I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12922

Policy Evaluation Using Causal Inference Methods

Denis Fougère
Nicolas Jacquemet

JANUARY 2020

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 12922

Policy Evaluation Using Causal Inference Methods

Denis Fougère
CNRS, OSC/LIEPP, CEPR and IZA

Nicolas Jacquemet
Universite Paris 1 Pantheon-Sorbonne and Paris School of Economics

JANUARY 2020

Any opinions expressed in this paper are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but IZA takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The IZA Institute of Labor Economics is an independent economic research institute that conducts research in labor economics and offers evidence-based policy advice on labor market issues. Supported by the Deutsche Post Foundation, IZA runs the world's largest network of economists, whose research aims to provide answers to the global labor market challenges of our time. Our key objective is to build bridges between academic research, policymakers and society.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Policy Evaluation Using Causal Inference Methods*

This chapter describes the main impact evaluation methods, both experimental and quasi-experimental, and the statistical model underlying them. Some of the most important methodological advances to have recently been put forward in this field of research are presented. 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 assessments, 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

Corresponding author:
Denis Fougère
LIEPP - Sciences Po
27, rue Saint-Guillaume
75337 Paris Cedex 07
France
E-mail: denis.fougere@sciencespo.fr

* 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).

1 Introduction

Over the past twenty years, the number of impact evaluation studies, whether experimental or quasi-experimental, has increased exponentially. 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 (see Figure 1). Quasi-experimental evaluations have followed a similar trend and nowadays together constitute what some have called an "empirical revolution" (Angrist and Pischke, 2010). These studies and the quantitative assessments that they contain are resources of prime importance when it comes to choosing, designing and implementing public policies.

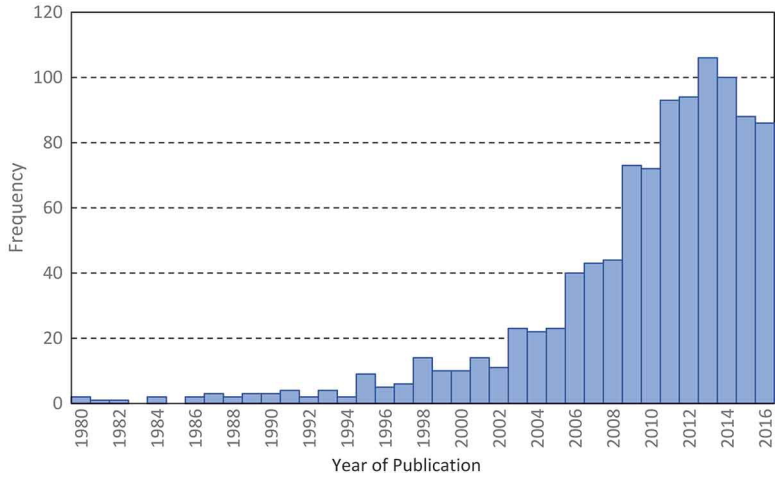
The recent publication of several reference articles and books also shows just how developed and diverse econometric evaluation methods have become. These include the books by Imbens and Rubin (2015), Lee (2016), Sperlich and Frölich (2019), which follow on from the survey papers by Angrist and Krueger (1999), Heckman, Lalonde, and Smith (1999), Heckman and Vytlacil (2007a), Heckman and Vytlacil (2007b), Abbring and Heckman (2007), and Imbens and Wooldridge (2009). The *Handbook of Field Experiments* published by Duflo and Banerjee (2017) is the reference book on randomised field experiments. For laboratory experiments, the book by Jacquemet and L'haridon (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.

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 the 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, firms, students, etc.) benefiting from that policy. 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.[1]

This chapter is devoted to presenting the developments that are particularly promising in the field of impact evaluation methods. Section 2 presents in a non-technical way these methods, both experimental and quasi-experimental, in order to familiarize the reader with their general principles. Section 3 describes the canonical statistical model of potential outcomes which is the basis of all these methods. Then it introduces the main causal estimands (i.e., parameters) and

[1] An extensive review of recent developments and future research directions can be found in the papers written by Athey and Imbens (2017a,b), and Abadie and Cattaneo (2018).

Figure 1: Publication trends of RCTs



Note. Number of randomised controlled trials conducted between 1980 and 2016 in the field of educational policy that have been published in international scientific journals, according to Connolly, Keenan, and Urbanska (2018).

presents more formally each impact evaluation method. Section 4 discusses the external validity of impact evaluation methods. The synthetic control approach, which overcomes some limitation of the difference-in-differences method, is presented in Section 5. Section 6 underlines the role and the choice of explanatory variables (i.e., covariates). Beyond the importance of the choice of the parameter to be estimated (which must take precedence over the choice of the identification method) and the choice of relevant covariates to be incorporated into the model, 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. That is why Section 7 is devoted to methods dealing with heterogeneous effects of interventions. In Section 8, we focus on the need to pay particular attention to the accuracy of the estimated effects. Section 9 emphasizes the need of replication studies, carried out by experimentation or quasi-experimentation, in order to distinguish false positives from proven effects. Finally, Section 10 is devoted to the crucial issue of interference in experimental and quasi-experimental studies.

2 A Non-Technical Review of Policy Evaluation 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.

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 market 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 for a long time used the instrumental variable method. 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 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", does not. The employment policy is implemented between the two dates under consideration. The method relies on the following assumption: in the absence of that 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)[2] 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 article

[2]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".

published by LaLonde (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, Levy, and Myers (2003), Chabé-Ferret (2015), Wong, Valentine, and Miller-Bains (2017), and Chaplin, Cook, Zurovac, Coopersmith, Finucane, Vollmer, and Morris (2018).

3 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 (or treatment) $\mathcal{T}_i \in \{0, 1\}$. In this model, each observation unit is characterized by two "potential outcomes": y_{0_i} is the outcome that would be observed for the unit i in the absence of the intervention, and y_{1_i} 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 in a sample of size N: $\Delta_i = y_{1_i} - y_{0_i}$, $\forall i = 1, \ldots, N$. The first step of the evaluation process is therefore to choose how this distribution will be approached based on the limited information delivered by the sample. Among the many parameters that could be used to summarize this distribution, the most common are the average treatment effect and the average treatment effect on the treated.

The average treatment effect (ATE), $\Delta^{\text{ATE}} = \mathbb{E}(y_{1_i} - y_{0_i})$ measures the average change in outcome for an individual randomly selected from the population. The average treatment effect on the treated (ATT), $\Delta^{\text{ATT}} = \mathbb{E}(y_{1_i} - y_{0_i} | \mathcal{T}_i = 1)$, is specific to the sub-population of individuals who actually benefit from the program. The two parameters are only equal under very restrictive assumptions. For example, they trivially coincide if the intervention is applied to 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 $\Delta_i = \Delta (= \Delta^{\text{ATE}} = \Delta^{\text{ATT}})$, $\forall i$. 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 which is chosen strongly influences the parameter that can be measured.

3.1 Causal effects and endogenous participation

The observed outcome y_i is linked to the potential outcomes through the individual's treatment status: $y_i = \mathcal{T}_i y_{1_i} + (1 - \mathcal{T}_i) y_{0_i}$. This relationship is sufficient to highlight the two issues that evaluation must overcome: a missing data problem, and the endogenity of the treatment. The first issue arises because data only deliver partial information about either y_{1_i} for individuals who received the treatment, or y_{0_i} for those who did not. Such data allows to compute the distribution of the treatment in the population, $\ell(\mathcal{T}_i)$, as well as the distribution of the outcome resulting

from the treatment among beneficiaries, $\ell(y_{1_i}|\mathcal{T}_i=1)=\ell(y_i|\mathcal{T}_i=1)$, and the distribution of the outcome absent the treatment for non-beneficiaries, $\ell(y_{0_i}|\mathcal{T}_i=0)=\ell(y_i|\mathcal{T}_i=0)$. But it does not allow to compute any of the relevant policy parameters. Evaluation methods must thus rely on counterfactual observations — information on what would have happened in a state of the world that is not observed.

Finding counterfactuals is even more challenging because of the second issue, i.e., the endogenous program participation, which implies that these conditional distributions are different from the conditional ones:

$$\ell(y_{1_i}|\mathcal{T}_i=1)\neq\ell(y_{1_i}|\mathcal{T}_i=0)\neq\ell(y_{1_i}),\ \ell(y_{0_i}|\mathcal{T}_i=0)\neq\ell(y_{0_i}|\mathcal{T}_i=1)\neq\ell(y_{0_i})$$

in such a way that the distribution of the outcome, both before and after the implementation of the treatment, is correlated with the treatment status. The main consequence is that the distribution of the outcome among untreated cannot be used as a counterfactual for the outcome among treated individuals. The observed sample provides information on the following parameters:

$$\begin{aligned}\mathbb{E}(y_i|\mathcal{T}_i=0) &= \mathbb{E}(y_{0_i}|\mathcal{T}_i=0)\\ \mathbb{E}(y_i|\mathcal{T}_i=1) &= \mathbb{E}(y_{1_i}|\mathcal{T}_i=1)=\mathbb{E}(y_{0_i}|\mathcal{T}_i=0)+\mathcal{T}_i\left[\mathbb{E}(y_{1_i}|\mathcal{T}_i=1)-\mathbb{E}(y_{0_i}|\mathcal{T}_i=1)\right]\end{aligned}$$

The naive (or cross-section) estimator, which relies on this counterfactual, $\widehat{\Delta}^{\text{C}}=\overline{y}^{\mathcal{T}_i=1}-\overline{y}^{\mathcal{T}_i=0}$ is linked to the distribution of the potential outcomes since $\widehat{\Delta}^{\text{C}}\xrightarrow{\mathbb{P}}\Delta^{\text{C}}=\mathbb{E}(y_i|\mathcal{T}_i=1)-\mathbb{E}(y_i|\mathcal{T}_i=0)$, where:

$$\Delta^{\text{C}}=\mathbb{E}(y_{1_i}|\mathcal{T}_i=1)-\mathbb{E}(y_{0_i}|\mathcal{T}_i=1)+\mathbb{E}(y_{0_i}|\mathcal{T}_i=1)-\mathbb{E}(y_{0_i}|\mathcal{T}_i=0)=\Delta^{\text{ATT}}+\mathcal{B}^{\text{C}} \tag{1}$$

