HAL
open science

Causal Inference and Impact Evaluation

Denis Fougère, Nicolas Jacquemet

To cite this version:

Denis Fougère, Nicolas Jacquemet. Causal Inference and Impact Evaluation. Economie et Statistique / Economics and Statistics, 2019, Special Issue 50th Anniversary, 510-511-512, pp.181-200. 10.24187/ecostat.2019.510t.1996 . hal-02866828

HAL Id: hal-02866828

https://hal.science/hal-02866828v1

Submitted on 12 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Causal Inference and Impact Evaluation

Denis Fougère* and Nicolas Jacquemet**

Abstract – This paper describes, in a non-technical way, the main impact evaluation methods, both experimental and quasi-experimental, and the statistical model underlying them. In the first part, we provide a brief survey of the papers making use of those methods that have been published by the journal *Economie et Statistique / Economics and Statistics* over the past fifteen years. In the second part, some of the most important methodological advances to have recently been put forward in this field of research are presented. To finish, we focus not only on the need to pay particular attention to the accuracy of the estimated effects, but also on the requirement to replicate evaluations, carried out by experimentation or quasi-experimentation, in order to distinguish false positives from proven effects.

JEL Classification: C1, C2, C3, C54
Keywords: causal inference, evaluation methods, causal effects

Reminder:

The opinions and analyses in this article are those of the author(s) and do not necessarily reflect their institution's or Insee's views.

** CNRS, Sciences Po Paris OSC/LIEPP, CEPR and IZA (denis.fougere@sciencespo.fr)*
*** University of Paris 1 Panthéon-Sorbonne (Sorbonne Economics Centre) and Paris School of Economics (nicolas.jacquemet@univ-paris1.fr)*

We would like to thank an anonymous reviewer for his/her comments on the first version of this article, that have contributed to significant improvements. This project is supported by the *Agence nationale de la recherche* (National Research Agency - ANR) and the French Government under the LIEPP Labex investment programme for the future (ANR-11-LABX-0091, ANR-11-IDEX-0005-02).

Translated from the original version: "Inférence causale et évaluation d'impact"

Citation: Fougère, D. & Jacquemet, N. (2019). Causal Inference and Impact Evaluation. *Economie et Statistique / Economics and Statistics*, 510-511-512, 181–200.
https://doi.org/10.24187/ecostat.2019.510t.1996

Over the past twenty years, the number of impact evaluation studies, whether experimental or quasi-experimental, has increased exponentially. These methods make it possible to identify, using individual survey data, relationships between variables that can be rigorously interpreted as cause-and-effect relationships. They are based on observation and research schemes that ensure that estimated differences in outcomes (e.g. in terms of earnings, employability, productivity or educational results) are mainly due to the intervention or policy implemented, and that selection and self-selection biases that tarnish many empirical studies are significantly reduced or even eliminated. In particular, these methods aim to statistically identify so-called "counterfactual" outcomes, i.e. those that would have occurred had the intervention in question not been implemented. The identification of the causal effect of the intervention on the outcome variable (its "impact") is then deduced by comparing the observed outcomes for the statistical units (unemployed people, employees, companies, students, etc.) benefiting from that policy.

A Short Review of the Standard Techniques

To achieve this goal, the simplest experimental method, which consists in randomly drawing units that benefit from the policy to be evaluated and comparing their post-intervention situation with that of the units (individuals or firms) that do not benefit from this policy, ensures that a causal relationship between the policy and the observed effect is demonstrated, without the analyst having to make overly restrictive assumptions. The other methods, known as quasi-experimental methods, seek to identify situations where, depending on a certain number of factors, the fact of benefiting from the intervention is independent of the characteristics, observable or not, of the units targeted by that intervention. These methods can be grouped into four categories, which are presented below in a non-technical manner.[1]

Instrumental Variables

Let us suppose that we observe the wages of two groups of workers, the first group having recently benefited from an active labour policy such as a training program, the other group having not benefited from it. Using the linear regression method, it is possible to estimate not only the effects of several variables characterizing the workers, such as age, gender, family situation, level of education, place of residence, etc., but also the effect of the participation in the training program on the post-program wage, i.e., the wage received at the time of the survey. However, this simple method may produce biased estimates. The problem is that participation in the training program is not exogenous: it can not only be correlated with the observed characteristics that we have just mentioned, but also with variables not observed by the analyst, such as a desire to change profession, a desire to learn new skills, the employee's productivity as assessed by his/her employer, etc. Consequently, the fact of having participated in the training program is likely to be correlated with the error term of the regression, the value of that error term generally being dependent on these unobserved characteristics. This correlation is the cause of the so-called "endogeneity" bias. To deal with this problem, econometricians have used the instrumental variable method for a long time. By definition, an instrumental variable must have a very significant impact on access to the program being evaluated – in this case, the training program – without directly affecting the wage level received after participating in that program. The estimation method used in this case is the so-called "two-stage-least-squares" technique. The first step consists in regressing participation in the training program on all exogenous variables (age, gender, etc.) but also on the value of the instrumental variable (which can be, for example, the date of a significant amendment made to the conditions governing access to this program). In a second step, individual wages must be regressed on the same exogenous variables and on participation in training program, not as actually observed, but as predicted by the first regression. The coefficient associated with this "instrumented" value can be interpreted, under certain very restrictive conditions, as "the causal effect" of the training program on trainees' wages.

Matching Methods

The main purpose here is to compare beneficiaries and non-beneficiaries by neutralising the differences due to the distribution of observable characteristics. These methods are based on two assumptions. The first stipulates that the

1. These methods are described in detail, for example, in Crépon & Jacquemet (2018), Chapter 9.

assignment to the group of beneficiaries depends exclusively on observable exogenous characteristics and not on the anticipated outcomes of the intervention: this assumption is called the "conditional independence assumption". The second assumption is that any individual or firm has a non-zero probability (comprised strictly between 0 and 1) of being *a priori* a beneficiary of the intervention, whatever the characteristics of that individual or firm, or whether or not that the individual or the firm is actually (i.e. *a posteriori)* a beneficiary of the intervention: this assumption is called the "overlap assumption". When these two assumptions are valid, the method consists in comparing the outcome for each beneficiary with the average of the outcomes for the non-beneficiaries who are "close" in terms of the observable characteristics (age, gender, level of education, etc.), and then averaging all these differences among the group of beneficiaries. Proximity to the beneficiary under consideration, i.e. the choice of his/her "closest neighbours", can be made using a distance (such as the Euclidean distance or the Mahalanobis distance), or even more simply using a propensity score, defined as the probability of being a beneficiary of the intervention given the observable variables characterising the individual; this probability can be estimated in a first step, using for example a logit or a probit model, independently of the value of the observed outcome variables.

Difference-in-Differences Methods

These methods are based on a simple assumption. Suppose that we observe the variations between two dates of an outcome variable such as the wage within two distinct groups. The first of these groups, called the "target group", "treated group" or "treatment group", benefits from a given intervention or an employment policy; the second, called the "control group",[2] does not. The employment policy is implemented between the two dates under consideration. The method relies on the following assumption: in the absence of this policy, the average wage change for individuals in the treated group would have been identical to that observed in the control group (the "parallel trends" assumption). The validity of this assumption, which cannot be verified, can be confirmed by the fact that, before the policy was implemented, wages evolved in the same way in both groups (that is the so-called "common pre-trends" assumption). Unlike the previous assumption, this second one can be tested on the basis of data observed prior to the implementation of the intervention, provided that repeated observations are available during this period. This method thus exploits the longitudinal (or pseudo-longitudinal[3]) dimension of the data.

The Regression Discontinuity Method

This method can be applied when the access to an intervention or a public policy is dependent on an exogenous threshold set by the authorities in charge of that policy. This threshold may be an age condition (for retirement, for example), an employment level threshold (for example, a tax reduction policy for firms with less than 20 employees), or a level of resources giving access to a scholarship or a tax credit. In its simplest form, regression discontinuity makes it possible to compare the average value of the outcome variable in the group of beneficiaries, for example those whose income or age is just below the eligibility threshold, with the average value of this variable in the comparable control group, composed of those whose income or age is just above that threshold. The underlying assumption is that, for people who otherwise have the same characteristics in terms of employment skills, level of education or gender, those just below and above the threshold are identical. Only sheer chance, for instance a date of birth, distinguishes them. Under these conditions, a simple difference between the means of the outcome variable (for example, the level of wage or education after the policy is implemented) makes it possible to estimate the causal effect of the intervention in question. However, this difference is only a local measure, close to the threshold, and its extrapolation to income levels or ages far from that threshold has no scientific validity. For this reason, it is said that regression discontinuity makes it possible to estimate a local average treatment effect (discussed in detail below).

Each type of method therefore corresponds to very specific assumptions. In practice, particularly when it is not possible to conduct a randomized experiment, it is important to recognise the information available to the analyst and to know which of these assumptions are most likely in order to choose the method which is best suited to the data available. Since the pioneering

2. These expressions are the same in each of the causal inference methods used.

3. The repeated observations may not be those concerning the same individuals but may be repetitions of random samples taken from the same population and form a "pseudo panel".

