Causality*

Michele Pellizzari
(GSEM-University of Geneva, CEPR, IZA)

1. Introduction

Does immigration reduce the employment opportunities of local residents? Does going to university increase one's potential salary in the job market? Does offering tax cuts on hiring motivate firms to employ more personnel? Does introducing a minimum wage reduce employment? Questions like these are frequently discussed in both academic and public debates. It is more than likely that you have seen politicians and experts discussing them on TV, or read articles about these questions in the press.

For example, at the time of writing this chapter, there is a heated debate in Italy on the so-called *Jobs Act*, a labour market reform project proposed by the government. Despite this provision containing several modifications of the labour legislation, the debate is polarised on the famous *Article 18* of the Workers' Statute (*Statuto dei Lavoratori*). This article advocates the reinstatement of workers who have been unlawfully dismissed. The *Jobs Act* proposes replacing the act of reinstating the worker, with financial compensation paid by the employer to the unfairly dismissed worker. Beyond any ethical, philosophical, or moral consideration of such measure, the public debate focuses, above all, on the effect that *Article 18* may have on the employment rate. *This subject inspires heated debate with divisive opinion.*. To resolve the issue, the following question must be answered: Does *Article 18*, and more generally, employment protection legislation, reduce employment by restricting the number of hirings?

Another interesting example is the *Youth guarantee*, a programme funded by the European Commission (EC) that would provide all young people in the European Union under 25 with a job or training opportunity within 4 months since the start of an unemployment spell (or inactivity unrelated to studies). This programme comes with substantial financial support from the EC and its aim is to drastically reduce youth unemployment and inactivity. The programme officially started on May 1, 2014. We therefore ask, is it possible to know whether and how the *Youth guarantee* reduces youth unemployment?

While *Article 18* was being discussed in Italy and the *Youth guarantee* in Europe, the President of the United States, Barack Obama, proposed to increase the federal minimum wage to 10 USD per hour (currently 7.25 USD/hour, as of 2014). In the United States, federal states may independently set the minimum wage, as long as it remains above the current federal minimum. Several states have already autonomously adopted the proposal of the president. Critics raised concerns that such a measure could reduce employment. President Obama was convinced of the contrary. Who is right? What is the effect of the minimum wage on employment?

A substantial share of the work of labour economists is devoted precisely to finding answers to questions such as those raised by the previous examples. Such questions refer to the effect of new legislation or reforms, or of programmes targeting specific categories (unemployed, youth, senior citizens, etc.). In general terms, the issue is really a philosophical one because it consists in understanding if and how much a phenomenon *causes* another one. Does weakening employment protection norms, for instance by abolishing the famous *Article 18*, *cause* an increase in hiring and/or in employment? Does a programme such as *Youth guarantee cause* a decrease in youth unemployment? Does raising the minimum wage *cause* a lower employment rate?

At a closer look, the question of causality can be found in all areas of economics. The theory of the

---

- This document is the English translation of the Chapter 9 "Causalità" from the book "Manuale di Economia del Lavoro", by Brucchi Luchino and Pepi De Caleo. It can only be distributed under permission of the authors.

supply of labour assumes the existence of a causal relationship between the market salary and the number of people willing to work (or the number of hours of work offered). To put it in different terms, the theory of the supply of labour describes a mechanism by which an increase in the market salary *causes* workers to offer more labour. The same applies to the demand for labour: an increase in the market salary *causes* a decrease in the number of workers that firms are willing to hire. The theory of human capital assumes that an increase in the level of education *causes* an increase in the individual's earnings potential.

If you have felt lost at least once in the confusing discussions among politicians on TV, or among columnists in newspapers and blogs, it must be evident that providing credible answers to such questions is a service of enormous value that economists can offer to society. It helps prevent people from making the wrong choice concerning their education, their savings, their decision to change work, or place of residence. It is helpful in preventing firms from making bad decisions concerning hiring and dismissal, investments and productive technologies.