The estimator thus measures a mix between the ATT and $\mathcal{B}^{\text{C}}=\mathbb{E}(y_{0_i}|\mathcal{T}_i=1)-\mathbb{E}(y_{0_i}|\mathcal{T}_i=0)$, the selection bias. This bias can be due to either self-selection into the treatment (if individuals who expect to benefit the most from it decide to participate, namely to be treated) or to the implementation of the treatment itself, if the policy aims to target a specific sub-population. Estimating the ATE is even more demanding, as:

$$\begin{aligned}\Delta^{\text{C}} &= \mathbb{E}(y_{1_i})-\mathbb{E}(y_{0_i}) &+&\ \left[1-\mathbb{P}(\mathcal{T}_i=1)\right]\left[\mathbb{E}(y_{1_i}|\mathcal{T}_i=1)-\mathbb{E}(y_{1_i}|\mathcal{T}_i=0)\right]\\ &&+&\ \mathbb{P}(\mathcal{T}_i=1)\left[\mathbb{E}(y_{0_i}|\mathcal{T}_i=1)-\mathbb{E}(y_{0_i}|\mathcal{T}_i=0)\right]\\ &= \Delta^{\text{ATE}} &+&\ \mathcal{B}^{\text{C}_{\text{ATE}}}\end{aligned}$$

from which another source of bias arises:

$$\mathcal{B}^{\text{C}_{\text{ATE}}}=\mathcal{B}^{\text{C}}+\left[1-\mathbb{P}(\mathcal{T}_i=1)\right]\left[\mathbb{E}(y_{1_i}-y_{0_i}|\mathcal{T}_i=1)-\mathbb{E}(y_{1_i}-y_{0_i}|\mathcal{T}_i=0)\right]$$

On top of the selection bias, $\mathcal{B}^{\text{C}}$, the evaluation must now take into account the heterogeneity in the response to the treatment.

3.2 Identification through randomization

Randomization of individuals into the treatment allows to break the link between the outcome variable and the treatment status giving rise to the selection bias. If the value of $\mathcal{T}_i$, $\forall i$, is decided randomly, then $\mathbb{E}(y_{0_i}|\mathcal{T}_i=1)=\mathbb{E}(y_{0_i}|\mathcal{T}_i=0)$, and hence:

$$\mathbb{E}(y_{1_i}|\mathcal{T}_i=1)-\mathbb{E}(y_{0_i}|\mathcal{T}_i=0)=\mathbb{E}(y_{1_i}|\mathcal{T}_i=1)-\mathbb{E}(y_{0_i}|\mathcal{T}_i=1)=\mathbb{E}(y_{1_i}-y_{0_i}|\mathcal{T}_i=1)$$

But randomization also breaks the correlation between the treatment status and the outcome from the treatment, $\mathbb{E}(y_{1_i} - y_{0_i} | \mathcal{T}_i = 1) = \mathbb{E}(y_{1_i} - y_{0_i} | \mathcal{T}_i = 0) = \mathbb{E}(y_{1_i} - y_{0_i})$, in a such a way that: $\mathbb{E}(y_{1_i} | \mathcal{T}_i = 1) - \mathbb{E}(y_{0_i} | \mathcal{T}_i = 0) = \Delta^{\text{ATT}} = \Delta^{\text{ATE}}$. Randomized experiments thus 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. They cannot deliver the ATT unless some of the selected individuals refuse to take part in the experiment (in which case the treatment effect is conditional on endogenous refusal, which might select a subpopulation which is different from the one generated by endogenous participation), or if randomization is performed within a well-chosen sub-population.

3.3 Difference estimators

Several difference estimators overcome the missing data issue by relying on different counterfactuals. They are typically applied to data produced by the implementation of a policy, which is treated as a "natural experiment". In this context, individuals are observed at different points in time and $\mathcal{T}_{it} \in \{0, 1\}$ depending on whether i receives or does not receive the treatment in t. Time periods before and after the implementation of the policy in t_0, respectively denoted $\bar{t} = \{t > t_0\}$ and $\underline{t} = \{t < t_0\}$, define the dummy variable $\mathcal{P}_{it} = \mathbb{1}_{\{t>t_0\}}$ such that $\mathcal{P}_{i\bar{t}} = 1 - \mathcal{P}_{i\underline{t}} = 1$. In this context, two pairs of potential outcomes must be considered, depending on the date at which the outcome is observed:

$$y_{it} = \begin{cases} y_{0_{i\underline{t}}} & if \quad \mathcal{P}_{it} = 0, \mathcal{T}_{i\bar{t}} = 0 \\ y_{0_{i\bar{t}}} & if \quad \mathcal{P}_{it} = 1, \mathcal{T}_{i\bar{t}} = 0 \\ y_{0_{i\underline{t}}} & if \quad \mathcal{P}_{it} = 0, \mathcal{T}_{i\bar{t}} = 1 \\ y_{1_{i\bar{t}}} & if \quad \mathcal{P}_{it} = 1, \mathcal{T}_{i\bar{t}} = 1 \end{cases}, \forall i, t$$

3.3.1 The cross-section estimator

The cross-section estimator only makes use of post-implementation data, and compares treated individuals to untreated ones. Formally, it is similar to the naive estimator defined above (Section 3.1), and may be written as:

$$\widehat{\Delta}^{\text{C}} = \overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=1, \mathcal{P}_{it}=1} - \overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=0, \mathcal{P}_{it}=1} \xrightarrow{\mathbb{P}} \Delta^{\text{C}} = \mathbb{E}\left(y_{it} | \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 1\right) - \mathbb{E}\left(y_{it} | \mathcal{T}_{i\bar{t}} = 0, \mathcal{P}_{it} = 1\right)$$

The estimator is biased because of selection into the treatment since the resulting outcome can be decomposed as:

$$\Delta^{\text{C}} = \Delta^{\text{ATT}} + \mathbb{E}\left(y_{0_{it}} | \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 1\right) - \mathbb{E}\left(y_{0_{it}} | \mathcal{T}_{i\bar{t}} = 0, \mathcal{P}_{it} = 1\right) = \Delta^{\text{ATT}} + \mathcal{B}^{\text{C}}\left(\mathcal{P}_{it} = 1\right)$$

but does not depend on selection on the outcome in the absence of the treatment, $\mathcal{B}^{\text{C}}\left(\mathcal{P}_{it} = 0\right)$.

3.3.2 The Before-After (BA) estimator

The BA estimator focuses on treated individuals ($\forall i, \mathcal{T}_{i\bar{t}} = 1$); it uses their outcome before the implementation of the treatment to define a counterfactual of their post-treatment situation:

$$\widehat{\Delta}^{\text{BA}} = \overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=1, \mathcal{P}_{it}=1} - \overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=1, \mathcal{P}_{it}=0} \xrightarrow{\mathbb{P}} \Delta^{\text{BA}} = \mathbb{E}\left(y_{it} | \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 1\right) - \mathbb{E}\left(y_{it} | \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 0\right)$$

It can be written as the OLS estimator of Δ^{BA} in the linear model:

$$y_{it} = b_0 + \Delta^{\text{BA}} \mathcal{P}_{it} + u_{it},\ \forall t,\ \forall i : \mathcal{T}_{i\bar{t}} = 1$$

Given the observation rule that applies to y_{it}, the parameter may be written as:

$$\Delta^{\text{BA}} = \Delta^{\text{ATT}} + \mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 1\right.\right) - \mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 0\right.\right)$$

It thus measures a combination of the ATT and:

$$\mathcal{B}^{\text{BA}}\left(\mathcal{T}_{i\bar{t}} = 1\right) = \mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 1\right.\right) - \mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 0\right.\right) = \mathbb{E}\left(y_{0_{i\bar{t}}} - y_{0_{i\underline{t}}} \left| \mathcal{T}_{i\bar{t}} = 1\right.\right)$$

which measures the change in the outcome that would have happened over time even without the implementation of the treatment — this corresponds to a simultaneity bias. It thus relies on the identifying assumption that the outcome only changes as a result of the treatment. It does not depend on whether the same assumption applies to untreated individuals, $\mathcal{B}^{\text{BA}}\left(\mathcal{T}_{i\bar{t}} = 0\right)$.