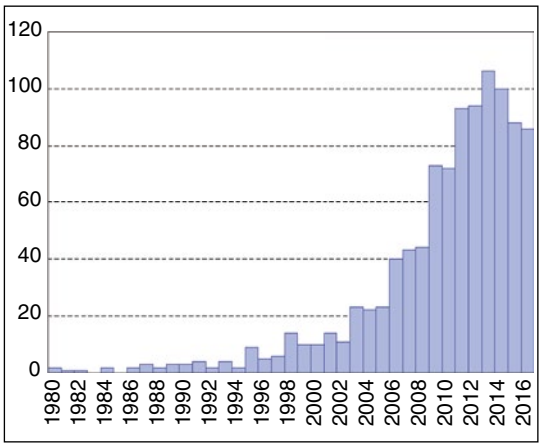
article published by LaLonde in 1986, several studies have been devoted to the comparison of evaluations carried out using experimental and quasi-experimental methods, and in particular to the estimation biases that may result from using quasi-experimental methods. Due to space constraints, it is not possible to summarize the results of those comparisons here. On this topic, the reader may consult, for example, papers written by Glazerman *et al*. (2003), Hill (2008), Chabé-Ferret (2015), Wong *et al.* (2017), and Chaplin *et al.* (2018).

A Flourishing International Scientific Literature

These methods have been applied in many research fields. For example, in the field of educational policy, the number of randomized controlled trials (RCTs) that have resulted in international publications has increased from just a few in 1980 to more than 80 per year since 2010 (Figure I). Quasi-experimental evaluations have followed a similar trend and nowadays, constitute together what some have called an "empirical revolution".[4] These studies and the quantitative assessments that they contain are resources of prime importance when it comes to choosing, designing and implementing public policies.

The recent publication of several reference articles and books also shows just how developed and diverse econometric evaluation methods have become. These include the books by Imbens & Rubin (2015), Lee (2016), and Frölich & Sperlich (2019), which follow on from the survey papers by Angrist & Krueger (1999), Heckman *et al.* (1999), Heckman & Vytlacil (2007a, 2007b), Abbring & Heckman (2007), and Imbens & Wooldrige (2009). The *Handbook of Field Experiments* published by Duflo & Banerjee in 2017 is the reference book on randomised field experiments. For laboratory experiments, Jacquemet & L'Haridon's book (2018) is the most recent reference. Finally, the list of papers on causal inference methods published in the best international economic or statistical journals over the past 30 years is too long to be included here. The interested reader will find it in the bibliographies of the above-mentioned works. Summaries in French (more or less formalised) are also available. These include papers by Brodaty *et al.* (2007), Givord (2014) and Chabé-Ferret *et al.* (2017).

Figure I
Number of randomised controlled trials conducted between 1980 and 2016 in the field of educational policy published in an international scientific journal, from Connolly *et al.* (2018)



Many Evaluations Studies Were Published in *Économie et Statistique*

The journal *Économie et Statistique* (not "/ *Economics and Statistics*" at the time) has accompanied this progress and these developments over the past twenty years, frequently publishing papers applying econometric evaluation methods to French data, mainly produced by public statistics departments and agencies. Some of these papers have found a real resonance in the public debate. It is admittedly risky to draw up an exhaustive list of them, since some of these publications may have escaped our attention. However, some of them may be cited by grouping them according to the methods used.

The instrumental variable technique was used by Crépon *et al.* (2004) to measure the effects of reduced working time on firms' productivity and employment. Leclair & Roux (2007) then used it to measure relative productivity and the use of short-term jobs in firms. Instrumental variables were also used by Beffy *et al.* (2009) to estimate the effects of students' paid work on their success in higher education, and by Fougère & Poulhès (2014) to study the influence of ownership on the household financial portfolio.

The reader will find applications of the difference-in-differences method in several papers published

4. *Angrist & Pischke (2010).*

in the journal. The first publications to use this method are the papers by Bénabou *et al.* (2004), devoted to the evaluation of priority education zones, and Behaghel *et al.* (2004), who sought to estimate the effects of the Delalande tax on employees' transitions between employment and unemployment. Fack & Landais (2009) used it to assess the effectiveness of tax incentives for donations. Carbonnier (2009) assessed the incentive-based and redistributive consequences of tax incentives for the employment of a home-based employee. The method made it possible for Bozio (2011) to measure the impact of the increase in insurance duration following the 1993 pension reform. Geniaux & Napoleone (2011) used a difference-in-differences method coupled with a matching method to assess the effects of environmental zoning on urban growth and agricultural activity. Again using the difference-in-differences method, Simonnet & Danzin (2014) assessed the effect of income support on the return to work of recipients, and Bérard & Trannoy (2018) measured the impact of the 2014 increase in real estate transfer taxes on the French housing market.

The papers that have applied matching methods include, in particular, those written by Crépon & Desplatz (2001) who used such a method to estimate the effects of payroll tax relief on low-wage workers' employment, by Even & Klein (2007) who estimated the medium-term effects of subsidized jobs on the employment of beneficiaries, by Rathelot & Sillard (2008) who assessed the effects of the urban tax-free zone policy on paid employment and the setting-up of new undertakings, and by Bunel *et al*. (2009) who focused their study on the effects of social security contribution reliefs on employment and wages.

The regression discontinuity method first appeared in *Économie et Statistique* by Lorenceau (2009), who estimates the effects of lower payroll charges granted in rural regeneration areas on the setting-up of new undertakings and employment level. It was also used by Baraton *et al.* (2011) to assess the effects of the 2003 reform on the retirement age of secondary-school teachers.

To our knowledge, *Economie et Statistique / Economics and Statistics* has, strictly speaking, not yet published any papers on randomized trials. This does not mean that French economists have not written high-quality research papers in this field. On the contrary, under the influence and sometimes with the collaboration of Esther Duflo, Professor of Economics at the MIT, French economists have published papers on randomized trials in the best international journals, particularly in the field of employment or education policies. The reader will find notable examples of such papers in the works of Crépon *et al.* (2013, 2015), Avvisati *et al.* (2014), Goux *et al.* (2017), or Barone *et al.* (2019). However, *Économie et Statistique* has published three papers on audit experiments, which, while being random experiments, cannot be considered as randomized field experiments. An audit study is a form of social experimentation in a real situation, mainly designed to detect a situation of discrimination. In the simplest case, the statistician compares the behaviour of a third party, usually an employer or a landlord, towards two people with exactly the same profile concerning all the relevant characteristics, except for the one suspected of giving rise to discrimination, for instance ethnic origin, disability, religion, age, gender, sexual orientation, etc. The paper by Petit *et al.* (2011) on the effects of an individual's place of residence on his/her access to employment, as well as those by Petit *et al.* (2013) and Edo & Jacquemet (2013) on the effects of gender and origin on discrimination in the workplace, are particularly representative of this type of approach, the limitations of which, both methodological and conceptual, were mentioned by Aeberhardt *et al.* (2011) in a comment published in the journal following the paper written by Petit *et al.* (2011).

The list of publications, particularly international publications, using statistical methods of causal inference is growing day by day. In addition to studies directly applying them with experimental or quasi-experimental data, much work has been devoted in the last ten years to refining these methods, or to coming up with solutions to overcome some of their limitations. The rest of this paper is devoted to presenting the developments that we believe are particularly promising in this area. Due to space constraints, we have not been able to address all the emerging themes here, including, in particular, social interactions and interference in randomised trials. This subject, which has unfortunately been relatively neglected to date, is addressed, for example, in the papers written by Hudgens & Halloran (2008), Aronow (2012), Manski (2013), Liu & Hudgens (2014), and Baird *et al.* (2018). An extensive review of recent developments and future research directions can be found in the papers written by Athey & Imbens (2017a, 2017b) and Abadie & Cattaneo (2018).

The Canonical Impact Evaluation Model

From its original formulation by Rubin (1974), the canonical impact evaluation model emphasises the heterogeneity of the response of economic agents following an intervention concerning them[5]. In this model, each observation unit is characterised by two "potential outcomes" specific thereto: y_{i0} is the outcome that would be observed for the unit i in the absence of the intervention, and y_{i1} is the outcome that would be observed as a result of the intervention. For each unit, only one of these two effects is observed. Rather than a "causal effect", the intervention is therefore associated with a distribution of situational changes $\Delta_i = y_{i1} - y_{i0}$, $i = 1,\ldots,N$, N here being the sample size. The evaluation process therefore requires choosing the parameter of this distribution that the analyst seeks to identify. Among the parameters summarising the distribution of the effect of the intervention (or treatment), the most common are the average treatment effect and the average treatment effect on the treated.

The average treatment effect (ATE) corresponds to the mathematical expectation of this distribution: it therefore measures the average change in outcome for an individual randomly selected from the population. The average treatment effect on the treated (ATT), for its part, is specific to the sub-population of individuals who actually benefit from the program (and formally corresponds to the conditional expectation to be actually treated). The two parameters are only equal under very restrictive assumptions. For example, they match each other trivially if the intervention concerns the whole population (for instance, an increase in the minimum age for leaving the school system, a measure that concerns all pupils), or if the treatment is supposed to act in the same way on all the individuals ($\Delta_i = \Delta,\ i = 1,\ldots,N$). In all other circumstances, these two parameters are distinct. They provide different information on the distribution of the causal effect: the average treatment effect on the treated measures the effectiveness of the program through the change in the beneficiaries' outcome, while the average treatment effect indicates how effective it would be if the program were to be applied to the entire population. The evaluation method chosen strongly influences the parameter that can be measured. Randomized experiments make it possible to estimate the ATE provided that the random assignment to experimental groups is made in the entire population and that all individuals selected to take part in the experiment actually do so. However, they can be used to estimate the ATT only when some of the selected individuals refuse to take part in the experiment or, more generally, when only a non-random sub-sample of the collected sample is observed (see Chabé-Ferret *et al.*, 2017, for an illustration). The difference-in-differences estimator or the matching estimators, for their part, measure the change in the situation specific to the beneficiaries, i.e. the ATT.

