, both programmes may have positive effects, but the effect of TEMP is much larger.

5.3 Synthesis: main hypotheses, data requirements, pros and cons

Propensity score matching (PSM)			
Main hypotheses	**Data requirements**	**Pros**	**Cons**
Selection into treatment determined only by observable characteristics.	The dataset contains all the variables describing the characteristics determining the selection into treatment.	Long tradition in many fields (i.e. medicine, environmental studies, health economics, labour economics and the economics of education).	In fields like the economics of education or labour economics the selection on observables hypothesis is difficult to defend.
	For each value of the propensity score there is at least one treated unit and one non-treated unit.	It controls for a set of covariates that simultaneously determine the selection into treatment.	It is a data-consuming procedure: its conclusions hold only on the subset of matched units.
	(Traditionally) cross-sectional data.	Is a non-parametric approach, therefore very flexible (i.e. matching does not require any functional form assumptions for relationship linking the outcome variable with the covariates).	The external validity (generalisability) of its results decreases when the share of unmatched units increases.
			The estimated variance of the treatment effect should include the variance of the estimated propensity score: to our knowledge, there is no established procedure to do that.

6. Regression discontinuity design (RDD)

Regression discontinuity (RDD) design has many of the assets of a randomised experiment, but can be used when random assignment is not feasible. It is a popular quasi-experimental design that exploits precise knowledge of the rules determining the eligibility into treatment.

According to this design, assignment is solely based on pre-intervention variables observable by the analyst and the probability of participation changes discontinuously as a function of these variables. To fix ideas, consider the case in which a pool of individuals willing to participate is split into two groups according to whether a pre-intervention measure is above or below a known threshold. Those individuals scoring above the threshold are exposed to the intervention, while those who scoring below are denied it (Battistin and Rettore, 2008). RDD can be of two types: sharp and fuzzy.

Sharp RDD is used when the treatment status is a deterministic discontinuous function of a covariate, x (i.e. age, wage, length of the unemployment period, test scores, etc)[8]. Suppose, for example, that:

$$[9] \qquad D_i = \begin{cases} 0 & if \quad x_i < S \\ 1 & if \quad x_i \geq S \end{cases}$$

where S is a known threshold or cutoff[9]. This eligibility mechanism is a deterministic function of x because once we know x_i we know also D_i. Treatment is a discontinuous function of x because no matter how close x gets to S, the treatment status is unchanged until $x = S$ (Angrist and Pischke, 2009)[10]. The main idea behind this design is that units in the target population just below the cutoff (which do not receive the intervention) are good comparisons to those just above the cutoff (which are exposed to the intervention). Thus, in this setting, the analyst can evaluate the impact of an intervention by comparing the average outcomes for the recipients just above the cutoff with those of non-recipients just below it. That is, under certain comparability conditions, the assignment near the cutoff can be seen almost as random. These conditions go under the so called local continuity assumption requiring that:

$$[10] \qquad E(Y_{0i}|x_i = S) \qquad \text{and} \qquad E(Y_{1i}|x_i = S)$$

are continuous in S.

[8] In this section, the variable that determines the eligibility to treatment is represented by x, one of the observed variables X. x_i is the individual realisation of this variable.

[9] This threshold may represent a single characteristic or a composite indicator constructed using multiple characteristics.

[10] From this it follows that there is no value of x_i at which we observe both treated and non-treated units; therefore, although sharp RD can be seen as a special case of selection on observables, common solutions to such selection problem (i.e. propensity score matching methods) are not applicable here because there is no region of common support.

The continuity assumption rules out coincidental functional discontinuities in the relationship between the selection variable x and the outcome of interest Y such as those caused by other interventions that use assignment mechanisms based on the same exact assignment variable and cutoff (Klaauw, 2008).

From this discussion, it follows that the sharp RDD design captures the effect of an intervention only for the subpopulation with values of x near the threshold point (the so called Local Average Treatment Effect, LATE). In case of heterogeneous impacts, the local effect may be very different from the effect at values further away from the threshold. Hence, this approach, while characterised by high internal validity, may produce results that cannot be generalised to the entire target population.

Example. Lemieux and Milligan (2005) studied a peculiar social policy implemented in the Canadian province of Quebec. This policy is peculiar in the sense that it pays much lower social assistance benefits to individuals without children who had not yet attained the age of 30 and much higher social benefits to individuals without children aged 30 and above. Thus, in this case "age" is the selection variable and "age=30" is the cutoff. The authors used the regression discontinuity design approach to estimate the effect of the increased social assistance on employment within these age cohorts.

Fuzzy RDD is used when there is no full compliance to the eligibility rule. In this type of setting, the analyst may observe treated units when $x_i < S$ or non-treated units when $x_i \geq S$ (one way non-compliance) or both, treated units when $x_i < S$ and non-treated units when $x_i \geq S$ (two way non-compliance). In fuzzy RDD the propensity score function $\Pr(D_i = 1|X_i)$ is discontinuous at S as in the case of sharp RDD but, instead of a 0-1 step function, the treatment probability as a function of x now contains a jump at the cutoff that is less than 1. The Fuzzy design discontinuity is highly correlated with treatment leading to an instrumental variable type of setup. More specifically, in this setup the assignment variable x is used as an instrumental variable for program participation (see section 8).

To conclude, both sharp and fuzzy approaches may be invalid research designs if the assignment variable can be manipulated by the units in the target population: in this case the existence of a treatment that is a discontinuous function of an assignment variable is not sufficient to justify the validity of a RDD design.

The remainder of the section is organised as follow. Subsection 6.1 introduces the steps an analyst has to follow to implement RDD; subsection 6.2 presents some recent applications of this

method; finally, subsection 6.3 highlights the main hypotheses, data requirement and pro and cons of the RDD.

6.1 Implementation steps[11]

This section describes the implementation steps an analyst has to follow to apply the regression discontinuity design (Figure 3). The illustration of these steps is based on Serrano-Valerde's application *"The Financing Structure of Corporate R&D - Evidence from Regression Discontinuity Design"* (Serrano-Valerede, 2008) aiming at identifying the effect of R&D subsidies given by the French ANVAR programme (created in 1979) on firms' R&D intensity.

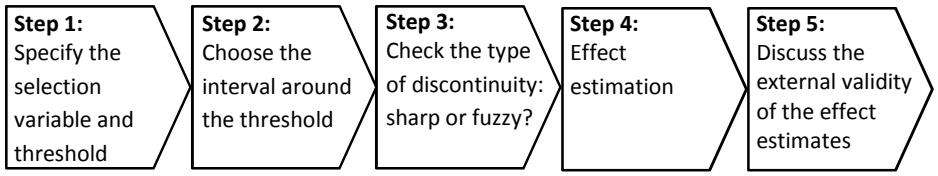


Figure 3 - RDD: implementation steps

Step 1: Specify the selection variable and the threshold. The precondition for the applicability of the RDD is the presence of a continuous selection variable with a cutoff point splitting the units that are eligible for the treatment from the units that are not eligible.