3.3.3 The difference-in-differences (DD) estimator

The DD estimator makes use of the entire data produced by $\mathcal{P}_{it}$ and $\mathcal{T}_{it}$. It is based on the change over time in the untreated individuals' situation as a conterfactual for the change observed in the treated group. It thus amounts to compute the difference between two difference estimators: the BA computed in both groups generated by $\mathcal{T}_{i\bar{t}}$, or, equivalently, the cross-section estimators at both time periods defined by $\mathcal{P}_{it}$.

$$\begin{aligned}
\widehat{\Delta}^{\text{DD}} &= \left[\overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=1,\mathcal{P}_{it}=1} - \overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=1,\mathcal{P}_{it}=0}\right] - \left[\overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=0,\mathcal{P}_{it}=1} - \overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=0,\mathcal{P}_{it}=0}\right] = \widehat{\Delta}^{\text{BA}}_{\mathcal{T}_{i\bar{t}}=1} - \widehat{\Delta}^{\text{BA}}_{\mathcal{T}_{i\underline{t}}=0} \\
&= \left[\overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=1,\mathcal{P}_{it}=1} - \overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=0,\mathcal{P}_{it}=1}\right] - \left[\overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=1,\mathcal{P}_{it}=0} - \overline{y}_{it}^{\mathcal{T}_{i\bar{t}}=0,\mathcal{P}_{it}=0}\right] = \widehat{\Delta}^{\text{C}}_{\bar{t}} - \widehat{\Delta}^{\text{C}}_{\underline{t}}
\end{aligned}$$

Thanks to this double differentiation, the DD estimator gets rid of the biases affecting each kind of simple-difference estimator provided the bias remains constant in the dimension over which the differencing is performed. The estimator measures:

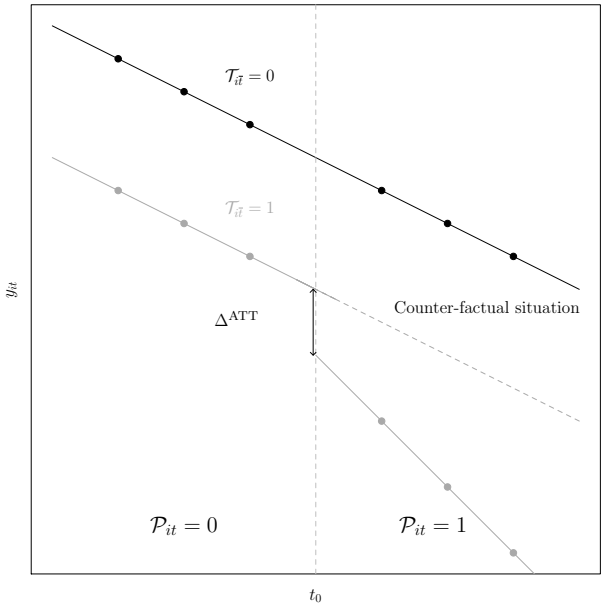
$$\begin{aligned}
\widehat{\Delta}^{\text{DD}} \xrightarrow{\mathbb{P}} \Delta^{\text{DD}} = \Delta^{\text{ATE}} &+ \left[\mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 1\right.\right) - \mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 0\right.\right)\right] \\
&- \left[\mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 0, \mathcal{P}_{it} = 1\right.\right) - \mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 0, \mathcal{P}_{it} = 0\right.\right)\right] = \Delta^{\text{ATE}} + \mathcal{B}^{\text{DD}}
\end{aligned}$$

where the bias affecting the DD estimator, $\mathcal{B}^{\text{DD}} = \mathbb{E}\left(y_{0_{i\bar{t}}} - y_{0_{i\underline{t}}} \left| \mathcal{T}_{i\bar{t}} = 1\right.\right) - \mathbb{E}\left(y_{0_{i\bar{t}}} - y_{0_{i\underline{t}}} \left| \mathcal{T}_{i\bar{t}} = 0\right.\right)$, measures the difference in trends between the control and treatment group, s.t. $\mathcal{B}^{\text{BA}}\left(\mathcal{T}_{i\bar{t}} = 1\right) \neq \mathcal{B}^{\text{BA}}\left(\mathcal{T}_{i\bar{t}} = 0\right)$. This bias can be equivalently interpreted as the change over time in the selection bias: $\mathcal{B}^{\text{C}}\left(\mathcal{P}_{it} = 1\right) \neq \mathcal{B}^{\text{C}}\left(\mathcal{P}_{it} = 0\right)$. The identification property of the estimator thus relies on the parallel trends assumption:

$$\begin{aligned}
\left[\mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 0, \mathcal{P}_{it} = 1\right.\right)\right. &- \left.\mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 0, \mathcal{P}_{it} = 0\right.\right)\right] \\
&= \left[\mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 1\right.\right) - \mathbb{E}\left(y_{0_{it}} \left| \mathcal{T}_{i\bar{t}} = 1, \mathcal{P}_{it} = 0\right.\right)\right]
\end{aligned}$$

It is worth noting that this assumption does not require either the lack of selection into the treatment group — $\mathcal{B}^{\text{C}}\left(\mathcal{P}_{it} = 1\right) \neq 0$ — or the lack of simultaneity bias — $\mathcal{B}^{\text{BA}}\left(\mathcal{T}_{i\bar{t}} = 1\right) \neq 0$. It only

Figure 2: The parallel trends assumption



Note. The figure illustrates the assumption that the trends in the treatment ($\mathcal{T}_i = 1$) and control ($\mathcal{T}_i = 0$) groups would be parallel over time ($\mathcal{P}_{it}$) with no treatment (dashed line). The double difference then estimate the ATT, Δ^{ATT}.

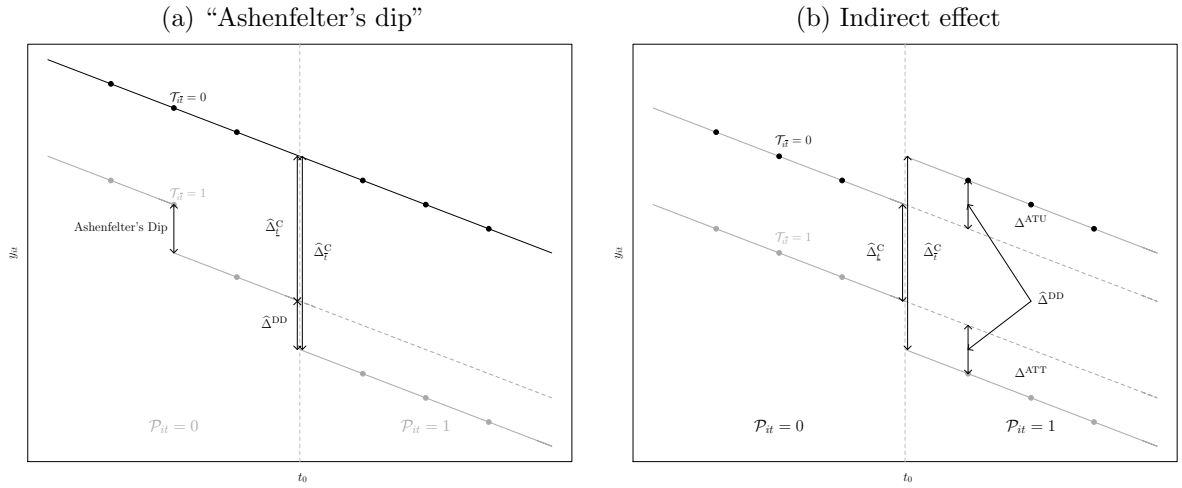
requires that such confounding variations are balanced over time — $\mathcal{B}^{\text{C}}(\mathcal{P}_{it} = 1) = \mathcal{B}^{\text{C}}(\mathcal{P}_{it} = 0)$ — or, equivalently, between groups — $\mathcal{B}^{\text{BA}}(\mathcal{T}_{i\bar{t}} = 1) = \mathcal{B}^{\text{BA}}(\mathcal{T}_{i\bar{t}} = 0)$.

As such, the DD estimator relies on the trend experienced by individuals from the control group as a counterfactual for the trend that individuals from the treatment group would have experienced without the intervention. Figure 2 provides a graphical illustration of this identifying assumption: the two parallel lines show the evolution of $y_{0_{it}}$ over time within each group (the dotted part on the right thus provides the counterfactual evolution of this quantity in the treatment group: it reproduces the trend observed in the control group). Under this assumption, the double difference measures the ATT of the intervention. Although this assumption is less demanding than the identifying assumptions of either the cross-section or the BA estimators, Figure 3 describes two typical sources of its failure. Figure 3.a illustrates the Ashenfelter (1978)'s "dip", which refers to a change in the outcome among treated individuals before the policy is implemented (for instance because the expected benefits from the intervention lead treated individuals to reduce the resources devoted to the outcome). In Figure 3.b, the failure is rather due to the behavior of individuals from the control group: due to an indirect effect of the policy, the average treatment on the untreated (Δ^{ATU}) is non-zero.

3.4 Conditional independence

Instead of eliminating confounding heterogeneity through differentiation, identification can also be achieved by finding covariates conditional on which the assignment can be assumed to be random. To expose this method, we consider a sample of N individuals for whom the pair $\{y_i, \mathcal{T}_i\}$

Figure 3: Well-known failures of the parallel trends assumption



Note. The figure displays two typical failures of the assumption that the trends in the treatment ($\mathcal{T}_i = 1$) and control ($\mathcal{T}_i = 0$) groups would be parallel over time ($\mathcal{P}_{it}$) with no treatment; resulting in biased difference-in-difference estimates, $\widehat{\Delta}^{\mathrm{DD}}$.

is observed. The model of interest is:

$$y_i = b_0 + b_1 \mathcal{T}_i + u_i, \, \forall i$$

in which unobserved heterogeneity is correlated to the treatment variable due to endogenous participation in the program. Control variables are observed covariates, $\boldsymbol{x}$, whose aim is to wash out the noise from confounding factors. Denoting $u_i = \boldsymbol{x}_i \boldsymbol{c} + v_i$, the model may now be written as:

$$y_i = b_0 + b_1 \mathcal{T}_i + \boldsymbol{x}_i \boldsymbol{c} + v_i$$

Those covariates are sufficient to achieve identification if they lead to the independence between the outcome variable and the treatment variable, conditional on the values of the control variables. Control variables thus achieve a quasi-experiment, with statistically identical individuals in both groups. This independence condition is weak if $\ell\left(y_{0_i} \left| \mathcal{T}_i, \boldsymbol{x}_i \right.\right) = \ell\left(y_{0_i} \left| \boldsymbol{x}_i \right.\right)$, in which case the output observed among untreated individuals provides a counterfactual for the situation of treated individuals: $\mathbb{E}\left(y_{0_i} \left| \mathcal{T}_i, \boldsymbol{x}_i \right.\right) = \mathbb{E}\left(y_{0_i} \left| \boldsymbol{x}_i \right.\right)$, which is sufficient to identify Δ^{ATT}. Independence is strong if this property also concerns the output induced by the treatment $\ell\left(y_{1_i}, y_{0_i} \left| \mathcal{T}_i, \boldsymbol{x}_i \right.\right) = \ell\left(y_{1_i}, y_{0_i} \left| \boldsymbol{x}_i \right.\right)$, which means that we also impose that $\mathbb{E}\left(y_{1_i} \left| \mathcal{T}_i, \boldsymbol{x}_i \right.\right) = \mathbb{E}\left(y_{1_i} \left| \boldsymbol{x}_i \right.\right)$, and the output comparison between groups thus identifies Δ^{ATE}.