Beyond the importance of the choice of the parameter to be estimated (which must take precedence over the choice of the identification method), the heterogeneity of the treatment effect constitutes a significant limitation to the ability to generalise the estimated effects of an intervention in the context of a particular empirical study (see below).

The Local Average Treatment Effect (LATE)

Since the work of Imbens & Angrist (1994), who introduced the local average treatment effect (LATE) estimator, the interpretation of the instrumental variable estimator as the "average treatment effect on the treated" has been called into question. It is only valid if the effect of the program is the same for all individuals, regardless of their age, gender, experience, etc., which is obviously a very unrealistic assumption. Imbens & Angrist (1994), and many econometricians following them, show that if the effect of an intervention or public policy is likely to vary from one group of individuals to another, and more generally to be heterogeneous within a given population, only a local estimator can be produced for those individuals who decide to benefit from the program when it becomes available as a result of a variation of the instrument. Those individuals are called "compliers", i.e. people who comply or adhere to the programme when the value of the instrument changes. The group of compliers is probably best defined when confronted with people who systematically refuse the program ("never-takers") and those who are always willing to take participate in it ("always-takers"), regardless of the value of the instrument. The implementation

5. This model is different from the model introduced by Judea Pearl, which uses the formalism of directed acyclic graphs, which are often used in epidemiology or psychometry (see Peters et al., 2017, or Pearl & Mackenzie, 2018).

of the LATE estimator assumes that there are no individuals who would be willing to take part in the program when it is not offered, but who would refuse to do so once the program is rolled out. This group of people, who are called "defiers", is assumed not to exist: this assumption corresponds to what Imbens & Angrist (1994) call the "monotonicity assumption". The LATE estimator therefore measures the effect of the intervention only on the group of compliers, which unfortunately cannot always be identified. When it is, for instance when a lottery or a random procedure changes the assignment to the treatment (i.e., the proposed intervention or program), the LATE estimator can be obtained using the two-stage least squares procedure. Angrist & Imbens (1995) propose a more general method that takes into account the effect of other exogenous variables (such as age) in the implementation of the LATE. Angrist *et al.* (2000) apply this approach to the estimation of simultaneous equation models.

The External Validity of Impact Evaluation Methods

Several of the methods cited above are characterised by strong internal validity: they provide credible estimators of the average effects of interventions for the samples under consideration. However, the possibility of extrapolating their outcomes to a larger population, i.e., their external validity, is often called into question.

In the case of randomized trials, this criticism is based on the fact that the samples are generally quite small and concern particular groups, for example people living in some given environments or with specific characteristics; they are not representative of the population as a whole, or at the very least of all the potentially eligible people. The issue of external validity is fundamentally linked to the heterogeneity of the effects of interventions (see below). Suppose that a trial is conducted in a setting A, which may correspond to a given location, period, or sub-population of individuals. How do the estimates of the effects of this particular intervention conducted in this particular setting inform us of what the effects of the same intervention would be in another location, in a different period, for a different group of individuals, i.e., in a setting B that is different from setting A? The differences may result from observed and unobserved characteristics of those other locations, periods or individuals, and possibly from changes (no matter how slight they are) in the intervention procedures. To answer these questions, it is useful to have access to the results of multiple trials, carried out in different settings, and if possible, with fairly large samples representative of the eligible population (at least in terms of the main observable characteristics). Microfinance represents a particularly interesting example. For instance, Meager (2019) analyzed the results of seven trials conducted on this topic, and found that the estimated effects were remarkably consistent.

Another approach is to explicitly take account of the differences between the distributions of the characteristics specific to the groups or periods in question. Hotz *et al*. (2005) and Imbens (2010) propose a theoretical setting in which the differences in effects observed within a group of several locations stem from the fact that the units established in these locations have different characteristics. By means of an adjustment procedure that consists in reweighting individual units (persons, households, firms, etc.), they can compare the effects of the intervention in question in these different locations. This technique is close to the inverse probability weighting methods[6] recommended by Stuart and co-authors (Imai *et al.*, 2008; Stuart *et al.*, 2011; Stuart *et al.*, 2015).

It should be recalled that the instrumental variable estimator is often interpreted as a local estimator of the average treatment effect, i.e., as a LATE estimator that measures the average treatment effect for those members of the population (the compliers) whose assignment to the treatment is modified by a change in the value of the instrument. Under what conditions can this estimator be interpreted as the average treatment effect for the entire population? In other words, what are the conditions that ensure its external validity? Two groups are never affected by the instrumental variable: the always-takers who always receive the treatment, and the never-takers who never receive it. To answer the question, Angrist (2004) suggests testing whether the difference between the average outcomes of the always-takers and the never-takers is equal to the average treatment effect on the outcome of the compliers. Angrist & Fernandez-Val (2013) seek to exploit a conditional effect ignorability assumption stipulating that, conditional on certain exogenous variables, the average effect

6. Inverse probability weighting is a statistical technique for calculating standardized statistics for a pseudo-population that is different from the one from which the data were collected.

for compliers is identical to the average effect for always-takers and never-takers. Bertanha & Imbens (2019) suggest testing the combination of two equalities, namely the equality of the average outcomes of untreated compliers and never-takers, and the equality of the average outcomes of treated compliers and always-takers.

In the case of regression discontinuity, the lack of external validity is mainly due to the fact that this method produces local estimators, which are only valid around the considered eligibility threshold. If, for example, that threshold is an age condition, regression discontinuity does not make it possible to infer what the average effect of the intervention would be for people whose age differs significantly from the age defining the eligibility threshold. Under what conditions can the estimated effects obtained through regression discontinuity be generalized? Dong & Lewbel (2015) note that in many cases, the variable that defines the eligibility threshold (called the "forcing variable") is a continuous variable such as age or income level. These authors point out that in this case, beyond the extent of the discontinuity of the outcome variable in the vicinity of the threshold, it is also possible to estimate the variation of the first derivative of the regression function, and even of higher-order derivatives. This makes it possible to extrapolate the causal effects of the treatment to values of the forcing variable further away from the eligibility threshold. Angrist & Rokkanen (2015) propose to test whether, conditional on additional exogenous variables, the correlation between the forcing variable and the outcome variable disappears. Such a result would mean that the allocation to treatment could be considered independent of the potential outcomes (this is called the unconfoundedness property)[7] conditional on those additional exogenous variables, which would again allow the result to be extrapolated to values of the forcing variable further from the threshold. Bertanha & Imbens (2019) propose an approach based on the fuzzy regression discontinuity design.[8] They suggest testing the continuity of the conditional expectation of the outcome variable, for a given value of the treatment and of the forcing variable at the threshold level, adjusted by variations in exogenous characteristics.

Difference-In-Differences and Synthetic Control

As noted above, the implementation of the difference-in-differences method requires there to be a control group whose evolution over time reflects what the treatment group would have experienced in the absence of any intervention. This assumption cannot be tested over the period following the intervention, during which differences in outcome between groups also reflect the effect of the policy. A testable component of this assumption is that the past evolution of the outcome variable (before the policy being evaluated is implemented) is on average similar to that of the same variable in the treatment group. When it is rejected, it is possible to create an artificial control ("synthetic control") unit, based on the observations of the control group, using an appropriate weighting system. This synthetic control is constructed in such a way that the past evolution of the outcome variable within it is identical to that of this variable in the treatment group.

The method was introduced by Abadie & Gardeazabal (2003) in a study aimed at assessing the effect of ETA terrorist activity on the development of the Basque Country's GDP between 1975 and 2000, a period when the Basque separatist terrorist organisation was most active, frequently committing extreme acts of violence. The problem is that between 1960 and 1969, the decade preceding the beginning of the period of terrorist activity, the Basque Region's GDP evolved very differently from the average GDP of the other sixteen Spanish regions, leading to the assumption of a common pre-treatment trend being rejected. Abadie & Gardeazabal (2003) then proposed to construct a synthetic control region whose GDP evolution between 1960 and 1969 would be similar to that of the Basque Country's GDP. This can be achieved by minimizing the distance between the annual observations of the Basque Country's GDP between 1960 and 1969 and those of this synthetic region. More formally, the annual GDP values in the Basque Country between 1960 and 1969 are denoted $y_{1,t}$ (t = 1960,…,1969) and grouped together in a vector $\boldsymbol{Y}_{1,0} = [Y_{1,1960} \ldots Y_{1,1969}]$. Similarly, the annual observations concerning the GDP of each of the other sixteen Spanish regions are denoted $Y_{j,t}\left(j = 2,\ldots,17; t = 1960,\ldots,1969\right)$ and stored in a matrix denoted $\boldsymbol{Y}_{0,0}$ of dimension $\left(10 \times 16\right)$. The synthetic control region is constructed from a

7. *"The unconfoundedness assumption states that assignment is free from dependence on the potential outcomes" (Imbens & Rubin, 2015, p. 257).*
8. *The sharp regression discontinuity design corresponds to the case where nobody can derogate from the constraint of the eligibility threshold. This case is opposite to that of the fuzzy regression discontinuity design, in which treated individuals, or untreated individuals, are observed on both sides of the threshold.*

weighting vector $\boldsymbol{w} = [w_1, \ldots, w_{16}]'$ of dimension (16×1) which minimizes the following weighted Euclidean norm for a given matrix $\boldsymbol{V}$:

$$\left\| \boldsymbol{Y}_{1,0} - \boldsymbol{Y}_{0,0}\boldsymbol{w} \right\| = \sqrt{\left(\boldsymbol{Y}_{1,0} - \boldsymbol{Y}_{0,0}\boldsymbol{w}\right)' \boldsymbol{V} \left(\boldsymbol{Y}_{1,0} - \boldsymbol{Y}_{0,0}\boldsymbol{w}\right)}$$