Luckily enough, by applying methods and knowledge from Statistics and Econometrics, in the last decades, labour economists have made enormous progress in this field and have been able to provide answers to a large set of causal questions of high relevance for economic policy. The aim of this document is to introduce you to the concept of causality and to describe some of the most commonly used statistical methods for identifying and estimating causal effects. This will allow you to assess the validity of the various causal statements that circulate in the public debate, and possibly also to directly propose empirical evidence of interesting causal effects.

2. The definition of causality

The first thing to do when facing a causal question is to define exactly what we intend to measure. To this end, let us consider the general case of a population of individuals (but it could also be a population of firms, countries, regions, or schools) and indicate with the index i a generic individual belonging to this population. For the moment, we can omit any sampling problem and assume to be able to observe all the individuals belonging to the population of interest.

For each individual i, we define a dichotomous variable D_i with value 1 if the individual i is subject to a certain event or, technically speaking, to a certain *treatment*, and zero otherwise. The treatment is to be defined on the basis of the phenomenon whose effect we intend to measure. For instance, if the aim of our analysis is to estimate the causal effect of a programme supporting employment such as *Youth guarantee*, we would define $D_i=1$ if individual i is taking part in the programme, and $D_i=0$ for those individuals who are not participating in it. Similarly, if we want to measure the effect of university education, they will define $D_i=1$ for those individuals who hold a university degree, and $D_i=0$ for those who do not. In these examples, and for the sake of simplicity in the rest of this document, we will treat D_i as a dichotomous variable, which can only have a value of one or zero. However, most of the analysis applies without much modification to the more general case, where D_i may assume more than two values, or may also be a continuous variable.

[2.1] $$D_i = \begin{cases} 1 \text{ if } i \text{ is exposed to the treatment} \\ 0 \text{ if } i \text{ is not exposed to the treatment} \end{cases}$$

We refer to the set of treated individuals as the *treatment group*, and the set of non-treated individuals as the *control group*.

[2.2] $$\text{Treatment group } T = \{i : D_i=1\}$$
$$\text{Control group } C = \{i : D_i=0\}$$

We now define an outcome variable Y_i, which indicates the process on which we want to measure the effect of the treatment. Following the previous examples, a rather obvious outcome on which we may want to evaluate a programme such as the *Youth guarantee* could be the employment status of the participants (at a specified date since entrance into the programme). So, Y_i could be an indicator taking value 1 if the individual i is employed and 0 otherwise. Similarly, it could be interesting to measure the causal effect of university education on wages and, in that case, we would define Y_i as the wage of individual i. The exact definition of the outcome variable Y_i depends upon which phenomenon is considered the most interesting, and upon whether and how such phenomenon is measured. We are frequently interested in measuring the effect of a given treatment on several outcome variables, but for simplicity, in this document we will focus on just one outcome.

[2.3] Y_i = observed value of the outcome variable for individual i

The exact definition of the variables D_i and Y_i is important because it allows us to clarify the objective of the causal analysis. In other words, these definitions help us formalise the object of the causal question and define with precision the effect *of what*, *on what* we want to measure.

The indicator D_i, by identifying who is subject to the treatment and who is not, defines the *"of what"* of the causal effect we want to measure, whereas the variable Y_i defines the *"on what"*.

In order to precisely define what is meant by causal effect, it is necessary to elaborate on the notion of the outcome variable. In particular, it is useful to abstractly think of the existence of two possible outcomes, also called *potential outcomes*, which we indicate with $Y_i(0)$ and $Y_i(1)$ [Rubin 2005]. With $Y_i(1)$ we indicate the value of Y that the individual i would experience if she was subject to the treatment, i.e. if the value of D_i was one. Similarly, $Y_i(0)$ indicates the value of the outcome variable that would be observed for the same individual i, if she was not subject to the treatment, i.e. if D_i was zero.