3.4.1 Regression discontinuity

One of the most popular application of the identification based on conditional independence is regression discontinuity. Denote $\mathcal{D}_i$ the observed variable for which the threshold rule applies, and $\overline{d}$ the threshold which determines participation in the program, $\mathcal{T}_i = \mathbb{1}\{\mathcal{D}_i > \overline{d}\}$. There is a "regression discontinuity" if there exists $\epsilon > 0$ such that:

$$\ell\left(y_{0_i}, y_{1_i} \left| \mathcal{D}_i = \overline{d} - \epsilon \right.\right) = \ell\left(y_{0_i}, y_{1_i} \left| \mathcal{D}_i = \overline{d} + \epsilon \right.\right) \Leftrightarrow \ell\left(y_{0_i}, y_{1_i} \left| \mathcal{T}_i, \mathcal{D}_i = \overline{d} \right.\right) = \ell\left(y_{0_i}, y_{1_i} \left| \mathcal{D}_i = \overline{d} \right.\right)$$

This rule thus produces a quasi-experiment in the neighborhood of the threshold $\overline{d}$, as individuals can be assumed to be randomly assigned to the treatment conditional on $\mathcal{D}_i \in [\overline{d}-\epsilon; \overline{d}+\epsilon]$. Thus the distance to the threshold $\mathcal{D}_i - \overline{d}$ is a control variable which generates conditional independence.

3.4.2 Matching estimators

In the more general case of a vector of control variables $\boldsymbol{x}_i$ generating conditional independence, the estimation of the causal parameter relies on finding for each treated individual i a counterfactual $\mathbb{E}(y_{0_i} | \mathcal{T}_i = 1, \boldsymbol{x}_i)$. The matching estimator (Rubin, 1977) implements this strategy by matching each treated individual, characterized by control variables $\boldsymbol{x}_i$, with a statistical twin belonging to the control group, namely an untreated individual $j(i)$ whose observed values of the control variables are the same, $j(i) \in \{j | \mathcal{T}_j = 0, \boldsymbol{x}_j = \boldsymbol{x}_i\}$. Under such circumstances, the outcome $y_{j(i)}$ fulfills the weak independence condition:

$$\mathbb{E}(y_i | \mathcal{T}_i = 0, \boldsymbol{x} = \boldsymbol{x}_i) = \mathbb{E}(y_{0_i} | \mathcal{T}_i = 0, \boldsymbol{x} = \boldsymbol{x}_i) = \mathbb{E}(y_{0_i} | \boldsymbol{x} = \boldsymbol{x}_i) = \mathbb{E}(y_{0_i} | \mathcal{T}_i = 1, \boldsymbol{x} = \boldsymbol{x}_i)$$

This allows to measure the individual change in the outcome induced by the treatment as $\widehat{\Delta}_i = y_i - y_{j(i)}$, which yields a consistent estimator for the ATT:

$$\widehat{\Delta}^{\text{ATT}} = \frac{1}{N_{\mathcal{T}_i=1}} \sum_{i \in \mathcal{T}_i = 1} \left(y_i - y_{j(i)}\right) \xrightarrow{\mathbb{P}} \mathbb{E}(\Delta_i | \mathcal{T}_i = 1) = \Delta^{\text{ATT}}$$

where $N_{\mathcal{T}_i=1}$ denotes the size of the treatment group. Strong independence allows to extend the matching strategy to individuals from the control group, $\mathcal{T}_i = 0$, with statistical twins $j(i) \in \{j | \mathcal{T}_j = 1, \boldsymbol{x}_j = \boldsymbol{x}_i\}$, so as to estimate the ATE we can calculate:

$$\widehat{\Delta}^{\text{ATE}} = \frac{1}{N} \left[\sum_{i \in \mathcal{T}_i = 1} \left(y_i - y_{j(i)}\right) + \sum_{i \in \mathcal{T}_i = 0} \left(y_{j(i)} - y_i\right) \right] \xrightarrow{\mathbb{P}} \mathbb{E}(\Delta_i) = \Delta^{\text{ATE}}$$

The conditional independence assumption required to apply this strategy only holds if control variables bring enough information to conditionally break the endogeneity of the treatment. This concern very often leads to include many control variables, which makes the perfect matching described above more and more challenging as the number of cells increases. An alternative is to define counterfactuals based on the closest neighbours, rather than finding a perfect statistical twin. The increase in the number of control variables also reduces the number of degrees of freedom of the model. This can be addressed by aggregating the control variables, which is the aim of the propensity score.

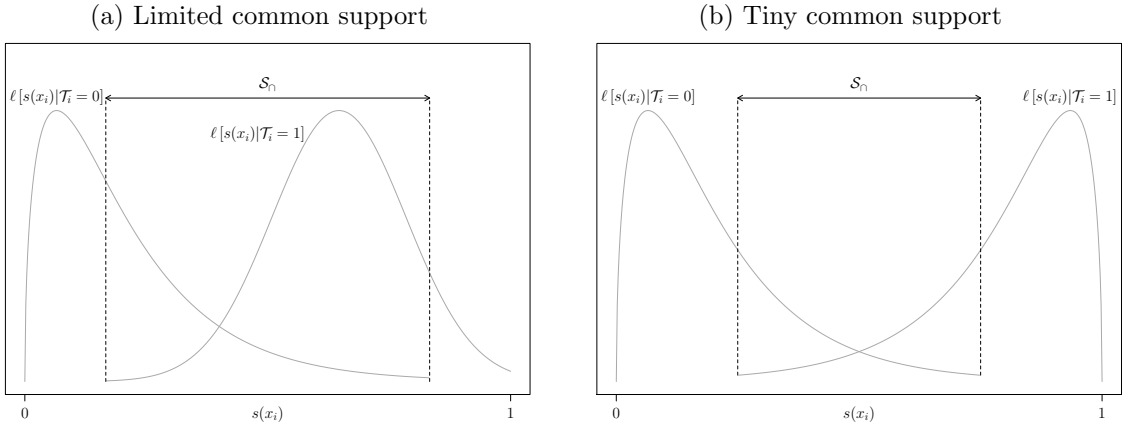
3.4.3 Aggregating control variables: the propensity score

The propensity score summarizes the information provided by the entire set of control variables through the change in the probability of being treated (Rosenbaum and Rubin, 1983): $s(\boldsymbol{x}_i) = \mathbb{P}(\mathcal{T}_i = 1 | \boldsymbol{x}_i)$. This score aims to balance the distribution of the control variables across the treatment and control groups.[3] This allows to compute a counterfactual for each treated individual

[3] This *balancing score* property implies that $\boldsymbol{x}_i \perp \mathcal{T}_i | s(\boldsymbol{x}_i)$. Given that $s(\boldsymbol{x}_i) = \ell(\mathcal{T}_i | \boldsymbol{x}_i)$, and hence: $\ell[\mathcal{T}_i | s(\boldsymbol{x}_i)] = \ell[\mathcal{T}_i | \boldsymbol{x}_i, s(\boldsymbol{x}_i)]$, conditional independence implies:

$$\ell[\boldsymbol{x}_i, \mathcal{T}_i | s(\boldsymbol{x}_i)] = \ell[\boldsymbol{x}_i | \mathcal{T}_i, s(\boldsymbol{x}_i)] \ell[\mathcal{T}_i | s(\boldsymbol{x}_i)] = \ell[\mathcal{T}_i | \boldsymbol{x}_i, s(\boldsymbol{x}_i)] \ell[\boldsymbol{x}_i | s(\boldsymbol{x}_i)]$$

Figure 4: The common support property of the propensity score



Note. The figure plots the distributions of the score, $s(\boldsymbol{x}_i)$, conditional on the value of the treatment variable, $\mathcal{T}_i$, in two cases with varying size of the common support, $\mathcal{S}_\cap$.

based on $\mathbb{E}\left[y_{j(i)} \left| s\left(\boldsymbol{x}_{j(i)}\right) = s\left(\boldsymbol{x}_i\right), \mathcal{T}_{j(i)} = 0\right.\right]$, and then to compute the ATT. By computing a counterfactual for untreated individuals, $\mathbb{E}\left[y_{j(i)} \left| s\left(\boldsymbol{x}_{j(i)}\right) = s\left(\boldsymbol{x}_i\right), \mathcal{T}_{j(i)} = 1\right.\right]$, we can obtain an estimator of the ATE.

Such counterfactuals can only be found if it is possible to observe for each individual whose score is $s\left(\boldsymbol{x}_i\right)$ a counterfactual individual (untreated if $\mathcal{T}_i = 1$, treated otherwise) with a similar score. This possibility depends on the extent to which the support of the two distributions $\ell\left[s\left(\boldsymbol{x}_i\right)|\mathcal{T}_i = 1\right]$ and $\ell\left[s\left(\boldsymbol{x}_i\right)|\mathcal{T}_i = 0\right]$ overlap: $\mathcal{S}_\cap = \ell\left[s\left(\boldsymbol{x}_i\right)|\mathcal{T}_i = 1\right] \cap \ell\left[s\left(\boldsymbol{x}_i\right)|\mathcal{T}_i = 0\right]$. This common support property has important consequences on the consistency of the estimators derived from this conditional independence assumption (see Heckman, Ichimura, and Todd, 1998). Identification is indeed restricted to the range of values of the score belonging to the common support: $\mathbb{E}\left[\Delta_i | s\left(\boldsymbol{x}_i\right), \mathcal{T}_i = 1\right] = \mathbb{E}\left[\Delta_i | s\left(\boldsymbol{x}_i\right) \in \mathcal{S}_\cap, \mathcal{T}_i = 1\right]$ and $\mathbb{E}\left[\Delta_i | s\left(\boldsymbol{x}_i\right)\right] = \mathbb{E}\left[\Delta_i | s\left(\boldsymbol{x}_i\right) \in \mathcal{S}_\cap\right]$. As illustrated in Figure 4, this condition is more likely to be restrictive when the explanatory power of the score is higher: on each graph, the score gives more probability to the conditioning outcome, as expected, but the common support is much thiner in Figure 4.b than in Figure 4.a due to the improved accuracy of the score.