In a first simple application, Abadie & Gardeazabal (2003) choose the identity matrix as the matrix $\boldsymbol{V}$. This allows them to easily find the weighting system $\boldsymbol{w}^*$ that minimizes this norm.[9] They verify that the ten annual GDPs of that synthetic region, which are calculated as $\boldsymbol{Y}_{0,0}^* = \boldsymbol{Y}_{0,0} \times \boldsymbol{w}^*$ during the 1960-1969 period, are similar to the yearly GDPs of the Basque region observed during the same period. This allows them to then calculate the counterfactual GDPs of the Basque region during the period of terrorist activity (1975-2000). These counterfactual GDPs are denoted $\boldsymbol{Y}_{0,1}^*$ and are calculated in the dimension vector (26×1) $\boldsymbol{Y}_{0,1}^* = \boldsymbol{Y}_{0,1} \times \boldsymbol{w}^*$, where $\boldsymbol{Y}_{0,1}$ is the dimension matrix (26×16) which groups together the observations concerning the 26 annual GDPs[10] of each of the sixteen Spanish regions other than the Basque Country. The causal effect of terrorism on the Basque GDP is then measured as $\boldsymbol{Y}_{1,1} - \boldsymbol{Y}_{0,1}^*$ where $\boldsymbol{Y}_{1,1}$ is the dimension matrix (26×1) which groups together the 26 annual observations of the Basque GDP from 1975 to 2000.

In general, $\boldsymbol{V}$ is a diagonal matrix with non-negative diagonal elements. In an extended version of this method, Abadie & Gardeazabal (2003) and Abadie *et al.* (2010, 2015) propose to choose matrices $\boldsymbol{V}$ whose elements are data driven. The number of units treated may be greater than one: in this case, a synthetic control must be calculated for each unit treated. However, when the number of units treated is very large, the synthetic control of a treated unit may not be unique. Abadie & L'Hour (2019) propose a variant that takes this difficulty into account. Their estimator is written:

$$\left\| \boldsymbol{Y}_{1,0} - \boldsymbol{Y}_{0,0}\boldsymbol{w} \right\|^2 + \lambda \sum_{j=2}^{J+1} w_j \left\| \boldsymbol{Y}_{j,0} - \boldsymbol{Y}_{1,0} \right\|^2, \text{ with } \lambda > 0$$

In this expression, $\boldsymbol{Y}_{j,0}$ is the vector whose elements are the observed values of the outcome variable for the control unit j $(j = 2, \ldots, J+1)$ during each of the periods preceding the implementation of the intervention. The estimator proposed by Abadie & L'Hour (2019) includes a penalty λ for differences between the values of the outcome variable of a treated unit and those of each control unit in the period before the intervention was implemented. Abadie & L'Hour (2019) show that, under these conditions, and except in a few specific cases, their estimator provides a single synthetic control.

Extended versions of the synthetic control estimator have also been proposed by Amjad *et al.* (2018) and Athey *et al.* (2018), who suggest the use of matrix completion techniques, but also by Hahn & Shi (2017), who base their approach on sampling-based inferential methods.

The Role and Choice of Explanatory Variables

Regardless of the type of intervention or evaluation method chosen by the researcher, the individuals, households, firms, etc. sampled, whether or not they are beneficiaries of the intervention, whether they are members of the target group (i.e. the treatment group) or the control group, may still differ in terms of some exogenous characteristics (such as age, gender, number of years of labour market experience, etc., for individuals, or number of employees, date of creation, short-term debt level, etc., for a firm). In the case of a non-stratified randomized controlled trial or a sharp regression discontinuity design, a simple regression of the observed outcome variable on a constant and a treatment group dummy variable is sufficient to obtain a convergent estimator of the average treatment effect in the sample. The addition of exogenous variables to this regression will mainly improve, in theory, the precision of the estimator of the average treatment effect.

However, in cases other than non-stratified randomization or sharp regression discontinuity design, it is necessary to add assumptions about the role of exogenous variables in order to obtain consistent estimators. The most commonly used assumption is that of conditional independence. This assumption states that the assignment to the treatment group, represented by a random variable T, and the potential outcomes of the intervention, denoted y_{1i} for a treated individual and y_{0i} for an untreated individual, are independent conditional on all relevant exogenous variables $\boldsymbol{x}$, i.e. all those affecting the probability of benefiting from the intervention. This assumption is crucial for implementing a technique such as matching. Once this hypothesis is accepted, if the sample is large enough

9. The only regions with weights well above zero are Madrid and Catalonia.
10. 2000 – 1974 = 26 years.

and/or the number of exogenous variables is not too high, it is possible to implement an exact matching method: this is based on comparing the outcome of each treated individual with that of an untreated individual having exactly the same observable characteristics. When this method cannot be implemented, particularly when the number of exogenous variables is too high, this exact matching is often replaced by a distance criterion making it possible to associate to each treated individual his/her "closest neighbour" in the sense of the chosen distance, or to implement the technique of the propensity score, as defined above: the outcome of each treated individual is compared with that of the untreated individual who has a propensity score whose value is very close to that of the treated individual's propensity score.[11] Exogenous variables that can be used to construct a valid propensity score should be conditionally independent of the assignment to the treatment group for a given value of this score.[12] The set of these exogenous variables is potentially extremely large. In addition to these variables, it is possible to include in this set some of their interactions, dichotomous indicators for those with multiple modalities (e.g. levels of education or socioprofessional categories), some transformations of these variables such as their powers or logarithms, etc.

Faced with the multiplicity of exogenous variables that can be mobilised, several recent studies have recommended the implementation of model and variable selection methods such as machine learning methods (McCaffrey *et al.*, 2004; Wyss *et al.*, 2014; Athey & Imbens, 2017a; Chernozhukov *et al.*, 2018), and LASSO[13] methods (Belloni *et al.*, 2014, 2017; Farrell, 2015). For example, McCaffrey *et al.* (2004), like Wyss *et al.* (2014), combine the method of random forests[14] with the LASSO technique in order to estimate the propensity score. It should be noted that these methods can be applied to evaluation methods other than matching. This is the case, in particular, of the method proposed by Belloni *et al.* (2017), which consists of a double variable selection procedure. The LASSO regression is used first to select the variables that are correlated with the outcome variable, and then again to select those that are correlated with the treatment dummy variable. After that, ordinary least squares can be applied by combining these two sets of variables, which improves the properties of the usual estimators of the average treatment effect, especially compared to simpler regularised regression techniques such as ridge regression.

The Heterogeneity of the Effects of an Intervention

Recent work has often focused on the heterogeneity of the effects of an intervention between groups of eligible individuals. Figure II illustrates this situation using a fictional example drawn from Leamer (1983). To make it easier to depict graphically, the heterogeneity of the treatment effect is assumed to be related to a variable x, the values of which differentiate individuals from each other. The left-hand side of Figure II describes the identification of the causal effect using a sample of individuals for whom the values of the exogenous variable, plotted on the x-axis, are dispersed only to a low extent. The variation in the outcome variable between individuals in the control group and those in the treatment group (i.e., the heterogeneity of the treatment effect) is measured by the slope of the regression line $\Delta(\mathbb{x})$, but it does not allow to disentangle between the many possible generalizations of the effect to other ranges of heterogeneity (of which two examples are drawn on Figure II). Looking also at the right-hand side of Figure II shows that having access to additional data, corresponding to greater heterogeneity among individuals ($x \in \mathbb{x} \cup \mathbb{x}'$), allows the analysis to be refined and pin down the distortion of the treatment effect in the population.

A wider range of observed situations therefore makes it possible to refine the estimation of the causal effect of the treatment, and to characterize its heterogeneity according to the observable characteristics of the individuals. As rich as the available data may be, however, the identification of the distribution of the treatment effect cannot be solved empirically. As an illustration, Figure III presents various measurements of the effect of a treatment, estimated for a wide range of values of the exogenous variable x. Nevertheless, these point values of the treatment effect are compatible with an infinite number of underlying distributions, of which Figure III presents three examples: $\Delta_a(x), \Delta_b(x)$, et $\Delta_c(x)$.

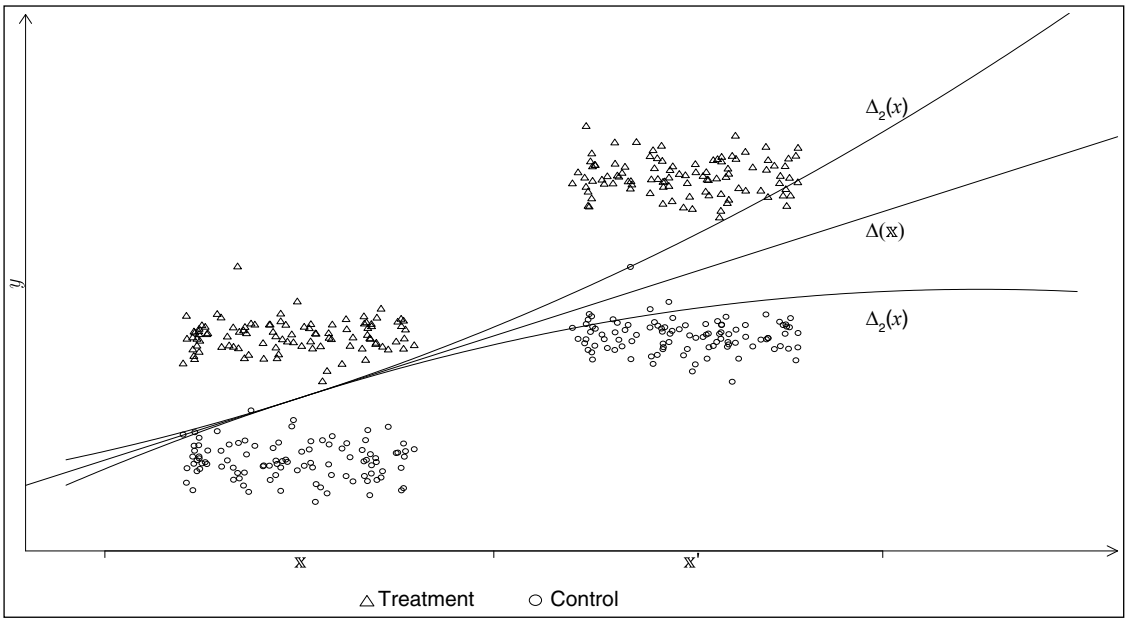
11. It is sometimes preferable to compare it with a weighted average of the outcomes of untreated individuals whose propensity scores have similar values. This is the principle that is implemented in the case of kernel matching.