In order for a firm to be eligible for the ANVAR programme it has to be independent from a large business group (henceforth referred to as LBG). Independence is defined with respect to firms' ownership structure, which becomes the selection variable. According to the French law, a firm is independent if less than 25% of its capital is owned by a LBG. Thus 25% becomes the eligibility threshold. A firm owned at 26% by a LBG will be considered ineligible in this setting.

Step 2: Choose the interval around the threshold. The author restricts the analysis to the subsample of private and non-agricultural firms which have $0\% < x < 50\%$ ownership by a LBG. Thus, a firm is eligible for the treatment when the share of its capital that is owned by a LBG does not exceed 25%; similarly, a firm is ineligible when the share of its capital that is owned by a LBG is between 26% and 50%.

[11] This section is heavily based on DG REGIO's EVALSED project. (http://ec.europa.eu/regional_policy/sources/docgener/evaluation/evalsed/sourcebooks/method_techniques/index_en.htm)

The study further distinguishes between four bandwidths around the threshold: Large ($0\% < x < 50\%$), Intermediate ($5\% < x < 45\%$), Small ($10\% < x < 40\%$), and Very Small ($15\% < x < 35\%$). The smaller the bandwidth, the more likely are the conditions of a quasi-experiment; but it should be remembered that the smaller the bandwidth, the smaller the external-validity of the results of the analysis will be.

Step 3: Check the type of discontinuity: sharp or fuzzy? Serrano-Valerede's work represents an intermediate case between sharp and fuzzy RDD designs. In this setting, ineligible firms have zero probability of receiving the subsidy, while some firms in the treatment group may not receive it (because they do not apply for it or because they do not get it). The group of eligible firms that are not treated are often referred to as "no shows" (Bloom, 1984). In other words, eligible firms have a positive assignment to treatment less than one, in the sense that only a small fraction of eligible firms takes up the subsidy. Battistin and Rettore (2008) show that the conditions required to achieve identification in this setting (often called Fuzzy type 1 RDD) are the same as in the sharp design. They show formally that, thanks to the discontinuity, eligible non-treated and ineligible firms are valid counterfactuals for supported firms, independently on how these supported firms self-select into the programme.

Step 4: Effect estimation. The author estimates a particular type of regression model, defined as a quantile regression. The purpose is to go beyond the effect of the subsidy on the R&D expenditure of the average firms, estimating the effect on firms that are located at the 25%, 50% and 75% percentile of the R&D expenditure. The following table summarises the results of the estimates by bandwidth and by quantile.

Table 2. Quantile regression estimates

	Quantiles			
Bandwidth	25 percentile	50 percentile	75 percentile	Observations
Large (0%<X<50%)	1.12 (0.29)	-0.13 (1.16)	1.13 (2.01)	560
Intermediate (5%<X<45%)	1.17 (0.23)	-1.15 (1.5)	2.55 (4.93)	380
Small (10%<X<40%)	1.10 (0.23)	-.84 (1.5)	1.78 (4.93)	276
Very small (15%<X<35%)	1.33 (0.4)	-2.7 (1.6)	-4.0 (1.21)	189

Source: Serrano-Valerede, 2008, p. 21.

As shown in Table 2, the author finds a statistically significant positive effect of the R&D subsidy on private R&D investment only for firms at the lowest quartile of the private R&D investment distribution (25 percentile).

Step 5. Discuss the external validity of the effect estimates. In presence of heterogeneous effects, the RDD only permits the identification of the mean impact at the threshold for the selection. In the realistic situation of heterogeneous effects across units, the estimated local effect might be very different from the effect for units away from the threshold for selection (Battistin and Rettore, 2008). Thus, the analyst should ascertain whether, and under which conditions, the results at the threshold can be extended to the whole population of interest.

6.2 Selected applications

Schwartz, J., (2013). Do temporary extensions to unemployment insurance benefits matter? The effects of the US standby extended benefit program. *Applied Economics*, 45(9): 1167-1183.

Abstract. During the 2007-2010 economic downturn, the **US** temporarily increased the duration of Unemployment Insurance (UI) by 73 weeks, higher than any prior extension, raising concerns about UI's disincentive effects on job search. This article examines the **effect of temporary benefit extensions** using a Regression Discontinuity (RD) approach that addresses the endogeneity of benefit extensions and labour market conditions. Using data from the 1991 recession, the results indicate that the Stand-by Extended Benefit (SEB) program has a significant, although somewhat limited, impact on county unemployment rates and the duration of unemployment. The results suggest that the temporary nature of SEB benefit extensions may mitigate their effect on search behaviour.

Marie, O., Vall Castello, J., (2012). Measuring the (income) effect of disability insurance generosity on labour market participation. *Journal of Public Economics*, 96 (1-2): 198-210.

Abstract. We analyze the **employment effect of a law that provides for a 36% increase in the generosity of disability insurance (DI)** for claimants who are, as a result of their lack of skills and of the labour market conditions they face, deemed unlikely to find a job. The selection process for treatment is therefore conditional on having a low probability of employment, making evaluation of its effect intrinsically difficult. We exploit the fact that the benefit increase is only available to individuals aged 55 or older, estimating its impact using a regression discontinuity approach. Our first results indicate a large drop in employment for disabled individuals who receive the increase in the benefit. Testing for the linearity of covariates around the eligibility age threshold reveals that the age at which individuals start claiming DI is not

continuous: the benefit increase appears to accelerate the entry rate of individuals aged 55 or over. We obtain new estimates excluding this group of claimants, and find that the policy decreases the employment probability by 8%. We conclude that the observed DI generosity elasticity of 0.22 on labour market participation is mostly due to income effects since benefit receipt is not work contingent in the system studied.

Bargain, O., Doorley, K., (2011). Caught in the trap? Welfare's disincentive and the labor supply of single men. *Journal of Public Economics*, 95(9-10): 1096-1110.

Abstract. Youth unemployment is particularly large in many industrialized countries and has dramatic consequences in both the short and long-term. While there is abundant evidence about the labor supply of married women and single mothers, little is known about how young (childless) singles react to financial incentives. The **French minimum income** (Revenu Minimum d'Insertion, RMI), often accused of generating strong disincentives to work, offers a natural setting to study this question since childless single individuals, primarily males, constitute the core group of recipients. Exploiting the fact that childless adults under age 25 are not eligible for this program, we conduct a regression discontinuity analysis using French Census data. We find that the RMI reduces the participation of uneducated single men by 7-10% at age 25. We conduct an extensive robustness check and discuss the implications of our results for youth unemployment and current policy developments.

Lalive, R., (2008). How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of Econometrics*, 142 (2): 785-806.