3.5 Instrumental variable estimator

Instruments are exogenous observed variables which achieve identification thanks to their correlation with the treatment variable. Consider the most simple implementation, in which the binary treatment variable, $\mathcal{T}_i$, is instrumented by a binary instrument $z_i^e \in \{0, 1\}$, $\forall i$, in the linear model:

$$y_i = b_0 + b_1 \mathcal{T}_i + u_i, \ \mathbb{E}(\mathcal{T}_i u_i) \neq 0, \forall i$$

The variable z^e can be used as an instrument if it is exogenous, namely $\mathbb{E}(u_i|z_i^e) = 0$ (the "orthogonality condition"), in such a way that $\mathbb{E}(y_i|z_i^e) = b_0 + b_1\mathbb{E}(\mathcal{T}_i|z_i^e) + \mathbb{E}(u_i|z_i^e)$. Under

this assumption, the variations in the outcome resulting from changes in the instrument,

$$\begin{aligned}\mathbb{E}(y_i|z_i^e=1) &= b_0 + b_1\mathbb{E}(\mathcal{T}_i|z_i^e=1)\\ \mathbb{E}(y_i|z_i^e=0) &= b_0 + b_1\mathbb{E}(\mathcal{T}_i|z_i^e=0)\end{aligned}$$

identify the effect of the treatment:

$$b_1 = \frac{\mathbb{E}(y_i|z_i^e=1) - \mathbb{E}(y_i|z_i^e=0)}{\mathbb{E}(\mathcal{T}_i|z_i^e=1) - \mathbb{E}(\mathcal{T}_i|z_i^e=0)}$$

This relation only defines the parameter if changes in the instrument actually induce changes in the treatment, i.e. $\mathbb{E}(\mathcal{T}_i|z_i^e=1) - \mathbb{E}(\mathcal{T}_i|z_i^e=0) \neq 0$ (the "rank condition"). The closer this difference to 0, the weaker the instrument, leading to poor identification of the parameter. This target parameter can be estimated based on the Wald (1940) estimator (denoting $\overline{w}^{z_i^e=0/1} = \frac{1}{\sum_i \mathbb{1}_{\{z_i^e=0/1\}}}\sum_i w_i \times \mathbb{1}_{\{z_i^e=0/1\}}$, $w \equiv \{y, x\}$):

$$\widehat{\Delta}^{\text{WALD}} = \frac{\overline{y}^{z_i^e=1} - \overline{y}^{z_i^e=0}}{\overline{x}^{z_i^e=1} - \overline{x}^{z_i^e=0}} \xrightarrow{\mathbb{P}} \frac{\mathbb{E}(y_i|z_i^e=1) - \mathbb{E}(y_i|z_i^e=0)}{\mathbb{E}(\mathcal{T}_i|z_i^e=1) - \mathbb{E}(\mathcal{T}_i|z_i^e=0)} \tag{2}$$

3.5.1 The Local Average Treatment Effect (LATE)

To clarify the interpretation of the true parameter measured by the Wald estimator presented above, let us denote $\mathcal{T}_{1_i}$ and $\mathcal{T}_{0_i}$ the potential treatment status of individual i depending on the value of the instrument:

$$\mathcal{T}_i = \begin{cases} \mathcal{T}_{0_i} & \text{if } z_i^e = 0 \\ \mathcal{T}_{1_i} & \text{if } z_i^e = 1 \end{cases}$$

As shown in Table 1, this relationship between the treatment and the instrument defines four types of individuals. Two among them do not react to the instrument, i.e., $\mathcal{T}_{0_i} = \mathcal{T}_{1_i} = 0$ for *Never takers*, and $\mathcal{T}_{0_i} = \mathcal{T}_{1_i} = 1$ for *Always takers*, while the two other types do react to the instrument. The response is positive for *Compliers*, $\mathcal{T}_{1_i} > \mathcal{T}_{0_i}$, and negative for *Defiers*, $\mathcal{T}_{1_i} < \mathcal{T}_{0_i}$. The rank condition is meant to exclude the possibility that the sample is only made of the first two kinds of individuals. Defiers are often excluded by assumption by imposing monotonicity of the change in the treatment. The monotonicity assumption means that $\mathcal{T}_{1i} - \mathcal{T}_{0i} \geq 0$, $\forall i$. Under these assumptions, only compliers contribute to the identification of the effect of the treatment, and the Wald estimator measures the Local Average Treatment Effect (LATE):

$$\widehat{\Delta}^{\text{WALD}} \xrightarrow{\mathbb{P}} \frac{\mathbb{E}(y_i|z_i^e=1) - \mathbb{E}(y_i|z_i^e=0)}{\mathbb{E}(\mathcal{T}_i|z_i^e=1) - \mathbb{E}(\mathcal{T}_i|z_i^e=0)} = \mathbb{E}(y_{1_i} - y_{0_i}|\mathcal{T}_{1_i} - \mathcal{T}_{0_i} > 0) = \Delta^{\text{LATE}}$$

Since the work of Imbens and 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 and 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

Table 1: Response of the treatment status to a change in the (binary) instrumental variable

	$\mathcal{T}_{1_i} = 0$	$\mathcal{T}_{1_i} = 1$
$\mathcal{T}_{0_i} = 0$	Never taker	Complier
$\mathcal{T}_{0_i} = 1$	Defier	Always taker

Note. The table describes the four types of observations resulting from the combination of the potential treatment variables generated by a binary instrument, z^e: $\mathcal{T}_{0_i} = \mathcal{T}_i|_{z_i^e=0}; \mathcal{T}_{1_i} = \mathcal{T}_i|_{z_i^e=1}$.

from the program when it becomes available as a result of a variation of the instrument. Those individuals are called "compliers", i.e., they are people who comply or adhere to the program 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 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 and 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, Graddy, and Imbens (2000) apply this approach to the estimation of simultaneous equation models.

4 The External Validity of Impact Evaluation Methods

Several of the methods cited above are characterized 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, Imbens, and Mortimer (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 recommended by Stuart and co-authors (Imai, King, and Stuart, 2008; Stuart, Cole, Bradshaw, and Leaf, 2011; Stuart, Bradshaw, and Leaf, 2015).[4]

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? There are two groups that 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 and Fernandez-Val (2013) seek to exploit a conditional effect ignorability assumption stipulating that, conditional on certain exogenous variables, the average effect for compliers is identical to the average effect for always-takers and never-takers. Bertanha and 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 and 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 thresh-

[4]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.

old. Angrist and 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)[5] 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 and Imbens (2019) propose an approach based on the fuzzy regression discontinuity design.[6] 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.

5 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 and 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 and 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 $y_{1,t}$, $\forall t = 1960, \ldots, 1969$, 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 $\boldsymbol{y}_{j,t}$, $\forall j = 2, \ldots, 17, t = 1960, \ldots, 1969$, and stored in a (10×16) matrix denoted $\boldsymbol{Y}_{0,0}$. The synthetic control region is constructed from a (16×1) weighting vector $\boldsymbol{w} = [w_1, \ldots, w_{16}]'$, which minimizes the following weighted Euclidean norm for

[5] "The unconfoundedness assumption states that assignment is free from dependence on the potential outcomes" (Imbens and Rubin, 2015, p. 257).

[6] 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.

a given matrix $\boldsymbol{V}$:

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

In a first simple application, Abadie and 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.[7] They verify that the ten annual GDPs of that synthetic region, which are calculated as $\boldsymbol{Y}^*_{0,0} = \boldsymbol{Y}_{0,0}\boldsymbol{w}^*$ 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 (26×1) vector $\boldsymbol{y}^*_{0,1} = \boldsymbol{Y}_{0,1}\boldsymbol{w}^*$, where $\boldsymbol{Y}_{0,1}$ is the (26×16) matrix which groups together the observations concerning the 26 (= 2000–1974) annual GDPs 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 (26×1) vector 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 and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (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 and l'Hour (2017) propose a variant that takes this difficulty into account. Their estimator is written:

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

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, $\forall j = 2, \ldots, J+1$, during each of the periods preceding the implementation of the intervention. The estimator proposed by Abadie and l'Hour (2017) 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 and l'Hour (2017) 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, Shah, and Shen (2018) and Athey, Bayati, Doudchenko, Imbens, and Khosravi (2018), who suggest the use of matrix completion techniques, but also by Hahn and Shi (2017), who base their approach on sampling-based inferential methods.

6 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,

[7] The only regions with weights well above zero are Madrid and Catalonia.

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 accuracy of the estimator of the average treatment effect.