12. This property is called the "balancing score property".

13. LASSO stands for Least Absolute Shrinkage and Selection Operator. This method, introduced by Tibshirani (1996), is a method for shrinking regression coefficients that essentially involves estimating the coefficient vector by minimizing the sum of the squared residuals under an additional regularisation constraint.

14. To implement this technique, the reader can in particular use the R package randomForest (https://cran.r-project.org/web/packages/randomForest/index.html).

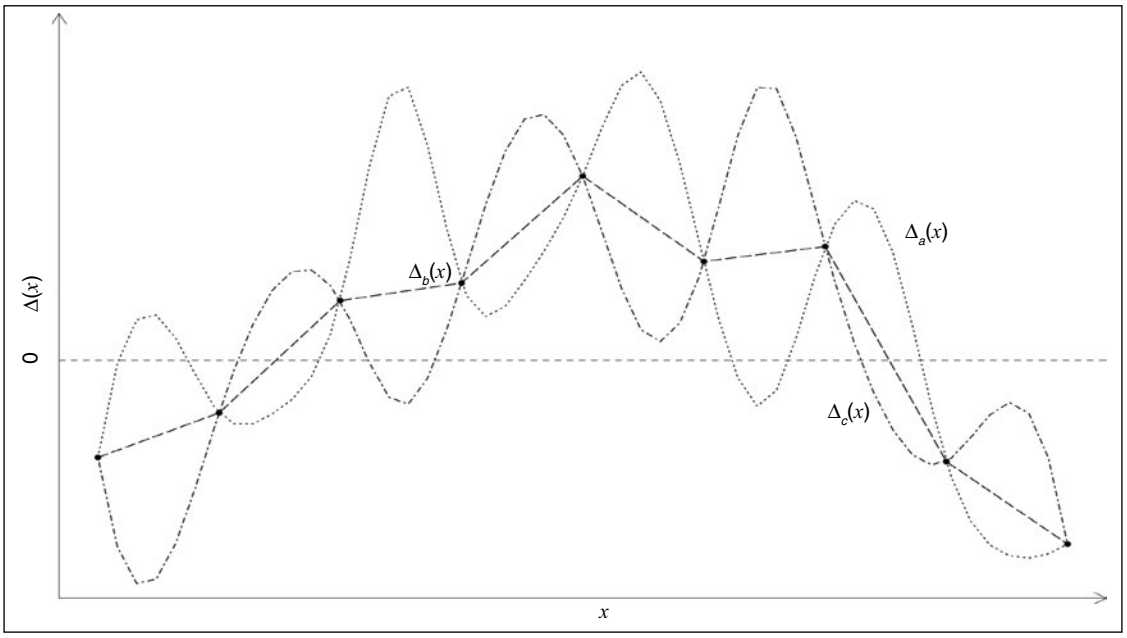
Figure II
Empirical identification of the effect of a treatment using an exogenous variable x with low ($x \in \mathbb{x}$) and high dispersion ($x \in \mathbb{x} \cup \mathbb{x}'$)



However fine the information provided by the data may be, and however heterogeneous the sample may be, the ability to describe the entire distribution of the treatment effect requires prior modelling to select the form of the relationship between the outcome variable and the treatment.

In the case where the sample is large and contains information on many variables, as it

Figure III
From the estimation to the identification of the distribution of the treatment effect



is the case with big data, it is possible to estimate heterogeneous treatment effects by combining quasi-experimental causal inference methods with LASSO methods and, more generally, with machine learning techniques (see, for example, Wager & Athey, 2018; Knaus *et al.*, 2017, 2018). This statistical approach can be generalised on a case-by-case basis with several treatments (Lechner, 2018).

Recent empirical work has focused on measuring the heterogeneity of effects, often in conjunction with the question of the external validity of the estimators used. Particularly compelling examples of this approach are given in the work of Dehejia *et al.* (2019) and Bisbee *et al.* (2017), who examine, using LATE-type estimators and data from more than a hundred international censuses, the causal link between fertility and female labour force participation. Their results are relatively convergent. Another example is provided by Allcott (2015), who assesses the variation in the effect of an energy reduction policy that has been gradually implemented at 111 sites in the United States: he finds that the effect of this policy has been stronger at the ten sites where the scheme was initially applied, suggesting that these first sites were selected because of their particular characteristics.

Precision of the Estimated Effects: The Quality of Identification beyond Unbiasedness

The attention paid to the estimation of causal effects in the policy evaluation literature has confined thoughts about identification to the unbiasedness of the estimated effects. In this context, the precision of the estimates is mainly addressed on the basis of the statistical significance of the estimated effects – an intervention being considered worthy of interest provided that its estimated effect is significantly different from 0.

A first limitation of statistical significance, which is well known but still largely overlooked in the empirical literature (see McCloskey & Ziliak, 1996; Ziliak & McCloskey, 2004), is that it does not make it possible to assess the quantitative importance of the measured effects. For each of these effects, statistical significance depends only on the precision of their estimation. A very small point estimate can thus be statistically very significant, while a very large effect can be insignificant due to its very low precision. In fact, hypothesis testing is nothing more than an alternative formulation of a confidence interval (provided the confidence level matches the level of the test). In this sense, statistical significance only provides information on whether the value zero belongs to the confidence interval built on the estimated parameter, i.e., to all the underlying effects compatible with the point estimate. Relying solely on statistical significance, whether to reject an intervention or to consider it beneficial, is tantamount to giving disproportionate weight to one of the many values within the confidence interval, many of which lead to a decision contrary to that indicated by statistical significance in the strict sense: in other words, a too wide confidence interval (i.e., a too imprecise estimation of an effect with a high point estimate) may lead to discard the intervention if this interval includes zero, or being considered beneficial if this interval, although gathering negligible values, is narrow enough to exclude zero (Amrhein *et al.*, 2019).

The attention paid to statistical precision must be just as close as the attention to the identification of causal effects. Improving precision requires in particular to minimize uncontrolled sources of variation. The control over the environment – i.e. blocking the sources of variation other than those of the variables of interest, such as the level of a "treatment" or the way it is administered – is an experimental approach that not only achieves identification but also increases the precision of the estimates (see the paper by Deaton & Cartwright, 2018, on this subject). Randomization, often presented in an excessive or even activist manner as the "golden rule" of policy evaluation, achieves identification of the causal effect based on the statistical similarity of the units belonging to the control and the treatment groups. It does not control, however, for all the unobserved factors that can add noise to the estimation.[15]

The importance given to the significance of the estimated effects may also lead to a certain number of deviations in the interpretation of the statistical tests. In particular, the limit value of the test statistic that leads to the rejection of the null hypothesis of no effect does not, in any way, measure the probability that the alternative hypothesis, stipulating the existence of an effect, is true. This probability is measured by the power of the test, the value of which is dependent on the distribution that

15. In a paper that is relatively critical of the mechanical applications of the randomized trial procedure, Deaton (2010) reviews the identification problems that remain despite random assignment to the treatment and control groups.

produces the test statistic when the alternative hypothesis is true, and therefore on the true (unknown) value from which the estimation results. An additional issue is that the p-value does not correspond either to the probability that the null hypothesis (i.e. the absence of effect) is true. This probability is indeed conditional on the null hypothesis: the distribution of the test statistic associated with the estimation is deduced from the value of the effect under the null hypothesis. If the calculated value of the test statistic is denoted $\hat{s}$ and the null hypothesis is denoted H_0, the p-value therefore formally measures the quantity $Pr(\hat{s} \mid H_0)$. The probability that the null hypothesis is true corresponds to the reverse conditioning, $Pr(H_0 \mid \hat{s})$. The confusion between these two probabilities can be illustrated by what the behavioural science literature calls the "prosecutor fallacy", introduced by Thompson & Schumann (1987): although, for example, the probability of winning at roulette without cheating is very low, it is obviously wrong to infer that a winner at roulette must be a cheater. Assessing the probability that the null hypothesis is true entails measuring the unconditional probability of this event, as illustrated in the next section.