Abstract. This paper studies a targeted program that **extends the maximum duration of unemployment benefits from 30 weeks to 209 weeks in Austria**. Sharp discontinuities in treatment assignment at age 50 and at the border between eligible regions and control regions identify the effect of extended benefits on unemployment duration. Results indicate that the duration of job search is prolonged by at least 0.09 weeks per additional week of benefits among men, whereas unemployment duration increases by at least 0.32 weeks per additional week of benefits for women. This finding is consistent with a lower early retirement age applying to women.

Chen, S., van der Klaauw, W., (2008). The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics*, 142(2): 757-784.

Abstract. In this paper we evaluate the **work disincentive effects of** the **disability insurance (DI)** program during the 1990s using comparison group and regression-discontinuity methods. The latter approach exploits a particular feature of the DI eligibility determination process to

estimate the program's impact on labor supply for an important subset of DI applicants. Using merged survey-administrative data, we find that during the 1990s the labor force participation rate of DI beneficiaries would have been at most 20 percentage points higher had none received benefits. In addition, we find even smaller labor supply responses for the subset of 'marginal' applicants whose disability determination is based on vocational factors.

6.3 Synthesis: main hypotheses, data requirements, pros and cons

Regression discontinuity design			
Main hypotheses	**Data requirements**	**Pros**	**Cons**
Assignment is solely based on pre-intervention variables observable to the analyst and the probability of participation changes discontinuously as a function of these variables.	The dataset contains the selection variable and observations on eligible and non-eligible units.	In a neighbourhood of the cut-off for selection a RDD presents of a pure experiment.	Limited external validity of the estimates.
	(Traditionally) cross-sectional data.	This design allows one to identify the programme's causal effect without imposing arbitrary exclusion restrictions, assumptions on the selection process, functional forms, or distributional assumptions on errors.	To extend the results at the threshold to the whole population one can only resort to a non-experimental estimator whose consistency for the intended impact intrinsically depends on behavioural (and non-testable) assumptions.

7. Difference-in-Differences (DID)

The Difference-in-Differences (DID) method explores the time dimension of the data to define the counterfactual. It requires having data for both treated and control groups, before and after the treatment takes place. The ATT_{DID} is estimated by comparing the difference in outcomes between treated and control groups in some period after the participants have completed the programme with the difference that existed before the programme. It acknowledges the presence of unobserved heterogeneity in the selection into treatment, ensuring the estimation of the true ATT if this selection bias is constant over time as it is differenced out. Longitudinal data, in which the same individuals are followed over time, is usually used but it can also be applied to repeated cross-sectional data. Compared to cross-section estimators it has the advantage of controlling for differences in unobservable characteristics that are fixed over time, i.e. a specific form of selection on unobservables.

The treatment effect is obtained by taking two differences between group means. In a first step, the before-after mean difference is computed for each group: $E(Y_{ai}^T - Y_{bi}^T | D_i = 1)$ and $E\left(Y_{ai}^C - Y_{bi}^C | D_i = 0\right)$, where the subscripts a and b denote "after" and "before" the policy intervention, and the superscripts T and C indicate the treatment and the control group, respectively. The average treatment effect on the treated is the difference of these two differences:

$$[11] \qquad ATT_{DID} = E(Y_{ai}^T - Y_{bi}^T | D_i = 1) - E\left(Y_{ai}^C - Y_{bi}^C | D_i = 0\right),$$

This method relies on two main assumptions:

1. The unobserved heterogeneity is time invariant and is cancelled out by comparing the before and after situations;
2. The so called *common trend*: in the absence of the treatment, both treated and control groups would have experienced over time the same trend in the outcome variable. Therefore, any deviation from the trend observed in the treated group can be interpreted as the effect of the treatment.

An Example

For graphical illustration consider the example in the following figure, where the mean outcomes are on the vertical axis and time is on the horizontal one.

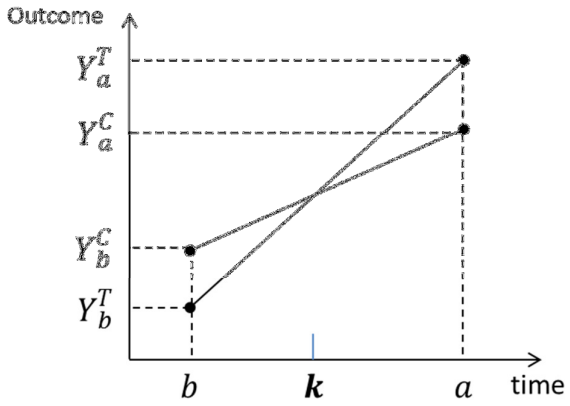


Figure 4 - An example of DID (part 1)

The period of the intervention is indicated as '*k*', and '*b*' and '*a*' are the before and after points in time for which data are available. The before and after mean outcomes for the treated units are Y_b^T and Y_a^T, respectively. The correspondent outcomes for the control group are Y_b^C and Y_a^C. The *ATT* effect is computed according to equation [11] above. In order to identify the *ATT* in the figure it is important to recall the *common trend* assumption, stating that the trend on the outcome would be the same for treated and control groups in the absence of the treatment (dashed dark line in figure 5 below). Accordingly, Y' in figure 5 is defined as the outcome of the treated individuals had they not received the treatment and the *common trend* assumption can be written as: $Y_a^C - Y_b^C = Y' - Y_b^T$. The *ATT* corresponds to the difference between the actual outcome of the treated group and the outcome they would have experienced had they not received the treatment: $Y_a^T - Y'$, i.e. the excess outcome change for the treated as compared to the non-treated.

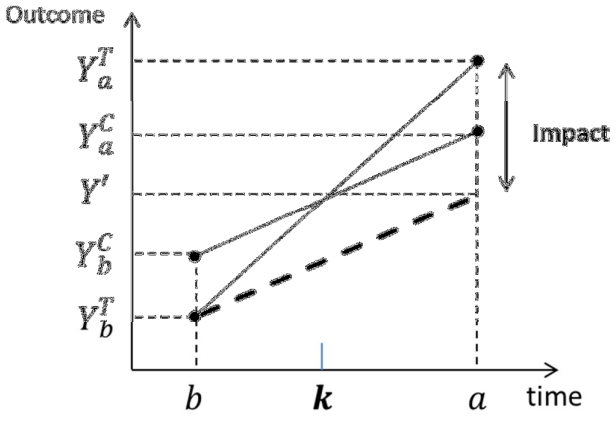


Figure 5 - An example of DID (part 2)

The remainder of the section is organised as follows: subsection 7.1 presents the steps for the implementation of this method; subsection 7.2 presents some recent applications; and, subsection 7.3 highlights the main hypotheses, data requirement and pros and cons of the DID approach.

7.1 Implementation steps

The following figure presents the estimation steps of the Difference-in-Differences method.

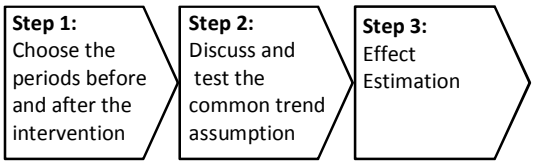


Figure 6 – DID implementation steps

Step 1: Choose the periods before and after the intervention

This choice is usually limited by data availability.