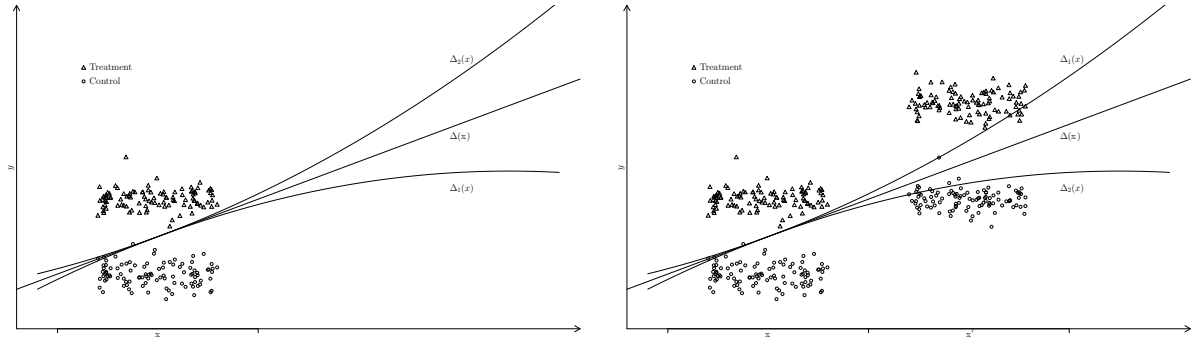
However, in cases other than non-stratified randomization or sharp regression discontinuity design, it is necessary to add assumptions about the role of exogenous variables in order to obtain consistent estimators. The most commonly used assumption is that of conditional independence. This assumption states that the assignment to the treatment group, represented by a random variable $\mathcal{T}$, and the potential outcomes of the intervention, denoted y_{i_1} for a treated individual and y_{i_0} 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 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.[8] 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 ("*balancing score property*"). 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, Ridgeway, and Morral, 2004; Wyss, Ellis, Brookhart, Girman, Funk, LoCasale, and Stürmer, 2014; Athey and Imbens, 2017b; Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018), and LASSO[9] methods (Belloni, Chernozhukov, and Hansen, 2014; Belloni, Chernozhukov, Fernández-Val, and Hansen, 2017; Farrell, 2015). For example, McCaffrey, Ridgeway, and Morral (2004), like Wyss, Ellis, Brookhart, Girman, Funk,

[8] 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.

[9] 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 residual under an additional regularisation constraint.

Figure 5: Empirical identification and generalizability



Note. The figure illustrates the empirical identification of the effect of a treatment using an exogenous variable x with low ($x \in \mathbb{x}$, left-hand side) and high ($x \in \{\mathbb{x} \cup \mathbb{x}'\}$, right-hand side) dispersion.

LoCasale, and Stürmer (2014), combine the method of random forests[10] 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, Chernozhukov, Fernández-Val, and Hansen (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.

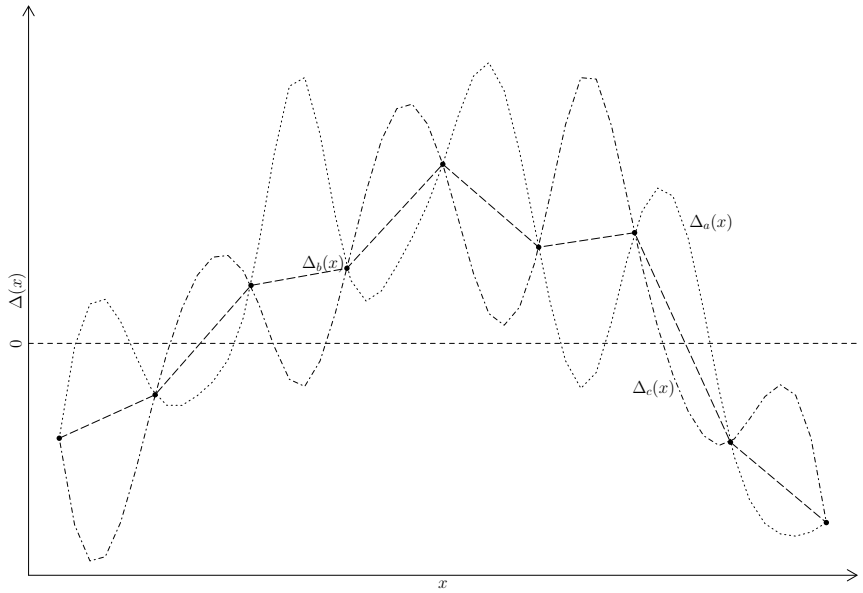
7 The Heterogeneity of the Effects of an Intervention

Recent work has often focused on the heterogeneity of the effects of an intervention between groups of eligible individuals. Figure 5 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. Figure 5.a 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 (denoted $\mathbb{x}$). 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, $\Delta_1(x)$ and $\Delta_2(x)$ are drawn on the figure). Figure 5.b 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 to be measured.

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

[10] To implement this technique, the reader can in particular use the R package randomForest (`https://cran.r?project.org/web/packages/ran?domForest/index.html`).

Figure 6: (Non-)Identification of the full distribution of the treatment effect based on (any) estimates



Note. The figure plots a large variety of point estimates of the treatment effect, each associated with a different value of the variable x, along with 3 examples of the many relationships between the treatment and the covariate that are consistent with this data.

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 6 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 6 presents three examples:$\Delta_a(x), \Delta_b(x)$ and $\Delta_c(x)$. 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 modeling 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 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 and Athey, 2018; Knaus, Lechner, and Strittmatter, 2017, 2018). This statistical approach can be generalised on a case-by-case basis with several treatments (Lechner, 2019).

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 Bisbee, Dehejia, Pop-Eleches, and Samii (2017) and Dehejia, Pop-Eleches, and Samii (2019), 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 10 sites where the scheme was initially applied, suggesting that these first sites were selected because of their particular characteristics.

8 Accuracy 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 and Ziliak, 1996; Ziliak and 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, Greenland, and McShane, 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 minimizing 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 Deaton and Cartwright, 2018). 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[11].

[11] 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

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 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 $\widehat{s}$ and the null hypothesis is denoted H_0, the p-value therefore formally measures the quantity $\mathbb{P}(\widehat{s}|H_0)$. The probability that the null hypothesis is true corresponds to the reverse conditioning, $\mathbb{P}(H_0|\widehat{s})$. The confusion between these two probabilities can be illustrated by what the behavioural science literature calls the "prosecutor fallacy", introduced by Thompson and 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.

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

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, $1-\beta$. 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.[12]

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, Chanock, Garcia-Closas, El ghormli, and Rothman (2004)

groups.

[12]The benchmark power level in applied work is 80%, although Ioannidis, Stanley, and Doucouliagos (2017) show that in more than half of applied economics work, the median power is 18% or even less.

Table 2: Components of the Probability of Occurrence of a False Positive

Truth of the alternative hypothesis	Significance of test		Total
	Significant	Not significant	
True association	$(1-\beta)\bar{y}$ [True positive]	$\beta\bar{y}$ [False negative]	$\bar{y}$
No association	$\alpha(1-\bar{y})$ [False positive]	$(1-\alpha)(1-\bar{y})$ [True negative]	$1-\bar{y}$
Total	$(1-\beta)\bar{y}+\alpha(1-\bar{y})$	$\beta\bar{y}+(1-\alpha)(1-\bar{y})$	$\bar{y}$

Note. 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 $\overline{y}$ of interventions that have a non-zero effect amongst all those evaluated. Sources: Wacholder, Chanock, Garcia-Closas, El ghormli, and Rothman (2004, p.440), Table 1.

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 $\overline{y}$, of interventions that have a non-zero effect amongst all the interventions that are being evaluated. Table 2 describes the probability of occurrence of the four types of possible situations: the legitimate detection of an absence (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:

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

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 lower 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 $\overline{y}$, and very difficult to collect objective information on this proportion. Still, the consequences of low values of $\overline{y}$ on the credibility that results from evaluations deserve may be very serious: under the extreme hypothesis that one intervention out of 1,000 has a non-zero effect ($\overline{y} = 0.001$), 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.[13] Ioannidis (2005) focuses in particular on two types of biases that increase the probability of reporting false positives, namely the publication bias and the 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 $\mathcal{B}$. Amongst the $\overline{y}$ 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 $(\mathcal{B} \times \beta)$ will appear to have an effect due to one of the types of biases. Similarly, a proportion α of interventions amongst the $(1-\overline{y})$ actually having zero effect will appear as having no effect, while a certain number $\mathcal{B} \times (1-\alpha)$ will appear as having a non-zero effect due to bias. In total, the FPRP becomes:

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

For the "credibility revolution" announced by some authors (Angrist and Pischke, 2010) to fully succeed, 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.

10 Interference between units in experimental and observational studies

Usually, the post-treatment outcome of one unit (either treated or control) is assumed to be unaffected by the potential outcomes of other units. This assumption of no interference, introduced

[13] 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, Bailey, Euzent, and Martin, 2001). Christensen and 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.

by Cox (1958), is called the SUTVA assumption (Stable Unit Treatment Value Assumption) by Rubin (1978) or the individualistic treatment response by Manski (2013). In the presence of interference, treatments may have direct effects on the units (individuals, schools, neighborhoods, cities, firms, etc.) receiving the treatment as well as indirect effects on units not receiving the treatment. In other terms, each unit may have several potential outcomes depending on the treatment status, the outcomes and the behaviors of other units. During the last fifteen years, a burgeoning academic literature has been devoted to measuring the effects of interference between units in experimental and observational studies. Among this body of research, we may distinguish works relative to partial interference and those concerning general interference. This section gives a brief overview of these developments.

It is previously useful to introduce some concepts. Following Halloran and Struchiner (1991, 1995), we can define direct, indirect, total and overall effects in the presence of interference in RCTs. To illustrate these definitions, let us consider two clusters, denoted A and B. In cluster A, a certain proportion of randomly chosen units is treated while the rest remains untreated. In cluster B, all units are untreated. The direct effect of the treatment is defined as the difference between the average outcomes of treated units and untreated ones in cluster A. The indirect effect is the difference between the average outcome of untreated units in cluster A and the average outcome of untreated units in cluster B. The total effect is defined as the difference between the average outcome of treated units in cluster A and the average outcome of untreated units in cluster B. In general, the total effect can be decomposed as a function of direct and indirect effects. The overall effect is defined by the contrast in the average outcome in the entire cluster A compared to the average outcome of the entire cluster B.

10.1 Partial interference

In the presence of interference, Halloran and Struchiner (1995) propose individual-level causal estimands by letting the potential outcomes for any individual depend on the vector of treatment assignments to other individuals. However, with a binary outcome, one treatment and one control, if there are N people in a population, there are 2^N possible treatment vectors and 2^N possible potential outcomes. Causal inference in the presence of so many potential outcomes is difficult, if not impossible, without making additional assumptions.