[2.5] $Y_i(1)$ = potential outcome with treatment
$Y_i(0)$ = potential outcome without treatment

$Y_i(0)$ and $Y_i(1)$ are defined potential outcomes because, despite both being theoretically possible, in reality, only one of them is realised. In fact, the outcome that is actually observed for any subject i is $Y_i(1)$ if treated, and $Y_i(0)$ otherwise. It is impossible that, for the same individual, both potential outcomes are simultaneously realised. For the sake of clarity of notation, it is useful to define the outcome variable that is actually realised as simply Y_i. We can describe the relation it has with the potential outcomes as follows:

[2.6] $$Y_i = Y_i(0) + [Y_i(1) - Y_i(0)]\, D_i$$

Definition 2.6 simply states that when the treatment indicator is 0, the realised outcome is $Y_i(0)$, whereas when $D_i=1$ the realised outcome is $Y_i(1)$.

An example may help to clarify the nature of the concept of potential outcome. Imagine that we want to measure the causal effect of university education on wages. As we already discussed, in this case D_i will take value one if individual i holds a university degeree and zero otherwise, whereas Y_i will be a measure of the salary actually earned. Let's take two specific cases. Consider two individuals, Luca and Aldo. Luca has obtained a degree, and hence his treatment indicator equals 1: $D_{Luca}=1$. $Y_{Luca}(0)$ is the salary (not observable) that he would have earned if he had not obtained a degree, whereas $Y_{Luca}(1)$ is his actual salary, given that he obtained a degree. Contrary to Luca, Aldo did not go university, hence $D_{Aldo}=0$, and his actual salary is $Y_{Aldo}(0)$, whereas $Y_{Aldo}(1)$ is the salary he would have earned if he had gone to university and obtained a degree. In this example, the wages that are actually realised and observed are $Y_{Luca}(1)$ and $Y_{Aldo}(0)$, whereas $Y_{Luca}(0)$ and $Y_{Aldo}(1)$ are only potential wages because they could have

potentially been realised if Luca and Aldo had made different choices concerning university.

Despite $Y_{Luca}(0)$ and $Y_{Aldo}(1)$ not being observable, they are well-defined concepts, i.e. the salaries that Luca and Aldo would have earned if they had made different choices concerning obtaining a university degree. They represent the outcomes that would have been realised in the hypothetical situation, contrary to the situation that has actually been realised, namely the hypothetical situation in which Luca had not gone to university and Aldo had.

This kind of parallel world that could have been a reality if the individual choices had been different is called, in technical terms, the *counterfactual*. Despite being essentially an abstract concept, the counterfactual has a very real connotation to the extent that we all refer to it on a daily basis. In the previous example, it is certainly true that the counterfactual is not observable, but presumably when Luca decided to go to university, he made that decision on the basis of a comparison between his expectations about the values of $Y_{Luca}(1)$ and $Y_{Luca}(0)$. Most likely, you yourselves have also used the counterfactual when deciding whether to enrol in a university programme. You must have, more or less consciously, compared how much you could have earned with a degree and how much you could have earned without one. One can discuss how people form their expectations about the counterfactual and how accurate such expectations are, but it is undeniable that it represents a key-concept in countless decisions we make every day.

BOX 1. *Sliding Doors*

Sliding Doors is an American film (1998) which exemplifies the concept of the counterfactual very well. The movie starts with the main character, Helen (played by the actress Gwyneth Paltrow), who is leaving the office in a hurry to catch the underground and arrives exactly when the train doors are closing. At that point, the plot splits: one part of the movie describes the events in the scenario in which Helen, at the very last moment, catches the train, arrives at home early, finds her boyfriend in bed with another woman, breaks up with him and her life takes a new and unexpected direction. In parallel, the film also tells the story of Helen's life in the scenario where she misses the train, arrives at home late and her boyfriend's lover already left. In this alternative or counterfactual life, Helen is unaware of her boyfriend's betrayal for a long period of time, and she remains stuck in an unrewarding and unhappy relationship.

Sliding Doors very effectively describes the causal effect of "missing the underground," and allows us to see what cannot normally be seen, i.e. the counterfactual.