The Increasing Risk of "False Positives" and the Need for Replication Work

Significance tests are subject to two types of risks of error: "false positives" are situations in which the estimation wrongly leads to thinking that a non-zero effect exists, and "false negatives" relate to the opposite situation, where the absence of an estimated relationship is only apparent. The respective probabilities of these cases correspond to the Type I error (also known as the "level" of the test), which is often denoted α and the most commonly chosen value of which is 5%, and the Type II error, β, which is the opposite of the power, $P = 1 - \beta$. The power measures the probability of detecting the effect of the intervention and depends on the intensity of that effect: it does not correspond to a probability, but to a function that also depends crucially on the sample size.[16]

An estimated effect is "statistically significant at the 5% threshold" if the probability of getting this estimate while the effect is actually zero is less than 5%. This property implies a 5% probability of making a mistake when concluding that the estimated effect of an intervention is statistically significant. This probability is often interpreted as measuring the proportion of statistically significant results that are incorrect. This conclusion is only true in very specific circumstances, and the consequences of Type I errors on the credibility of empirical work are in fact often much more serious than its value suggests.

To illustrate this point, Wacholder *et al.* (2004) describe the components of the False-Positive Report Probability (hereinafter denoted "FPRP") as a function of the statistical properties of significance tests. The FPRP is the probability that the effect of an intervention is actually zero, even though the estimation produces a statistically significant effect. The calculation of this probability involves an unknown quantity (which is not usually discussed, even though it is fundamental) that corresponds to the proportion, denoted $\bar{y}$, of interventions that have a non-zero effect amongst all the interventions that are being evaluated. Table 1 describes the probability of occurrence of the four types of possible situations: the legitimate detection of an absence

16. *The benchmark power level in applied work is 80%, although Ioannidis* et al. *(2017) show that in more than half of applied economics work, the median power is 18% or even less.*

Table 1
Components of the probability of occurrence of a false positive

Veracity of the alternative hypothesis	Statistical significance test		Total
	Significant	Insignificant	
Non-zero effect of the intervention	$(1-\beta)\bar{y}$ [True positive]	$\beta\bar{y}$ [False negative]	$\bar{y}$
Zero effect of the intervention	$\alpha(1-\bar{y})$ [False positive]	$(1-\alpha)(1-\bar{y})$ [True negative]	$(1-\bar{y})$
Total	$(1-\beta)\bar{y} + \alpha(1-\bar{y})$	$\beta\bar{y} + (1-\alpha)(1-\bar{y})$	1

Notes: Subject to the existence or absence of an intervention effect, each of the cells describes the probability that the estimated effect is statistically significant (first column) or statistically insignificant (second column), taking account of the level α of the test, its power β, and the proportion $\bar{y}$ of interventions that have a non-zero effect amongst all those evaluated.
Sources: Wacholder *et al.* (2004, p. 440).

(true negative) or presence (true positive) of an intervention effect, as well as the occurrence of false positives, or false negatives.

Given the probabilities of Type I and Type II errors, the probability of a false positive occurring (the proportion of effects that are only apparent amongst all the interventions having a significant effect) is measured by:

$$FPRP = \frac{\alpha(1-\bar{y})}{\alpha(1-\bar{y})+(1-\beta)\bar{y}}$$

Most of the commonly used statistical tests are consistent, i.e. their power tends towards one as the sample size increases. In this very favourable situation (where $\beta = 0$), this probability is less than the level α of the test only if at least half of all the interventions that are evaluated have a non-zero effect. If this frequency is higher, the probability of occurrence of false positives is lower than the level of the test. It is higher than this level under the opposite (and certainly more credible) hypothesis that, of all the interventions evaluated, less than one in two has a non-zero effect, a situation that is all the more likely to occur as more evaluations are undertaken. It is of course impossible to quantify $\bar{y}$, and very difficult to collect objective information on this proportion. Still, the consequences of the variations of $\bar{y}$ on the credibility attributed to the results of evaluations are not without importance: under the extreme hypothesis that one intervention out of 1,000 has a non-zero effect ($\bar{y}$ = 0,001), the probability of reporting false positives is greater than 98%.

This situation may be further aggravated by the conditions under which the results of the evaluation are made public.[17] Ioannidis (2005) focuses in particular on two types of bias that increase the probability of reporting false positives: publication bias and communication bias. Publication bias refers to the particular appeal of works highlighting a non-zero effect at all stages of the process – from project-funding decisions, to the results being communicated to the general public, after having been validated academically by being published in prestigious scientific journals. These publication biases lead to a distorted proportion of positive results. They are reinforced through communication biases, which consist in reporting on an evaluation only if it leads to significant effects, while at the same time not reporting evaluation results that conclude to no effect of other kinds of interventions. As stressed by Roth (1994), this risk is particularly high when an intervention is developed following a trial and error process, which leads to changes in the terms and conditions of a "pilot" intervention after it has been found to have no effect, until a final proposal is developed that gives rise to the expected significant effect on the outcome. This process is legitimate because it allows to design effective public policies; it does not affect the probability of reporting false positives if all trials are made public at the same time as the final evaluation. Conversely, this process leads to a communication bias as soon as only significant effects are made public, while previous unsuccessful attempts are ignored.

Publication biases, like communication biases, lead to an increase in the proportion of false positives. To illustrate this point, the proportion of positive results caused by one of these two types of bias is denoted B. Amongst the $\bar{y}$ interventions that actually have an effect, the analysis will make it possible to accurately conclude that there is a non-zero effect for a proportion $(1-\beta)$ of cases, while a certain number $(B \times \beta)$ will appear to have an effect due to one of the types of biases. Similarly, a proportion α of interventions amongst the $(1-\bar{y})$ actually having zero effect will appear as having no effect, while a certain number $B \times (1-\alpha)$ will appear as having a non-zero effect due to bias. In total, the FPRP becomes:

$$FPRP = \frac{(1-\bar{y})\left[\alpha + B(1-\alpha)\right]}{(1-\bar{y})\left[\alpha + B(1-\alpha)\right] + (1-\beta)\bar{y} + B\beta\bar{y}}$$

* *
*

For the "credibility revolution" announced by some authors (Angrist & Pischke, 2010) to be fully successful, public policy evaluation cannot be based solely on convincing identification strategies. The replication of policy evaluation results, making it possible to distinguish false positives from the proven effects of an intervention (Clemens, 2017), remains essential, as is the need to ensure the precision of the estimated effects. □

17. We have deliberately left out the issue of questionable practices that deliberately force the significance of results, for example by deliberately choosing the outcome variable from among all the variables on which the intervention may act, a practice that artificially increases the proportion of false positives (see, for example, List et al.*, 2001). Christensen & Miguel (2018) present an overview of practices that cause the credibility of empirical results in economics to be weakened, and list a certain number of possible solutions.*

BIBLIOGRAPHY

Abadie, A. & Cattaneo, M. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10, 465–503.
https://dx.doi.org/10.1146/annurev-economics-080217-053402

Abadie, A., Diamond, A. & Hainmueller, J. (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association*, 105(490), 493–505.
https://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., Diamond, A. & Hainmueller, J. (2015). Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2), 495–510.
https://doi.org/10.1111/ajps.12116

Abadie, A. & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review,* 93(1), 113–32.
https://doi.org/10.1257/000282803321455188

Abadie, A. & L'Hour, J. (2019). A penalized synthetic control estimator for disaggregated data. *Mimeo*.

Abbring, J. H. & Heckman, J. J. **(2007).** Econometric evaluation of social programs, Part III: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In: Heckman, J. J. & Leamer, E. (Eds.), *Handbook of Econometrics*, vol. 6, part B, chapter 72, pp. 5145–5303. Amsterdam: Elsevier

Amrhein, V., Greenland, S. & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567, 305–307.
https://doi.org/10.1038/d41586-019-00857-9

Aeberhardt, R., Fougère, D. & Rathelot, R. (2011). Les méthodes de testing permettent-elles d'identifier et de mesurer l'ampleur des discriminations ? *Économie et statistique*, 447, 97–101.
https://www.insee.fr/fr/statistiques/1377350?sommaire=1377352

Allcott, H. (2015). Site Selection Bias in Program Evaluation. *Quarterly Journal of Economics*, 130(3), 1117–1165.
https://doi.org/10.1093/qje/qjv015

Amjad, M. J., Shah, D. & Shen, D. (2017). Robust Synthetic Control. *Journal of Machine Learning Research*, 19(22), 1–51.
http://www.jmlr.org/papers/volume19/17-777/17-777.pdf

Angrist, J. (2004). Treatment Effect Heterogeneity In Theory And Practice. *Economic Journal*, 114(494), 52–83
https://doi.org/10.1111/j.0013-0133.2003.00195.x

Angrist, J. & Fernandez-Val, I. (2013). ExtrapoLATE-ing: External validity and overidentification in the LATE framework. In: Acemoglu, D., Arellano, M. & Dekel, E. (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress, Volume III: Econometrics*. Cambridge: Cambridge University Press.

Angrist, J., Graddy, K. & Imbens, G. (2000). The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish. *Review of Economic Studies*, 67(3), 499–527.
https://doi.org/10.1111/1467-937X.00141

Angrist, J. & Imbens, G. (1995). Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *Journal of the American Statistical Association*, 90(140), 431–442.
https://scholar.harvard.edu/imbens/publications/two-stage-least-squares-estimation-average-causal-effects-models-variable-treatm

Angrist, J. & Krueger, A. B. (1999). Empirical strategies in labor economics. In: Ashenfelter, O. C. & Card, D. (Eds.), *Handbook of Labor Economics*, vol. 3, part A, chapter 23, pp. 1277–1366. Amsterdam: Elsevier.