It is then possible to add some further assumptions which simplify the problem substantially. For instance, units can be partitioned into groups such as there is no interference between groups. This assumption is called partial interference by Sobel (2006) who considers interference in a housing mobility experiment in poor neighborhoods.

10.1.1 Two-stage randomization

Suppose that we consider two different treatment strategies. The first strategy might be to treat 50% of the units in a population, the other strategy might be to treat no one. What are the direct, indirect, total and overall effects of the 50% treatment strategy compared to the no treatment strategy? Assuming partial interference, Hudgens and Halloran (2008) define group- and population-level causal estimands for direct, indirect, total, and overall causal effects of treatment

under two different treatment allocations. To obtain unbiased estimators of the population-level causal estimands, Hudgens and Halloran (2008) propose a two-stage randomization scheme, the first stage at the group level, the second stage at the unit level within groups. For example, suppose there are 10 groups of units. At the first stage, we could randomize 5 groups to the 50% treatment strategy and the remaining groups to the no treatment strategy. In the second stage, within 5 groups, 50% of units are randomly assigned to the treatment, and in the other 5 groups no unit receives the treatment. Unbiased estimates of the direct, indirect, and total effects can be obtained by contrasts in average outcomes among treated and untreated units under the different strategies. Likewise, contrasts in average outcomes of all units under the two treatment strategies produce unbiased estimates of the overall effect.

Hudgens and Halloran (2008) propose variance estimators under an additional assumption called stratified interference, stipulating that the indirect treatment effects depend only on the proportion of other individuals in the group which receive treatment. More recently, Tchetgen and VanderWeele (2012), Liu and Hudgens (2014), and Rigdon and Hudgens (2015) calculate exact asymptotic confidence intervals.

In order to measure spillover effects in the context of economic experiments, Baird, Bohren, McIntosh, and Özler (2018) consider a similar two-stage randomized design. They refer to the level of (treatment) coverage in a cluster as the saturation level and their study design as the randomized saturation design. The intention to treat effect, the spillover on the non-treated effect, and the total causal effect, are analogous to the total, indirect, and overall effects defined above. Moreover, consider optimal design of two stage randomized trials, particularly choosing the optimal coverage level, to detect the different effects.

10.1.2 Partially randomized and observational studies

In studies which do not utilize two-stage randomization, the estimators described in the previous subsection are generally biased or inconsistent due to potential confounding. In the observational approach where the treatment assignment mechanism is not known (or not random) and there is no interference, the propensity score is one method to adjust for confounding (see section above). Propensity score methods have then been extended to the case where interference may be present. For example, Hong and Raudenbush (2006) consider the effect on reading scores of being retained in kindergarten versus being promoted to the first grade. They classify schools by whether they retain a high proportion or a low proportion of children. Interference within a school is summarized by the dichotomous dummy variable representing a high or low retention rate. Their study is observational at two levels: schools are not randomized to have high or low retention, and students at risk to be retained are not randomized to be retained. Assuming partial interference, Hong and Raudenbush (2006) use a multilevel propensity score stratification allowing for interference. The school propensity of adopting a high retention rate is estimated thanks to pretreatment information. Then the child propensity of repetition in high-retention schools and the child propensity of repetition in low-retention schools are also estimated by using pretreatment information. Estimation of causal effects is based on stratifications by the estimated school and child propensity scores.

Tchetgen and VanderWeele (2012) also use group-level propensity scores to propose inverse probability weighted (IPW) estimators of the direct, indirect, total and overall causal effects in observational studies in the absence of two-stage randomization. Lundin and Karlsson (2014) develop similar IPW estimators of direct, indirect, and total effects under interference where treatment assignment is randomized only at the first stage, while in some groups, all units are untreated. Perez-Heydrich, Hudgens, Halloran, Clemens, Ali, and Emch (2014), and Liu, Hudgens, and Becker-Dreps (2016) consider the asymptotic properties of different IPW estimators in the presence of partial interference.

10.2 General interference

In the general interference setting, each unit is assumed to have a unique set of other units whose treatment might affect the outcome of this unit. Research assessing treatment effects in the presence of general interference include papers by Rosenbaum and Rubin (1983), Toulis and Kao (2013), Ugander, Karrer, Backstrom, and Kleinberg (2013), Eckles, Karrer, and Ugander (2017), van der Laan (2014), Liu, Hudgens, and Becker-Dreps (2016), and Aronow, Samii, et al. (2017). These papers typically assume one finite population of N units. For each unit, there exists a set of other units which may interfere with that unit. This interference set is assumed to be known and fixed. It is also usually assumed that the number of units in these interference sets is smaller than N and that any indirect or spillover effect on a unit sets from this known interference set.

10.2.1 Randomized experiments

For instance, Ugander, Karrer, Backstrom, and Kleinberg (2013), and Eckles, Karrer, and Ugander (2017) consider randomized experiments on networks where treatment of one unit could have indirect/spillover effects on neighboring units. Their causal estimand of interest is the average of the difference in the potential outcomes under two extreme assignments, one where every unit in a network receives treatment and one where no units receive treatment. If units were independently randomized to treatment and control, for units which share many connections with other units, the probability that all their neighbors would *(i)* all receive treatment or *(ii)* all receive control would be low. Thus, Ugander, Karrer, Backstrom, and Kleinberg (2013) consider a partition of the network into clusters of units and propose randomizing some of the clusters to receive treatment and the remaining clusters to receive control. Then, they introduce the notion of network exposure wherein a unit is network exposed to treatment (control) if the unit's response under a specific assignment vector is the same as if every unit in the network had received the treatment (control). They derive an unbiased estimator of the average treatment effect using inverse probability weighting for any randomization design for which the network exposure probabilities can be explicitly computed.

Toulis and Kao (2013) propose causal estimands for peer influence (indirect) effects describing interference in a social network. For each unit, a neighborhood is defined by the other units with whom that unit shares some connections. If at least one neighbor receives treatment, then the unit is considered exposed to peer influence effects. The potential outcome for each unit can depend on its treatment and that of its neighbors. Then Toulis and Kao (2013) define two causal

estimands. The causal estimand for the primary effect is the average over the whole population of the difference in outcomes if a unit receives treatment versus receives control when every other unit in the neighborhood receives control. The main causal estimands for peer influence effects are defined by fixing the specific number of neighbors who receive treatment. For example, if k neighbors receive treatment, the $k-$level causal estimand for peer-influence effects is averaged over units with at least k neighbors. Two estimation procedures are proposed: a frequentist model-based estimator assuming some sequential randomization design and known network, and a Bayesian approach which accounts for uncertainty in the network topology.

10.2.2 Observational studies

In recent contributions, van der Laan (2014), and Sofrygin and van der Laan (2017) have considered statistical inference about causal effects in the presence of general interference in observational studies. They define the population of interest to be a set of (possibly) dependent units and assume that only a single draw from the true data generating process is observed. Namely, contrary to traditional statistical inference, multiple independent and identically distributed (hereafter, i.i.d.) replicates are not assumed. With partial interference, one may assume that the groups are i.i.d., which allows us to apply existing statistical theory. However, with general interference, the observation of i.i.d. replicates is generally not possible, such that standard large sample frequentist approaches do not apply. Thus, van der Laan (2014), and Sofrygin and van der Laan (2017) derive asymptotic properties of targeted maximum likelihood estimators in this setting, providing a method for statistical inference in the presence of general interference in observation studies.

References

Abadie, A., and M. D. Cattaneo (2018): "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503.

Abadie, A., A. Diamond, and J. Hainmueller (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.

——— (2015): "Comparative Politics and the Synthetic Control Method," *American Journal of Political Science*, 59(2), 495–510.

Abadie, A., and J. Gardeazabal (2003): "The Economic Costs of Conflict: A Case Study of the Basque Country," *American Economic Review*, 93(1), 113–132.

Abadie, A., and J. l'Hour (2017): "A Penalized Synthetic Control Estimator for Disaggregated Data," *MIT Working Paper*.

Abbring, J. H., and J. J. Heckman (2007): "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. 6, Part B, chap. 72, pp. 5145–5303. Elsevier.

Allcott, H. (2015): "Site Selection Bias in Program Evaluation," *Quarterly Journal of Economics*, 130(3), 1117–1165.

Amjad, M., D. Shah, and D. Shen (2018): "Robust Synthetic Control," *The Journal of Machine Learning Research*, 19(1), 802–852.

Amrhein, V., S. Greenland, and B. McShane (2019): "Scientists Rise up against Statistical Significance," *Nature*, 567, 305–307.

Angrist, J., and G. Imbens (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.

Angrist, J. D. (2004): "Treatment Effect Heterogeneity in Theory and Practice," *Economic Journal*, 114(494), 52–83.

Angrist, J. D., and I. Fernandez-Val (2013): "Extrapolate-Ing: External Validity and Overidentification in the Late Framework," in *Advances in Economics and Econometrics: Theory and Applications, Tenth World Congress*, ed. by D. Acemoglu, M. Arellano, and E. Dekel, vol. III of *Econometric Society Monographs*. Cambridge University Press / National Bureau of Economic Research.

Angrist, J. D., K. Graddy, and G. W. Imbens (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.

Angrist, J. D., and A. B. Krueger (1999): "Empirical Strategies in Labor Economics," in *Handbook of Labor Economics*, ed. by O. C. Ashenfelter, and D. Card, vol. 3, Part A, chap. 23, pp. 1277–1366. Elsevier.

Angrist, J. D., and J.-S. Pischke (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.

Angrist, J. D., and M. Rokkanen (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.

Aronow, P. M., C. Samii, et al. (2017): "Estimating Average Causal Effects under General Interference, with Application to a Social Network Experiment," *The Annals of Applied Statistics*, 11(4), 1912–1947.

Ashenfelter, O. (1978): "Estimating the Effect of Training Programs on Earnings," *Review of Economics and Statistics*, 60(1), 47–57.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2018): "Matrix Completion Methods for Causal Panel Data Models," *NBER Working Paper*, 25132.