Thanks to the notion of potential outcomes, it is now easy to formally define what is meant by causal effect:

[2.7] $\Delta_i = Y_i(1) - Y_i(0)$ **Causal effect of D on Y for individual *i***

In other words, the causal effect of a certain treatment on a certain outcome is defined as the difference between the value of the outcome in the case of treatment and the value of the outcome in the absence of the treatment. The causal effect of university education on Luca's salary is the difference between the salary he would earn with a degree and the salary he would earn without it. The causal effect of minimum wage on employment is defined as the difference between the employment level that would be realised in the presence of a minimum wage and that that would be realised in absence of a minimum wage.

It is worth noting that the causal effect in equation 2.7 is characterised by a subscript *i*, which indicates that the effect is specific for a single individual and that different individuals may have different causal effects. This is very important, because it offers a clear and simple explanation for why people make

different choices. The reason why some people, such as Luca in the previous example, decide to go to university whereas others, such as Aldo, prefer not to, is likely due to the fact that for some individuals the causal effect of university education is higher than for others. It is easy to understand why this is the case. There are people who have an aptitude for studying and who learn a lot during their university career; consequentially, they are highly valued by employers when they enter the labour market. Other people have a harder time learning new and complex material and, for them, university education may not result in a substantial improvement in their competences, even if they obtain their degree. For these individuals, the causal effect of a degree on their salary may be considerably limited, as employers will soon realise that they are not much more productive than high school graduates. We should then think of the causal effect Δ_i as a random variable with a certain distribution in the population. For some individuals, Δ_i will be high, for others, it will be low, and for many others it will be average.

BOX 2. A Numerical Example

A simple numerical example may help to clarify the definition of causal effect. Let us consider a population of 6 individuals; Table B2.1 reports the treatment indicator (D_i) for each of these individuals, the realised outcome (Y_i), the potential outcomes in the presence or absence of treatment ($Y_i(1)$ and $Y_i(0)$), and their difference, i.e. the causal effect of the treatment ($\Delta_i=Y_i(1)-Y_i(0)$).

Table B2.1. A Numerical Example

i	D	Y	Y(1)	Y(0)	Δ
1	1	2.3	2.3	2.0	0.3
2	1	2.2	2.2	1.6	0.6
3	0	1.0	1.2	1.0	0.2
4	0	0.9	0.9	0.9	0.0
5	0	1.7	1.8	1.7	0.1
6	1	2.6	2.6	2.3	0.3

The indicator of the individual $i=1,...,N$ is reported in the first column. Following the example in the main text, we can think of the outcome as the monthly salary in thousands of euros and the treatment as university education. The data in the table show that the causal effect is different for each of the 6 individuals. For some of them (for instance, $i=2$), the effect is substantial, for others (for instance, $i=5$) it is very small and for one of them ($i=4$) it is even zero. In each of the 6 cases, the realised outcome (Y) equals $Y(1)$ or $Y(0)$ depending on whether the treatment indicator is one or zero.

Unfortunately, in reality the data in the table B2.1 are not all observable. As already mentioned in the main text, the potential outcomes are never realised simultaneously, hence the observable variables are only those in the first three columns, indicated in the darker section (note that in the case of sample data, only a subgroup of individuals will be observable). As we will see more in details in the next section, this lack of information on potential outcomes is the origin of the fundamental problem of causal inference.

3. The fundamental problem of causal inference

In the previous section, we formally defined what is meant by a causal effect of a certain treatment D over an outcome variable Y (equation 2.7). The definition given is based on the concept of the counterfactual. The causal effect is the difference between the value of the outcome variable in the presence and in the absence of the treatment, and it can be different for each component of the population.

Unfortunately, it is not possible to directly measure the causal effect for each individual, since only one of the potential outcomes is realised. In other words, the actual outcome is only one – we have called it Y_i and we defined it in equation 2.6, showing that it equals $Y_i(0)$, if the individual i is not subject to the treatment, or $Y_i(1)$ if the individual is subject to the treatment.