Step 2: Discuss and test the common trend assumption

If data are available for the outcome of both groups in pre-programme periods other than the one used to estimate ATT_{DID} (see identity [11]), the validity of the *common trend* assumption may be tested. When applied to pre-programme periods only, the ATT_{DID} should be zero, as no programme has yet been implemented.

If, due to some unobservable characteristics, the treated and control groups respond differently to a common shock (e.g. a macroeconomic shock), the common trend assumption is violated and DID would either under or overestimate the ATT. To overcome this problem, Bell, Blundell and van Reenen (1999) suggest a 'trend adjustment' technique: a triple difference method to account for these differential time-trend effects. In this case, data on both treated and control groups are needed not only in the before and after periods but also on other two periods, say t and t' ($t' < t < b < k < a$). The triple differences estimator, DDD (Differences-in-Differences-in-Differences), is:

$$[12] \qquad ATT_{DDD} = \left[E(Y_{ai}^T - Y_{bi}^T | D_i = 1) - E\left(Y_{ai}^C - Y_{bi}^C | D_i = 0\right)\right] - \\ -\left[E\left(Y_{ti}^T - Y_{t'i}^T \middle| D_i = 1\right) - E\left(Y_{ti}^C - Y_{t'i}^C | D_i = 0\right)\right]$$

where the last term is the trend differential between the two groups, measured between t and t'.

Step 3: Effect estimation

The ATT_{DID} is usually estimated within a regression framework. The key independent variables are: an indicator for members of the treated group, T_{it}; an indicator of the post-treatment period, t; the interaction of treatment group status and post-programme period, $T_{it}t$.

$$[13] \qquad Y_{it} = \alpha + \beta T_{it}t + \gamma T_{it} + \pi t + \varepsilon_{it},$$

The coefficient of the interaction term, β, is the ATT_{DID} as it measures the difference between the two groups in the post-programme relative to the pre-programme. This parametric approach is convenient for two reasons: i) for the estimation of standard errors; and ii) because it is a more flexible approach that allows including other explanatory variables, namely those that reflect differences between the groups' initial conditions and those that would lead to differential time trends.

Heckman, Ichimura and Todd (1997) suggested a combination between DID and matching methods to guarantee that only treated and control units comparable in the pre-treatment stage are taken into account in the estimation of the treatment effect. Smith and Todd (2005) show that this "conditional DID" is more robust than the traditional cross-section matching estimators, as it allows selection on observables as well as time-invariant selection on unobservables. To implement this combined approach one should match treated with control units based on pre-programme characteristics, and only units falling within the common support are used to compute the treatment effect (see the section dedicated to Matching).

This combination of methods allows implementing DID in repeated cross-sectional data. Since with this type of data the units are not followed through time, the treated units after implementation must be matched with three groups of units: i) participants in the pre-programme period; ii) control units in the pre-programme period; and iii) control units in the post-programme period.

7.2 Selected Applications

Bell, B., Blundell, R. and van Reenen, J., (1999). Getting the unemployed back to work: the role of wage subsidies. *International Tax and Public Finance*, vol. 6 (3): 339-360

Abstract. This paper examines alternative approaches to **wage subsidy programmes**. It does this in the context of a recent active labour market reform for the young unemployed in Britain. This "New Deal" reform and the characteristics of the target group are examined in detail. We discuss theoretical considerations, the existing empirical evidence and propose two strategies for evaluation. The first suggests an ex-post 'trend adjusted difference in

difference' estimator. The second, relates to a model based ex-ante evaluation. We present the conditions for each to provide a reliable evaluation and fit some of the crucial parameters using data from the British Labour Force Survey. We stress that the success of this type of labour market programmes hinge on dynamic aspects of the youth labour market, in particular the pay-off to experience and training.

Bergemann, A., Fitzenberger, B. and Speckesser, S., (2009). Evaluating the dynamic employment effects of training programs in East Germany using conditional difference-in-differences. *Journal of Applied Econometrics*, vol. 24 (5): 797-82

Abstract. This study analyzes the employment effects of **training** in East Germany. We propose and apply an extension of the widely used conditional difference-in-differences estimator. Focusing on transition rates between nonemployment and employment, we take into account that employment is a state- and duration-dependent process. Our results show that using transition rates is more informative than using unconditional employment rates as commonly done in the literature. Moreover, the results indicate that due to the labor market turbulence during the East German transformation process the focus on labor market dynamics is important. Training as a first participation in a program of Active Labor Market Policies shows zero to positive effects both on re-employment probabilities and on probabilities of remaining employed with notable variation over the different start dates of the program.

Blundell, R., Costa Dias, M., Meghir, C. and van Reenen, J., (2004). Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association,* vol. 2 (4): 596-606

Abstract. This paper exploits area-based piloting and age-related eligibility rules to identify treatment effects of a labor market program—the New Deal for Young People in the UK. A central focus is on substitution/displacement effects and on equilibrium wage effects. The program includes extensive **job assistance and wage subsidies to employers**. We find that the impact of the program significantly raised transitions to employment by about 5 percentage points. The impact is robust to a wide variety of nonexperimental estimators. However, we present some evidence that this effect may not be as large in the longer run.

Boeri, T. and Jimeno, J., (2005). The effects of employment protection: learning from variable enforcement. *European Economic Review,* vol. 49: 2057-2077

Abstract. **Employment protection legislatio**n (EPL) is not enforced uniformly across the board. There are a number of exemptions to the coverage of these provisions: firms below a given threshold scale and workers with temporary contracts are not subject to the most restrictive provisions. This within-country variation in enforcement allows us to make inferences on the impact of EPL which go beyond the usual cross-country approach. In this paper we develop a simple model which explains why these exemptions are in place to start

with. Then we empirically assess the effects of EPL on dismissal probabilities and on the equilibrium size distribution of firms. Our results are in line with the predictions of the theoretical model. Workers under permanent contracts in firms with less restrictive EPL are more likely to be dismissed. However, there is no effect of the exemption threshold on the growth of firms.

Boockmann, B., Zwick, T., Ammermuller, A. and Maier, M., (2007). *Do hiring subsidies reduce unemployment among the elderly? Evidence from two natural experiments*. Discussion Paper n. 07-001, ZEW – Centre for European Economic Research (ftp://ftp.zew.de/pub/zew-docs/dp/dp07001.pdf)

Abstract. We estimate the effects of **hiring subsidies** for older workers on transitions from unemployment to employment in Germany. Using a natural experiment, our first set of estimates is based on a legal change extending the group of eligible unemployed persons. A subsequent legal change in the opposite direction is used to validate these results. Our data cover the population of unemployed jobseekers in Germany and was specifically made available for our purposes from administrative data. Consistent support for an employment effect of hiring subsidies can only be found for women in East Germany. Concerning other population groups, firms´ hiring behavior is hardly influenced by the program and hiring subsidies mainly lead to deadweight effects.