Angrist, J. & Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con Out of Econometrics. *Journal of Economic Perspectives*, 24(2), 3–30.
https://doi.org/10.1257/jep.24.2.3

Angrist, J. & Rokkanen, M. (2015). Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff. *Journal of the American Statistical Association*, 110(512), 1331–1344.
https://doi.org/10.1080/01621459.2015.1012259

Aronow, P. (2012). A General Method for Detecting Interference in Randomized Experiments. *Sociological Methods and Research*, 41(1), 3-16.
https://doi.org/10.1177%2F0049124112437535

Athey, S., Bayatiz, M., Doudchenkox, N., Imbens, G. & Khosravik, K. **(2018).** Matrix Completion Methods for Causal Panel Data Models.
2018, NBER *Working Paper* N° 25132

Athey, S. & Imbens, G. (2017a). The State of Applied Econometrics: Causality and Policy Evaluation. *Journal of Economic Perspectives*, 31(2), 3–32. https://doi.org/10.1257/jep.31.2.3

Athey, S. & Imbens, G. (2017b). Econometrics of randomized experiments. In: Banerjee, A. V. & Duflo, E. (Eds.), *Handbook of Economic Field Experiments*, vol. 1, chapter 3, pp. 73–140. Amsterdam : North-Holland.

Avvisati, F., Gurgand, M., Guyon, N. & Maurin, E. (2014). Getting parents involved: A field experiment in deprived schools. *Review of Economic Studies*, 81(1), 57–83, 2014. https://doi.org/10.1093/restud/rdt027

Baird, S., Bohren, J. A., McIntosh, C. & Özler, B. (2018). Optimal Design of Experiments in the Presence of Interference. *The Review of Economics and Statistics,* 100(5), 844–860. https://doi.org/10.1162/rest_a_00716

Baraton, M., Beffy, M. & Fougère, D. (2011). Une évaluation de l'effet de la réforme de 2003 sur les départs en retraite. Le cas des enseignants du second degré public. *Économie et statistique*, 441-442, 55–78. https://www.insee.fr/fr/statistiques/fichier/1377513/ES441D.pdf

Barone, C., Fougère, D. & Pin, C. (2019). Social origins, shared book reading and language skills in early childhood: evidence from an information experiment. *European Sociological Review,* forthcoming.

Beffy, M., Fougère, D. & Maurel, A. (2009). L'impact du travail salarié des étudiants sur la réussite et la poursuite des études universitaires. *Economie et statistique*, 422, 31–50. https://www.insee.fr/fr/statistiques/1376784?sommaire=1376788

Behaghel, L., Crépon, B. & Sédillot, B. (2004). Contribution Delalande et transitions sur le marché du travail. *Économie et statistique*, 372, 61–88. https://www.insee.fr/fr/statistiques/1376608?sommaire=1376612

Belloni, A., Chernozhukov, V. & Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2), 608–650. https://doi.org/10.1093/restud/rdt044

Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2017). Program Evaluation and Causal Inference with High-Dimensional Data. *Econometrica*, 85(1), 233–298. https://doi.org/10.3982/ECTA12723

Bénabou, R., Kramarz, F. & Prost, C. (2004). Zones d'éducation prioritaire : quels moyens pour quels résultats ? *Économie et Statistique*, 380, 3–29. https://www.insee.fr/fr/statistiques/1376492?sommaire=1376498

Bérard, G. & Trannoy, A. (2018). The impact of the 2014 increase in the real estate transfer taxes on the French housing market. *Economie et Statistique / Economics and Statistics*, 500-501-502, 179–200. https://www.insee.fr/en/statistiques/3622039?sommaire=3622133

Bertanha, M. & Imbens, G. (2019). External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, forthcoming.

Bisbee, J., Dehejia, R., Pop-Eleches, C. & Samii, C. (2017). Local Instruments, Global Extrapolation: External Validity of the Labor Supply-Fertility Local Average Treatment Effect. *Journal of Labor Economics*, 35(S1), S99–S147. https://doi.org/10.1086/691280

Bozio, A. (2011). La réforme des retraites de 1993 : l'impact de l'augmentation de la durée d'assurance. *Économie et Statistique*, 441-442, 39–53. https://www.insee.fr/fr/statistiques/1377511?sommaire=1377529

Brodaty, T., Crépon, B. & Fougère, D. (2007). Les méthodes micro-économétriques d'évaluation et leurs applications aux politiques actives de l'emploi. *Économie & prévision*, 177(1), 93–118. https://www.cairn.info/revue-economie-et-prevision-2007-1-page-93.htm

Bunel, M., Gilles, F. & L'Horty, Y. (2009). Les effets des allégements de cotisations sociales sur l'emploi et les salaires : une évaluation de la réforme de 2003. *Économie et Statistique*, 429-430, 77–105. https://www.insee.fr/fr/statistiques/1377396?sommaire=1377406

Carbonnier, C. (2009). Réduction et crédit d'impôt pour l'emploi d'un salarié à domicile, conséquences incitatives et redistributives. *Économie et Statistique*, 427-428, 67–100. https://www.insee.fr/fr/statistiques/1377124?sommaire=1377130

Chabé-Ferret, S. (2015). Analysis of the bias of matching and difference-in-difference under alternative earnings and selection processes. *Journal of Econometrics*, 185(1), 110–123. https://www.sciencedirect.com/science/article/pii/S0304407614002437

Chabé-Ferret, S., Dupont-Courtade, L. & Treich, N. (2017). Évaluation des politiques publiques : expérimentation randomisée et méthodes quasi-expérimentales. *Economie & prévision*, 211-212(2), 1–34. https://www.cairn.info/revue-economie-et-prevision-2017-2-page-1.htm.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N. & Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: a meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37(2), 403–429. https://doi.org/10.1002/pam.22051

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

Christensen, G. & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, 56(3), 920–980. https://doi.org/10.1257/jel.20171350

Clemens, M. A. (2017). The Meaning of Failed Replications: A Review and Proposal. *Journal of Economic Surveys*, 31(1), 326–342. https://doi.org/10.1111/joes.12139

Connolly, P., Keenan, C. & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomized controlled trials in education research 1980–2016. *Educational Research,* 60(3), 276–291. https://doi.org/10.1080/00131881.2018.1493353

Crépon, B. & Desplatz, R. (2001). Une nouvelle évaluation des effets des allégements de charges sociales sur les bas salaires. *Économie et statistique*, 348, 3–24. https://www.insee.fr/fr/statistiques/1376044?sommaire=1376054

Crépon, B., Devoto, F., Duflo, E. & Parienté, W. (2015). Estimating the impact of microcredit on those who take it up: evidence from a randomized experiment in Morocco. *American Economic Journal: Applied Economics*, 7(1), 123–150. https://doi.org/10.1080/00131881.2018.1493353

Crépon, B., Duflo, E., Gurgand, M., Rathelot, R. & Zamora, P. (2013). Do labor market policies have displacement effects: evidence from a clustered randomized experiment. *The Quarterly Journal of Economics*, 128(2), 531-580. https://doi.org/10.1093/qje/qjt001

Crépon, B. & Jacquemet, N. (2018). *Econométrie : Méthodes et Applications*, *2ème édition*. Louvain-la-Neuve : De Boeck Universités.

Crépon, B., Leclair, M. & Roux, S. (2004). RTT, productivité et emploi : nouvelles estimations sur données d'entreprises. *Économie et Statistique*, 376-377, 55–89. https://www.insee.fr/fr/statistiques/1376466?sommaire=1376476

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of Economic Literature*, 48(2), 424–55. https://doi.org/10.1257/jel.48.2.424

Deaton, A. & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Dehejia, R., Pop-Eleches, C. & Samii, C. (2019). From Local to Global: External Validity in a Fertility Natural Experiment. *Journal of Business & Economic Statistics*, forthcoming.

Dong, Y. & Lewbel, A. (2015). Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models. *Review of Economics and Statistics*, 97(5), 1081–1092. http://dx.doi.org/10.1162/REST_a_00510

Duflo, E. & Banerjee, A. (2017). *Handbook of Field Experiments, Vol. 1 & 2*. Amsterdam: North-Holland.

Edo, A. & Jacquemet, N. (2013). Discrimination à l'embauche selon l'origine et le genre : défiance indifférenciée ou ciblée sur certains groupes ? *Économie et Statistique*, 464-466, 155–172. https://www.insee.fr/fr/statistiques/1378023?sommaire=1378033

Even, K. & Klein, T. (2007). Les contrats et stages aidés : un profit à moyen terme pour les participants ? Les exemples du CIE, du CES et du Sife. *Économie et Statistique*, 408-409, 3–32. https://www.insee.fr/fr/statistiques/1377206?sommaire=1377217

Fack, G. & Landais, C. (2009). Les incitations fiscales aux dons sont-elles efficaces ? *Économie et Statistique*, 427-428, 101–121. https://www.insee.fr/fr/statistiques/1377126?sommaire=1377130

Farrell, M. H. (2015). Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations. *Journal of Econometrics*, 189(1), 1–23. https://dx.doi.org/10.2139/ssrn.2324292

Fougère, D. & Poulhès, M. (2014). La propriété immobilière : quelle influence sur le portefeuille financier des ménages ? *Économie et Statistique*, 472-473, 213–231.
https://www.insee.fr/fr/statistiques/1377779?sommaire=1377781

Frölich, M. & Sperlich, S. (2019). *Impact Evaluation: Treatment Effects and Causal Analysis*. Cambridge : Cambridge University Press.