Athey, S., and G. W. Imbens (2017a): "The Econometrics of Randomized Experiments," in *Handbook of Economic Field Experiments*, ed. by A. V. Banerjee, and E. Duflo, vol. 1, chap. 3, pp. 73–140. Elsevier.

——— (2017b): "The State of Applied Econometrics: Causality and Policy Evaluation," *Journal of Economic Perspectives*, 31(2), 3–32.

Baird, S., J. A. Bohren, C. McIntosh, and B. Özler (2018): "Optimal Design of Experiments in the Presence of Interference," *Review of Economics and Statistics*, 100(5), 844–860.

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

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81(2), 608–650.

Bertanha, M., and G. W. Imbens (2019): "External Validity in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, Forthcoming.

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

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.

Chaplin, D. D., T. D. Cook, J. Zurovac, J. S. Coopersmith, M. M. Finucane, L. N. Vollmer, and R. E. Morris (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.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018): "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21(1), C1–C68.

Christensen, G., and E. Miguel (2018): "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 56(3), 920–80.

Clemens, M. A. (2017): "The Meaning of Failed Replications: A Review and Proposal," *Journal of Economic Surveys*, 31(1), 326–342.

Connolly, P., C. Keenan, and K. Urbanska (2018): "The Trials of Evidence-Based Practice in Education: A Systematic Review of Randomised Controlled Trials in Education Research 1980–2016," *Educational Research*, 60(3), 276–291.

Cox, D. R. (1958): *Planning of Experiments.* Wiley.

Deaton, A. (2010): "Instruments, Randomization, and Learning about Development," *Journal of Economic Literature*, 48(2), 424–55.

Deaton, A., and N. Cartwright (2018): "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science & Medicine*, 210, 2–21.

Dehejia, R., C. Pop-Eleches, and C. Samii (2019): "From Local to Global: External Validity an A Fertility Natural Experiment," *Journal of Business & Economic Statistics*, Forthcoming.

Dong, Y., and A. Lewbel (2015): "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *Review of Economics and Statistics*, 97(5), 1081–1092.

Duflo, E., and A. Banerjee (2017): *Handbook of Field Experiments*, vol. 1. Elsevier.

Eckles, D., B. Karrer, and J. Ugander (2017): "Design and Analysis of Experiments in Networks: Reducing Bias from Interference," *Journal of Causal Inference*, 5(1).

Farrell, M. H. (2015): "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations," *Journal of Econometrics*, 189(1), 1–23.

Glazerman, S., D. M. Levy, and D. Myers (2003): "Nonexperimental versus Experimental Estimates of Earnings Impacts," *The Annals of the American Academy of Political and Social Science*, 589(1), 63–93.

Hahn, J., and R. Shi (2017): "Synthetic Control and Inference," *Econometrics*, 5(4), 52.

Halloran, M. E., and C. J. Struchiner (1991): "Study Designs for Dependent Happenings," *Epidemiology*, pp. 331–338.

——— (1995): "Causal Inference in Infectious Diseases," *Epidemiology*, pp. 142–151.

Heckman, J. J., H. Ichimura, and P. E. Todd (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65(2), 261–294.

Heckman, J. J., R. Lalonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in *Handbook of Labor Economics*, ed. by O. C. Ashenfelter, and D. Card, vol. 3, Part 1, chap. 31, pp. 1865–2097. Elsevier Science, Amsterdam.

Heckman, J. J., and E. J. Vytlacil (2007a): "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. 6, Part B, chap. 70, pp. 4779–4874. Elsevier.

——— (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 *Handbook of Econometrics*, ed. by J. J. Heckman, and E. Leamer, vol. 6, Part B, chap. 71, pp. 4875–5143. Elsevier.

Hong, G., and S. W. Raudenbush (2006): "Evaluating Kindergarten Retention Policy: A Case Study of Causal Inference for Multilevel Observational Data," *Journal of the American Statistical Association*, 101(475), 901–910.

Hotz, V. J., G. W. Imbens, and J. H. Mortimer (2005): "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations," *Journal of Econometrics*, 125(1-2), 241–270.

Hudgens, M. G., and M. E. Halloran (2008): "Toward Causal Inference with Interference," *Journal of the American Statistical Association*, 103(482), 832–842.

Imai, K., G. King, and E. A. Stuart (2008): "Misunderstandings between Experimentalists and Observationalists about Causal Inference," *Journal of the Royal Statistical Society: Series A*, 171(2), 481–502.

Imbens, G. W. (2010): "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)," *Journal of Economic Literature*, 48(2), 399–423.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–75.

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

Imbens, G. W., and J. M. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 5–86.

Ioannidis, J. P. A. (2005): "Why Most Published Research Findings Are False," *PLoS Med*, 2(8), e124.

Ioannidis, J. P. A., T. D. Stanley, and H. Doucouliagos (2017): "The Power of Bias in Economics Research," *Economic Journal*, 127(605), F236–F265.

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

Knaus, M., M. Lechner, and A. Strittmatter (2018): "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence," *CEPR Discussion Paper*, DP13402.

Knaus, M. C., M. Lechner, and A. Strittmatter (2017): "Heterogeneous Employment Effects of Job Search Programmes: A Machine Learning Approach," *IZA DP*, 10961.

LaLonde, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76(4), 604–620.

Leamer, E. E. (1983): "Let's Take the Con Out of Econometrics," *American Economic Review*, 73(1), 31–43.

Lechner, M. (2019): "Modified Causal Forests for Estimating Heterogeneous Causal Effects," *CEPR Discussion Paper*, DP13430.

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

List, J. A., C. Bailey, P. Euzent, and T. Martin (2001): "Academic Economists Behaving Badly? A Survey on Three Areas of Unethical Behavior," *Economic Inquiry*, 39(1), 162–170.

Liu, L., and M. G. Hudgens (2014): "Large Sample Randomization Inference of Causal Effects in the Presence of Interference," *Journal of the American Statistical Association*, 109(505), 288–301.

Liu, L., M. G. Hudgens, and S. Becker-Dreps (2016): "On Inverse Probability-Weighted Estimators in the Presence of Interference," *Biometrika*, 103(4), 829–842.

Lundin, M., and M. Karlsson (2014): "Estimation of Causal Effects in Observational Studies with Interference between Units," *Statistical Methods & Applications*, 23(3), 417–433.

Manski, C. F. (2013): "Identification of Treatment Response with Social Interactions," *The Econometrics Journal*, 16(1), S1–S23.

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004): "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies.," *Psychological Methods*, 9(4), 403.

McCloskey, D. N., and S. T. Ziliak (1996): "The Standard Error of Regressions," *Journal of Economic Literature*, 34(1), 97–114.

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.

Perez-Heydrich, C., M. G. Hudgens, M. E. Halloran, J. D. Clemens, M. Ali, and M. E. Emch (2014): "Assessing Effects of Cholera Vaccination in the Presence of Interference," *Biometrics*, 70(3), 731–741.

Rigdon, J., and M. G. Hudgens (2015): "Exact Confidence Intervals in the Presence of Interference," *Statistics & probability letters*, 105, 130–135.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

Roth, A. E. (1994): "Lets Keep the Con out of Experimental a Methodological Note," *Empirical Economics*, 19(2), 279–289.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66(5), 688–701.

——— (1977): "Assignment to Treatment Group on the Basis of a Covariate," *Journal of Educational and Behavioral Statistics*, 2(1), 1–26.

Rubin, D. B. (1978): "Bayesian Inference for Causal Effects: The Role of Randomization," *The Annals of Statistics*, pp. 34–58.

Sobel, M. E. (2006): "What Do Randomized Studies of Housing Mobility Demonstrate? Causal Inference in the Face of Interference," *Journal of the American Statistical Association*, 101(476), 1398–1407.

Sofrygin, O., and M. J. van der Laan (2017): "Semi-Parametric Estimation and Inference for the Mean Outcome of the Single Time-Point Intervention in a Causally Connected Population," *Journal of Causal Inference*, 5(1).

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

Stuart, E. A., C. P. Bradshaw, and P. J. Leaf (2015): "Assessing the Generalizability of Randomized Trial Results to Target Populations," *Prevention Science*, 16(3), 475–485.

Stuart, E. A., S. R. Cole, C. P. Bradshaw, and P. J. Leaf (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.

Tchetgen, E. J. T., and T. J. VanderWeele (2012): "On Causal Inference in the Presence of Interference," *Statistical methods in medical research*, 21(1), 55–75.

Thompson, W. C., and E. L. Schumann (1987): "Interpretation of Statistical Evidence in Criminal Trials," *Law and Human Behavior*, 11(3), 167–187.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society: Series B*, 58(1), 267–288.

Toulis, P., and E. Kao (2013): "Estimation of Causal Peer Influence Effects," *Proceedings of Machine Learning Research*, 28(3), 1489–1497.

Ugander, J., B. Karrer, L. Backstrom, and J. Kleinberg (2013): "Graph Cluster Randomization: Network Exposure to Multiple Universes," *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 329–337.

van der Laan, M. J. (2014): "Causal Inference for a Population of Causally Connected Units," *Journal of Causal Inference*, 2(1), 13–74.

Wacholder, S., S. Chanock, M. Garcia-Closas, L. El ghormli, and N. Rothman (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.

Wager, S., and S. Athey (2018): "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

Wald, A. (1940): "The Fitting of Straight Lines If Both Variables Are Subject to Error," *Annals of Mathematical Statistics*, 11(3), 284–300.

Wong, V. C., J. C. Valentine, and K. Miller-Bains (2017): "Empirical Performance of Covariates in Education Observational Studies," *Journal of Research on Educational Effectiveness*, 10(1), 207–236.

Wyss, R., A. R. Ellis, M. A. Brookhart, C. J. Girman, M. J. Funk, R. J. LoCasale, and T. Stürmer (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, 645–55.

Ziliak, S. T., and D. N. McCloskey (2004): "Size Matters: The Standard Error of Regressions in the American Economic Review," *Journal of Socio-Economics*, 33(5), 527–546.