Eichler, M. and Lechner, M., (2002). An Evaluation of public employment programmes in the East German state of Sachsen-Anhalt. *Labour Economics: An International Journal*, vol. 9: 143-186

Abstract. In East Germany, active labour market policies (ALMPs) have been used on a large scale to contain the widespread unemployment that emerged after unification. This paper evaluates the effects for participants in **public employment programmes** (PEPs), an important component of ALMP in the East German States (Länder). The paper focuses on individual unemployment probabilities. By concentrating on the state of Sachsen-Anhalt, the econometric analysis can use a large new panel data set available only for that state, the *Arbeitsmarktmonitor Sachsen-Anhalt*. We aim at nonparametric identification of the effects of PEPs by combining the use of comparison groups with differencing over time to correct for selection effects. Our results indicate that PEP participation reduces participants' probability of unemployment.

Forslund, A., Johansson, P. and Lindqvist, L., (2004). *Employment subsidies – A fast lane from unemployment to work?* IFAU Working Paper 2004, n. 18, Uppsala (http://www.ifau.se/upload/pdf/se/2004/wp04-18.pdf)

Abstract. The treatment effect of a Swedish **employment subsidy** is estimated using exact covariate-matching and instrumental variables methods. Our estimates suggest that the programme had a positive treatment effect for the participants. We also show how non-

parametric methods can be used to estimate the time profile of treatment effects as well as how to estimate the effect of entering the programme at different points in time in the unemployment spell. Our main results are derived using matching methods. However, as a sensitivity test, we apply instrumental variables difference-indifference methods. These estimates indicate that our matching results are robust.

7.3 Synthesis: main hypotheses, data requirements, pros and cons

Difference-in-Differences (DID)			
Main hypotheses	**Data requirements**	**Pros**	**Cons**
- The selection into treatment is based on time-invariant unobserved heterogeneity. - Common trend: in the absence of the treatment both groups would evolve similarly over time.	The method explores the time dimension of data, therefore requires either longitudinal or repeated cross-sectional data.	Allows for a specific form of unobserved heterogeneity: a time-invariant one.	The selection into treatment may be based on unobserved temporary individual characteristics which are not differenced out, leading to an error in the estimation of the ATT.
	The treatment must have occurred between two periods observed by the researcher.	Easy to test the common trend assumption if another pre-period data point is available (the DD estimator between these two pre-period time periods should be zero).	The common trend assumption might not be verified or might not be testable.
		Allows the combination with matching estimators to guarantee more comparability between treated and control groups.	More demanding on data availability: needs data from two data collection periods (longitudinal or repeated cross-section).
		It is a flexible approach allowing an illustrative interpretation.	

8. Instrumental Variables (IV)

The instrumental variable method deals directly with the selection on unobservables and is extensively discussed in Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996). The ATT_{IV} is identified if the researcher finds a variable, the *instrument*, which affects the selection into treatment but is not directly related with the outcome of interest or with the unobserved variables that determine it. The instrument is a source of exogenous variation that is used to approximate randomisation (Blundell and Costa Dias, 2008). The choice of the instrument is the most crucial step in the implementation of this method, and should be carefully motivated by economic intuition or theory.

The estimation of the ATT_{IV} is achieved through a linear regression model:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \varepsilon_i, \quad [14]$$

where the parameter of interest is β: the effect of the treatment on the outcome, keeping other pre-determined variables X_i constant.

In general, in non-experimental settings there is selection bias into treatment, namely selection on unobservable variables. Therefore, variables that affect simultaneously the outcome and the selection into treatment are often unobservable by the researcher (for instance innate ability or motivation). If the role of these unobservable variables is not taken into account, the *ATT* estimate will be wrong: it will either be over or underestimated. For example, consider a training programme targeted to unemployed individuals and suppose that motivation, a variable that cannot be observed by the researcher, affects both the probability of an individual applying to this programme and the future probability of finding a job (the outcome of interest). If this selection on unobservables is not dealt with, the *ATT* would be overestimated, as the role of motivation in finding a job would be attributed to the participation in the programme.

The *IV* approach aims at cleaning this selection on unobservables by using a so-called instrumental variable Z[12] that satisfies the following conditions:

1) It should affect the selection into treatment: $cov(Z, D) \neq 0$;
2) It does not make part of model [14]: $cov(Z, u) = 0$, also called *exclusion restriction*. This condition rules out any direct effect of the instrument on the outcome and any indirect effect through another variable other than D.

Following the example given above, the researcher must find an observed variable Z that: i) affects the decision to apply to the programme; ii) is not directly related to the probability of

[12] For simplicity the case with only one instrumental variable is presented, but there may exist more than one instrument.

finding a job in the future; and iii) is not related with the individual motivation. An example of such a variable might potentially be the distance the individual lives from the training centre, if observable. There is no obvious reason to believe that this distance is related to motivation or to probability of finding a job, but this belief should be discussed, explained and explored in the data in an exhaustive way.

The *IV* method identifies *ATT* using only the part of the variation in treatment status that is associated with Z. One of the most used implementation procedures is the two stage least squares (2SLS), in which the estimation of the parameter of interest occurs in two stages. First, the part of the treatment variable, D, that is independent of the unobserved characteristics affecting the outcome, ε_i, is isolated. This is done by estimating a first stage regression, in which the treatment variable is explained by the instrument, Z, and by the other exogenous covariates, X:

$$[15] \qquad D_i = \pi_0 + \pi_1 Z_i + \pi_2 X_i + \vartheta_i.$$

The fitted values of this regression $\widehat{D_i}$ reflect only the exogenous variation in the treatment. In the second stage, these fitted values are included in equation [14] as substitutes of D_i. The estimation of these two stage least squares (2SLS) should not be made manually by the researcher because, even though the coefficients would be correct, the standard errors would be wrong. This occurs because in the second stage it would not be taken into account that the $\widehat{D}$ were estimated. The estimation should be made using software package's specific commands (for instance the command 'ivregress' in STATA) to guarantee the correct estimation of standard errors.

It should be highlighted that the estimated *ATT* will depend on the particular instrumental variable used, as different instruments may induce different variation in the treatment variable.

Furthermore, if the impact of the treatment programme is the same for all individuals then the IV approach estimates the *ATT*. However, if the treatment has heterogeneous effects in the population, this method estimates a local treatment effect, the so called *Local Average Treatment Effect* (*LATE*). This estimator identifies the treatment effect only for the units that switch their treatment status from non-participation to participation due to the change in the instrument, but it does not identify the treatment effect for those who would always participate in the programme regardless of the instrument variation (Imbens and Angrist, 1994). While this fact is usually interpreted as a disadvantage of the *IV* method, the parameter estimated may be interesting for policy makers. For instance, when the instrument is a discrete variable, say a policy change, *LATE* will estimate the effect of treatment on individuals changing their treatment status in response to the policy change, which may give an important measure of the impact of the policy (Blundell and Costa Dias, 2008).