Geniaux, G. & Napoléone, C. (2011). Évaluation des effets des zonages environnementaux sur la croissance urbaine et l'activité agricole. *Économie et Statistique*, 444-445, 181–199.
https://www.insee.fr/fr/statistiques/1377857?sommaire=1377863

Givord, P. (2014). Méthodes économétriques pour l'évaluation de politiques publiques. *Économie & prévision,* 204-205(1), 1–28.
https://www.cairn.info/revue-economie-et-prevision-2014-1-page-1.htm

Glazerman, S., Levy, D. M. & Myers, D. (2003). Nonexperimental Versus Experimental Estimates of Earnings Impacts. *The Annals of the American Academy of Political and Social Science*, 589(1), 63–93.
https://doi.org/10.1177%2F0002716203254879

Goux, D., Gurgand, M. & Maurin, E. (2017). Adjusting Your Dreams? Highschool Plans and Dropout Behaviour. *Economic Journal*, 127(602), 1025–1046.
https://dx.doi.org/10.1111/ecoj.12317

Hahn, J. & Shi, R. (2017). Synthetic Control and Inference. *Econometrics*, 54(2), 52.
https://doi.org/10.3390/econometrics5040052

Hill, J. (2008). Comment. *Journal of the American Statistical Association*, 103(484), 1346–1350.
https://doi.org/10.1198/016214508000001002

Heckman, J. J., Lalonde, R. & Smith, J. (1999). The economics and econometrics of active labor market programs. In: Ashenfelter, O. C. & Card, D. (Eds.), *Handbook of Labor Economics*, vol. 3, part A, chapter 3, pp. 865–2097. Amsterdam: Elsevier.

Heckman, J. J. & Vytlacil, E. J. (2007a). Econometric evaluation of social programs, Part I: Causal models, structural models and econometric policy evaluation. In: Heckman, J. J. & Leamer, E. (Eds.), *Handbook of Econometrics*, vol. 6, part B, chapter 70, pp. 4779–4874.

Heckman, J. J. & Vytlacil, E. J. (2007b). Econometric evaluation of social programs, Part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. In: Heckman, J. J. & Leamer, E. (Eds.), *Handbook of Econometrics*, vol. 6, part B, chapter 71, pp. 4875–5143. Amsterdam: Elsevier.

Hotz, V. J., Imbens, G. & Mortimer, J. H. (2005). Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations. *Journal of Econometrics*, 125(1–2), 241–70.
http://dx.doi.org/10.1016/j.jeconom.2004.04.009

Hudgens, M. & Halloran, E. (2008). Towards Causal inference With Interference. *Journal of the American Statistical Association*, 103(482), 832–842.
https://doi.org/10.1198/016214508000000292

Imai, K., King, G. & Stuart, E. (2008). Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A*, 171(2), 481–502.
http://dx.doi.org/10.1111/j.1467-985X.2007.00527.x

Imbens, G. (2010). Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of Economic Literature*, 48(2), 399–423.
https://doi.org/10.1257/jel.48.2.399

Imbens, G. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity: A review. *Review of Economics and Statistics*, 86(1), 4–29.
http://dx.doi.org/10.1162/003465304323023651

Imbens, G. & Angrist, J. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2), 467–475.
https://doi.org/10.2307/2951620

Imbens, G. & Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge : Cambridge University Press.

Imbens, G. & Wooldridge, J. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5–86.
http://dx.doi.org/10.1257/jel.47.1.5

Ioannidis, J. P. A. (2005). Why Most Published Research Findings are False. *PLoS Med*, 2(8), e124.
https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A., Stanley, T. D. & Doucouliagos, H. (2017). The Power of Bias in Economics Research. *Economic Journal*, 127(605), F236–F265.
https://doi.org/10.1111/ecoj.12461

Jacquemet, N. & L'Haridon, O. (2018). *Experimental Economics: Method and Applications*. Cambridge : Cambridge University Press.

Knaus, M. C., Lechner, M. & Strittmatter, A. (2017). Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach. IZA *Discussion Paper* N° 10961. https://ssrn.com/abstract=3029832

Knaus, M. C., Lechner, M. & Strittmatter, A. (2018). Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence. IZA *Discussion Paper* N° 12039. https://ssrn.com/abstract=3318814

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4), 604–620. https://www.jstor.org/stable/1806062

Leclair, M. & Roux, S. (2007). Productivité relative et utilisation des emplois de courte durée dans les entreprises. *Économie et Statistique*, 405-406, 47–76. https://www.insee.fr/fr/statistiques/1376937?sommaire=1376947

Leamer, E. E. (1983). Let's take the con out of econometrics. *American Economic Review*, 73(1), 31–43. https://doi.org/10.1257/jep.24.2.3

Lechner, M. (2018). Modified Causal Forests for Estimating Heterogeneous Causal Effects. IZA *Discussion Paper* N° 12040. https://www.iza.org/publications/dp/12040/modified-causal-forests-for-estimating-heterogeneous-causal-effects

Lee, M. J. (2016). *Matching, Regression Discontinuity, Difference in Differences, and Beyond*. Oxford : Oxford University Press.

List, J. A., Bailey, C., Euzent, P. & Martin, T. (2001). Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior. *Economic Inquiry*, 39(1), 162–170. https://doi.org/10.1111/j.1465-7295.2001.tb00058.x

Liu, L. & Hudgens, M. (2014). Large Sample Randomization Inference of Causal Effects in the Presence of Interference. *Journal of the American Statistical Association*, 109(505), 288–301. http://dx.doi.org/10.1080/01621459.2013.844698

Lorenceau, A. (2009). L'impact d'exonérations fiscales sur la création d'établissements et l'emploi en France rurale : une approche par discontinuité de la régression. *Économie et Statistique,* 427-428, 27–62. https://www.insee.fr/fr/statistiques/1377120?sommaire=1377130

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1), S1–S23. https://doi.org/10.1111/j.1368-423X.2012.00368.x

McCaffrey, D. F., Ridgeway, G. & Morral, A. R. (2004). Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychological Methods*, 9(4), 403–425. https://psycnet.apa.org/doi/10.1037/1082-989X.9.4.403

McCloskey, D. N. & Ziliak, S. T. (1996). The Standard Error of Regressions. *Journal of Economic Literature*, 34(1), 97–114. https://www.jstor.org/stable/2729411

Meager, R. (2019). Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments. *American Economic Journal: Applied Economics*, 11(1), 57–91. https://doi.org/10.1257/app.20170299

Pearl, J. & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Peters, J., Janzing, D. & Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge : The MIT Press.

Petit, P., Duguet, E., L'Horty, Y., du Parquet, L. & Sari, F. (2013). Discrimination à l'embauche : les effets du genre et de l'origine se cumulent-ils systématiquement ? *Économie et Statistique*, 464-466, 141–153. https://doi.org/10.3406/estat.2013.10234

Petit, P., Sari, F., L'Horty, Y., Duguet, E. & du Parquet, L. (2011). Les effets du lieu de résidence sur l'accès à l'emploi : un test de discrimination auprès des jeunes qualifiés. *Économie et Statistique*, 447, 71–95. https://doi.org/10.3406/estat.2011.9711

Rathelot, R. & Sillard, P. (2008). Zones Franches Urbaines : quels effets sur l'emploi salarié et les créations d'établissements ? *Économie et Statistique*, 415-416, 81–96. https://doi.org/10.3406/estat.2008.7021

Roth, A. E. (1994). Let's Keep the Con Out of Experimental Econ.: A Methodological Note. *Empirical Economics*, 19(2), 279–289. https://econpapers.repec.org/RePEc:spr:empeco:v:19:y:1994:i:2:p:279-89

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology,* 66(5), 688–701. https://psycnet.apa.org/doi/10.1037/h0037350

Simonnet, V. & Danzin, E. (2014). L'effet du RSA sur le taux de retour à l'emploi des allocataires. Une analyse en double différence selon le nombre et l'âge des enfants. *Économie et Statistique*, 467-468, 91–116. https://doi.org/10.3406/estat.2014.10248

Stuart, E. A., Bradshaw, C. P. & Leaf, P. J. (2015). Assessing the Generalizability of Randomized Trial Results to Target Populations. *Prevention Science*, 16(3), 475–485.
http://dx.doi.org/10.1007/s11121-014-0513-z

Stuart, E. A., Cole, S. R., Bradshaw, C. P. & Leaf, P. J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society, Series A*, 174(2), 369–386.
https://doi.org/10.1111/j.1467-985X.2010.00673.x

Thompson, W. C. & Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials. *Law and Human Behavior*, 11(3), 167–187.
https://psycnet.apa.org/doi/10.1007/BF01044641

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, 58(1), 267–288.
https://www.jstor.org/stable/2346178

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L. & Rothman, N. (2004). Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies. *Journal of the National Cancer Institute*, 96(6), 434–442.
https://doi.org/10.1093/jnci/djh075

Wager, S. & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests. *Journal of the American Statistical Association*, 113(523), 1228–1242.
https://doi.org/10.1080/01621459.2017.1319839

Wong, V. C., Valentine, J. C. & Miller-Bains, K. (2017). Empirical Performance of Covariates in Education Observational Studies. *Journal of Research on Educational Effectiveness*, 10(1), 207–236.
https://doi.org/10.1080/19345747.2016.1164781

Wyss, R., Ellis, A., Brookhart, A., Girman, C., Jonsson Funk, M., LoCasale, R. & Stürmer, T. (2014). The Role of Prediction Modeling in Propensity Score Estimation: An Evaluation of Logistic Regression, bCART, and the Covariate-Balancing Propensity Score. *American Journal of Epidemiology*, 180(6), 645–655.
https://dx.doi.org/10.1093%2Faje%2Fkwu181

Ziliak, S. T. & McCloskey, D. N. (2004). Size matters: the standard error of regressions in the American Economic Review. *Journal of Socioeconomics*, 33(5), 527–546.
https://doi.org/10.1016/j.socec.2004.09.024