The remainder of the section is organised as follows: subsection 8.1 presents the steps for the implementation of this method; subsection 8.2 presents some recent applications; and subsection 8.3 highlights the main hypotheses, data requirement and pro and cons of the IV approach.

8.1 Implementation steps

Figure 7 presents the suggested steps for the estimation of the ATT using the instrumental variable method.

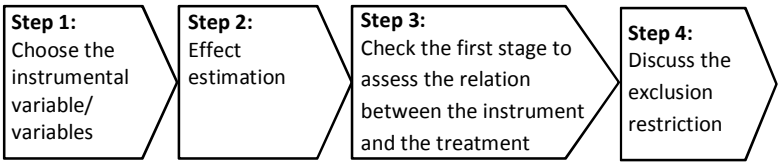


Figure 7 – IV implementation steps

Step1: Choice of the instrumental variables

The choice of the instrument is crucial to ensure the estimation of the causal treatment effect. In general, having detailed information on how the policy was targeted and implemented may reveal sources of exogenous variation that could be used as instrumental variables. Common sources of instruments include:

- *Policy geographical variation*
  For instance, if for exogenous reasons, the policy is implemented only in some regions and not in others, such that only part of the population is exposed to the policy
- *Exogenous shocks affecting the timing of policy implementation*
  For instance, if for exogenous reasons, the implementation of the policy was delayed in one region or for some group of the population.
- *Policy eligibility rules*
  If the policy is designed such that some units are eligible while others are not (parallel with the Regression Discontinuity Design method).

Step 2: Effect Estimation

Use specialised software (e.g. STATA or SAS) to estimate the two-stage least squares. The estimated parameter of interest is:

[16] $$ATT_{IV} = \frac{cov(Y,Z)}{cov(D,Z)} = ATT + \frac{cov(\varepsilon,Z)}{cov(D,Z)}.$$

This equation highlights the crucial role of the two assumptions presented above. If either is not met, the estimated ATT_{IV} will be different from the true one. Thus, it is important that the researcher discusses the reliability of these two assumptions.

Step 3: Check the first stage to assess the relation between the instrument and the treatment

The first assumption, stating that the instrument should be related with the treatment, is testable by analysing the first stage regression and assessing the strength of the relation between the treatment and the instrumental variables. If the instrument is weak in predicting the treatment, the estimation of the *ATT* is seriously affected.

The usual rule of thumb is that the F-statistic associated with the instrumental variable should be higher than 10 (Stock and Yogo, 2005). In case of weak instruments the bias of the estimated ATT could be even larger than the one obtained from not taking into account the selection on unobservables at all. Furthermore, even if $cov(\varepsilon_i, Z_i) = 0$ such that $ATT_{IV} = ATT$, the standard errors will increase because the treatment would be imprecisely predicted.

Step 4: Discuss the *exclusion restriction*

Even though the assumption that the instrument is not related with the error term is not directly testable, the researcher should discuss extensively why it is believed that the instrumental variable and the error term are not correlated, relying for instance on economic theory and intuition.

For instance, in the example above in which "distance" was suggested has a possible instrument, it could be argued that the individual could have moved to a particular part of the country where the training programmes are more usually offered. This would violate the assumption that motivation is not related with distance. Another example is when the training programmes are systematically given in more dynamic geographical areas from the economic point of view, for instance where it is more likely to find a job. If this is the case, there would exist a relation between distance and the outcome variable. These conjectures should be discussed and, if possible, proven to be right or wrong using the available data or alternative evidence.

8.2 Selected Applications

Abadie, A., Angrist, J. and Imbens, G., (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, vol. 70 (1): 91-117

Abstract: This paper reports estimates of the effects of **JTPA training programs** on the distribution of earnings. The estimation uses a new instrumental variable (IV) method that measures program impacts on quantiles. The quantile treatment effects (QTE) estimator reduces to quantile regression when selection for treatment is exogenously determined. QTE

can be computed as the solution to a convex linear programming problem, although this requires first-step estimation of a nuisance function. We develop distribution theory for the case where the first step is estimated nonparametrically. For women, the empirical results show that the JTPA program had the largest proportional impact at low quantiles. Perhaps surprisingly, however, JTPA training raised the quantiles of earnings for men only in the upper half of the trainee earnings distribution.

Carling, K. and Pastore, F. (1999). *Self-employment grants vs. subsidized employment: Is there a difference in the re-unemployment risk?* IFAU Working Paper 1999: 6, Uppsala (http://www.ifau.se/upload/pdf/se/to2000/wp99-6.pdf)

Abstract. **Self-employment grants and employment subsidies** are active labor market programs that aim at helping unemployed workers to escape unemployment by becoming self-employed or being hired at an initially reduced cost for the employer. In Sweden in the 1990's the participation rate in the self-employment program increased from virtually none to almost same as in the employment subsidy program. The advancement of the self-employment program is likely to be a result of (i) a change in the labor market program policy, and (ii) an increase in the supply of skilled unemployed workers. The justification for the policy change is unclear, however. The literature indicate that a rather specific group of unemployed workers may benefit from self-employment programs; Neither are there any strong reasons to believe in general that self-employment should be preferable to conventional employment through subsidies. We examine, *ex post*, the justification for the policy change by comparing the post-program duration of employment for the two programs. In addition, we focus in some detail on the outcome for female workers and workers of foreign citizenship. The reason for this is the explicit policy to direct those workers to self-employment. The data we study are the inflow to the two programs from June 1995 to December 1996. The program participants are followed to March 1999. The data contain detailed spell and background information on 9,043 unemployed workers who participated in the self-employment program and 14,142 who participated in the employment subsidy program. The second explanation, see (ii), for the increase in self-employment program implies a potentially serious selection problem. We discuss how the selection process may bias the effect estimate in the non-linear duration model that we use. Simulations help us to determine the magnitude of the selection bias in our application. Moreover, we exploit the existing behavioral heterogeneity across labor market offices to reduce the selection bias. We find that the risk of re-unemployment is more than twice as high for the subsidized employment program compared with the self-employment program. The large positive effect is, however, limited to male and female workers of Swedish origin. We thus conclude that the policy change in general has been successful, though we note that directing immigrant workers to self-employment is unlikely to improve the situation for this group of unfortunate workers on the Swedish labor market.

Forslund, A., Johansson, P. and Lindqvist, L., (2004). *Employment subsidies – A fast lane from unemployment to work?* IFAU Working Paper 2004, n. 18, Uppsala (http://www.ifau.se/upload/pdf/se/2004/wp04-18.pdf)

Abstract. The treatment effect of a **Swedish** employment subsidy is estimated using exact covariate-matching and instrumental variables methods. Our estimates suggest that the programme had a positive treatment effect for the participants. We also show how non-parametric methods can be used to estimate the time profile of treatment effects as well as how to estimate the effect of entering the programme at different points in time in the unemployment spell. Our main results are derived using matching methods. However, as a sensitivity test, we apply instrumental variables difference-indifference methods. These estimates indicate that our matching results are robust.

Frolich, M. and Lechner, M. (2004). *Regional treatment intensity as an instrument for the evaluation of labour market policies.* Discussion Paper n. 1095, Bonn, IZA (http://repec.iza.org/dp1095.pdf)

Abstract: The effects of active labour market policies (**ALMP**) on individual employment chances and earnings are evaluated by nonparametric instrumental variables based on Swiss administrative data with detailed regional information. Using an exogenous variation in the participation probabilities across fairly autonomous regional units (cantons) generated by the federal government, we identify the effects of ALMP by comparing individuals living in the same local labour market but in different cantons. Taking account of small sample problems occurring in IV estimation, our results suggest that ALMP increases individual employment probabilities by about 15% in the short term for a weighted subpopulation of compliers.

Stenberg, A., (2005). Comprehensive Education for the Unemployed — Evaluating the Effects on Unemployment of the Adult Education Initiative in Sweden. *Labour*, vol. 19 (1): 123-146

Abstract. This paper evaluates the effects on unemployment in Sweden of the **Adult Education Initiative** (AEI) which during its run from 1997 to 2002 offered adult education to the unemployed at compulsory or upper secondary level. The AEI is compared with the vocational part of Labor Market Training (LMT) using unemployment incidence and unemployment duration as outcome variables, both measured immediately after completion of the programs. For unemployment incidence, selection on unobservables is taken into account by using a bivariate probit model. The analysis of unemployment duration considers both selection bias and censored observations. The results indicate lower incidence following participation in the AEI, but also — significant at the 10 per cent level — longer duration.

Winter-Ebmer, R., (2006). Coping with a structural crisis: Evaluating an innovative redundancy-retraining project. *International Journal of Manpower*, vol. 27 (8): 700-721

Abstract. The purpose of this article is to evaluate a specific manpower **training program** in Austria; a program which was particularly designed for workers affected by a structural crisis in the steel industry. Microeconometric evaluation methods were used to assess earnings and employment probabilities up to five years after training. A treatment/control group approach was used together with instrumental variables estimates to control for selective entry into training. The results show considerable wage gains – even for a period of five years after leaving the Foundation – as well as improved employment prospects. The research has concentrated on a very specific project, which was exceptional in terms of broad training and counseling as well as in terms of funding and selection of trainees; therefore, it is not easily generalisable to other programs. The success of the program can be traced back to high incentives of all participants which, in turn, was caused by joint financing by local government, the workers themselves and the firm which made these workers redundant in the first place. Moreover, a combination of job counseling, search activities and training in capabilities which give presentable certificates turned out to be successful. The study will be valuable to those who look at specifics of job training programs as well as to those who are interested in designing programs for structural change.

8.3 Synthesis: main hypotheses, data requirements, pros and cons

Instrumental Variables (IV)			
Main hypotheses	**Data requirements**	**Pros**	**Cons**
The instrumental variable Z must satisfy two conditions: - Z affects the selection into treatment. - Z is not related with the unobserved variables and is not directly related with the outcome of interest.	Cross-sectional data, longitudinal or repeated cross-sectional data.	Deals directly with selection based on unobservables, either time invariant or not.	Very difficult to find an instrumental variable satisfying the two crucial assumptions.
		If a proper instrument is found, it guarantees that the estimated effect is causal.	In case of heterogeneous treatment effects, the IV method does not estimate ATT, but a local average treatment effect (LATE).
			The instrument may have insufficient variation in the selection into treatment → weak instrument.

9. Conclusions

This report presents the main methods used to evaluate the impacts of public policies using counterfactual methods. Our aim was not to discuss all the econometric details of these approaches – for which there are specialised books and papers – but rather to promote the culture of the evaluation of public policies within the European Commission showing the relevance and utility of the counterfactual framework. In order to put this framework into context, we have exemplified its applicability to Active Labour Market Policies. Yet it has to be highlighted that these tools are very flexible and can be used to evaluate the effect of a wide spectrum of public policies. For instance, a simple literature review conducted on Scopus using the search words "policy evaluation" and limiting the search to "Articles" or "Reviews" finds 2343 documents published between the 1974 and the 2013 in the area "Economics, Econometrics and Finance" (Figure 8). Limiting the search to the period between 2005 and 2013 we obtain 831 documents, most of them in heterogeneous fields like employment, education, health, development and environmental policies[13].

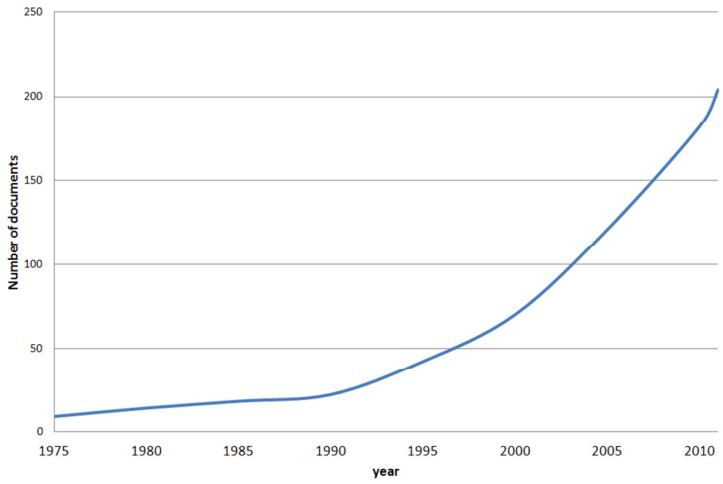


Figure 8 – Number of Articles or Reviews in the subject area "Economics, Econometrics and Finance" containing the words "policy evaluation" in their "Title, Abstract or Keywords" by year

Finally, it is worth stressing that the distinction between ex-post and ex-ante policy evaluation, as exists for instance in the European Commission, where different working groups are concerned with ex-ante (impact assessment) and ex-post (evaluation) analyses, should not be generalised to the methods applicable to these analyses.

The methods presented in this report are ex-post evaluation approaches because they evaluate the effects of a policy after its implementation. But, as discussed by Martini and Trivellato (2011), to be effective, any evaluation has to be designed before the implementation of the

[13] This literature review has been concluded in August 2012

policy to assess, i.e. ex-ante. Furthermore, the results of ex-post policy evaluations on a specific field are typically the inputs of studies aiming at assessing the impact of future policies on the same field. Last but not least, the application of the counterfactual approach to a unit of analysis, for example a territorial unit such as a region, can be used to justify the extension of the same policy, e.g. under logic of intervention, to similar units or set of units.

References

Abadie, A., Angrist, J. & Imbens, G.. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, vol. 70 (1), 91-117.

Angrist, J., Imbens, G. & Rubin, D., (1996). Identification of Causal Effects Using Instrumental Variables, Journal of the American Statistical Association, vol. 91: 444-472.

Angrist, J. & Pischke, J.S., (2009). *Mostly harmless econometrics. An empiricist's companion.* Princeton University Press.

Barnow, B.S., Cain, G.C. & Goldberger, A.S., (1980). Issues in the analysis of selectivity bias. In Stromsdorfer, E., Farkas (Eds.), *Education studies* (vol. 5, pp. 42-59). San Francisco: Sage.

Battistin, E. & Rettore, E., (2008). Ineligibles and eligible non-participants as a double comparison group in regression discontinuity designs. *Journal of Econometrics*, 142: 715-730.

Bell, B., Blundell, R. & Van Reenen, J., (1999). Getting the unemployed back to work: the role of wage subsidies. *International Tax and Public Finance.* Vol. 6 (3): 339-360.

Bloom, H., (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation review*, 8: 225-246.

Blundell, R. & Costa Dias, M., (2008). *Alternative approaches to evaluation in empirical microeconometrics*. Bonn, IZA: IZA DP No. 3800.

Blundell, R. & Costa Dias, M., (2002). Alternative Approaches to Evaluation in Empirical Microeconomic, *Portuguese Economic Journal*, vol. 1: 91-115.

Caliendo, M., Hujer, R. & Thomsen, S., (2005). *The employment effects of job creation schemes in Germany. A microeconomic evaluation*. Bonn, IZA: IZA DP No. 1512.

Caliendo, M. & Kopeinig, S., (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1): 31-72.

Card, D., Ibarrarán, P. & Villa, J.M., (2011). *Building in an evaluation component for active labor market programs: a practitioner's guide*. Bonn, IZA: IZA DP No. 6085.

Crépon, B., Gurgand, M., Rathelot, R. & Zamora, P., (2011). *Do Labor Market Policies have Displacement Effect? Evidence from a Clustered Randomized Experiment.*

Guo, S. & Fraser, M.W., (2010). *Propensity score analysis. Statistical methods and applications*. London: SAGE publications.

Heckman, J., Ichimura, H. & Todd, P., (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *The Review of Economics Studies*, vol. 64 (4): 605-654.

Hujer R. & Caliendo, M., (2000). *Evaluation of active labour market policy: methodological concepts and empirical estimates*. Bonn, IZA: IZA DP No. 236.

Imbens, G.W., (2004). Nonparametric estimation of average treatment effects under exogeneity: a Review. *Review of Economics and Statistics*, 86: 4-29.

Imbens, G. & Angrist, J., (1994). Identification and estimation of local average treatment effects. *Econometrica*, vol. 62 (2), 467-475.

Khandker, S., Koolwal, G. & Samad, H., (2010). *Handbook in Impact Evaluation: Quantitative Methods and Practices*. Washington D.C.: The World Bank.

Lechner, M., (1999). Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business and Economic Statistics*, 17: 74-90.

Lechner, M., (2001). *The Estimation of causal effects by differences-in-differences methods*. Discussion paper 2010-28, Department of Economics, University of St. Gallen

Lechner, M., (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *Journal of the Royal Statistical Society*, 165: 59-82.

Lemieux, T. & Milligan, K., (2008). Incentive effects of social assistance: a regression discontinuity approach. *Journal of Econometrics*, 142: 807-828.

Martini, A. & Trivellato, U., (2011). *Sono soldi ben spesi? Perché e come valutare l'efficacia delle politiche pubbliche*. Venezia: Marsilio Editore

Martini, A., (2008). How counterfactuals got lost on the way to Brussels. Prepared for the *Symposium "Policy and programme evaluation in Europe: cultures and prospects",* Strasbourg.

Rosenbaum P. & Rubin D., (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70: 41-50.

Rosenbaum, P.R. & Rubin, D.B., (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39: 33–38.

Rossi, P.H., Lipsey, M.W. & Freeman, H.E., (2004). *Evaluation. A systematic approach* (7th edition). Sage Publications, Inc.

Serrano-Valerede, N., (2008). *The Financing Structure of Corporate R&D - Evidence from Regression Discontinuity Design*. Working Paper, European University Institute, Department of Economics, Firenze.

Schochet, P., Burghardt, J. & McConnell, S., (2008). Does Job Corps work? Impact findings from the National Job Corps study. *American Economic Review*, 98(5): 1864-1886.

Sianesi, B., (2004). An Evaluation of the Swedish System of Active Labor Market Programs in the 1990s. *Review of Economics and Statistics*, 86(1): 133-155.

Shadish, W.R., Cook, T.D. & Campbell, D.T., (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Smith, J. & Todd, P., (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, vol. 125: 305-353.

Stock, J. & Yogo, M., (2005). Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometrics Models: Essays in Honor of Thomas Rothenberg*, ed. D. W. K. Andrews and J. H. Stock, 80-108. Cambridge: Cambridge University Press.

The authors

Massimo Loi is a research fellow at the Research Institute for the Evaluation of Public Policies (IRVAPP). He holds a Ph.D in Economics and Management from the University of Padua. Prior to this position, he was a post-doctoral fellow at the unit of Econometric and Applied Statistics of the European Commission (EU-JRC, unit G03) and at the department of economics of the University of Padua.

e-mail: maxloi@irvapp.it

Margarida Rodrigues is post-doctoral fellow at the unit of Econometric and Applied Statistics of the European Commission (EU-JRC, unit G03). She holds a Ph.D in Economics from the Nova School of Business and Economics, Universidade Nova of Lisbon.

e-mail: margarida.rodrigues@jrc.ec.europa.eu

European Commission
EUR 25519 EN – Joint Research Centre – Institute for the Protection and Security of the Citizen

Title: **A note on the impact of public policies: The counterfactual analysis**

Authors: Massimo Loi, Margarida Rodrigues

Luxembourg: Publications Office of the European Union

2012 – 54 pp. – 21.0 x 29.7 cm

EUR – Scientific and Technical Research series –ISSN 1831-9424 (online), ISSN 1018-5593 (print),

ISBN 978-92-79-26425-2

doi:10.2788/50327

Abstract

This report describes the policy evaluation framework and the different counterfactual analysis evaluation strategies: propensity score matching, regression discontinuity design, differences-in-differences and instrumental variables. For each method we present the main assumptions it relies on and the data requirements. These methodologies apply to any type of policy and, in general, to any type of intervention. A selection of papers applying this approach in the context of labour market interventions is also included.

LB-NA-25519-EN-N

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new standards, methods and tools, and sharing and transferring its know-how to the Member States and international community.

Key policy areas include: environment and climate change; energy and transport; agriculture and food security; health and consumer protection; information society and digital agenda; safety and security including nuclear; all supported through a cross-cutting and multi-disciplinary approach.

Publications Office

ISBN 978-92-79-26425-2

9 789279 264252