Recalling once again the example of university education, the observable wage was $Y_i(1)$ for graduates and $Y_i(0)$ for non-graduates – it is not possible to observe both outcomes for either group. Therefore, it is not possible to know how much a graduate would have earned if he had not graduated. We can make hypotheses or form an opinion on the matter, but it still holds that the wage that a person would have earned if s/he had not graduated is not observable. The same applies for non-graduates, for whom it is not possible to know the wage they would have earned if they had gone to university.

This point is extremely important, and it is worth clarifying a very frequent misunderstanding. The potential outcomes $Y_i(1)$ and $Y_i(0)$ are to be considered as values of the outcome variable in the presence and in the absence of the treatment, but with all other factors influencing the outcome held constant. For instance, you may have often heard, especially in the media, causal statements based on the comparison between intertemporal values of a certain variable, usually before and after a certain economic measure. Probably, the most frequent current case is that of the euro. Many claim that the euro has been the cause of a long series of economic difficulties for Italy (but not only for Italy) – such as high unemployment rates, high prices, or low growth of gross domestic product (GDP). Such claims are often substantiated on the basis of the comparison between the economic situation of the country before and after the introduction of the euro. Unfortunately, providing causal evidence is not that simple. According to our definition, the causal effect of the euro on, for instance, the unemployment rate, is the difference between the unemployment rate observed today and the unemployment rate we would have observed if the euro had not been introduced. Since the birth of the single currency, many things have happened and many of these things, including the euro, have probably had a causal effect on the evolution of unemployment. It is difficult to maintain that the differences observed over time are due just to – or mainly to – one factor. The comparison between potential outcomes is correct only if all the other conditions are held constant. The difficulty lies in the fact that there are countless conditions that may vary over time.

Recall once again the example of university education. Let us consider a situation in which a person has worked for some time before starting university; s/he then decides to stop working to fully dedicate their time to their studies and, once having obtained a degree, returns to the workforce. You could be tempted to measure the causal effect of university education by comparing the wage earned before graduation with that following graduation. However, such a comparison would not be holding all else constant because, between pre-education and post-education employment, many things have changed in addition to the individual's qualifications. For example, the general conditions of the economy or of the market may have changed, and the person is older, to cite the obvious. Thus, such a comparison would not be holding all else equal.

It should be clear now that the fundamental reason for which the causal effect of treatment D on an outcome Y for a generic individual i is not measurable is that for all individuals it is possible to observe only one of the potential outcomes, and never both.

Fundamental problem of causal inference: it is impossible to observe, for the same individual i, both the values $Y_i(1)$ and $Y_i(0)$, and it is therefore impossible to measure the causal effect $\Delta_i = Y_i(1) –$

$Y_i(0)$.

It is important to emphasise that the problem of causal inference is not a problem of sampling. In other words, the causal effects Δ_i are not measurable even when it is possible to collect all the data concerning all the individuals belonging to the population of interest. Simply put, the outcomes can never be observable simultaneously.

Given the impossibility to measure the causal effect of each single individual, it is legitimate to investigate oneself whether it is possible to measure some sort of synthetic indicator of the distribution Δ_i in the population of reference. For instance, it could be interesting to know the Average Treatment Effect (ATE).

[3.1] $E(\Delta_i) = E[Y_i(1) - Y_i(0)] = ATE$ **Average Treatment Effect (ATE)**

In many situations, it is more interesting to estimate a different moment of the distribution of Δ_i, such as the *Average Treatment Effect on the Treated* (ATT).

[3.2] $E(\Delta_i) = E[Y_i(1) - Y_i(0)|T_i=1] = ATT$ **Average Treatment Effect on the Treated (ATT)**

While the ATE is calculated as the mean of the individual causal effect on the whole population of interest, the ATT considers only the individuals belonging to the treatment group. Since the ATE synthesises the information on the whole population, one may think of it as the most interesting statistic. On the contrary, it is probably the ATT that is given more attention by researchers, for both technical and interpretative reasons. From a technical point of view, the estimation of the ATT requires weaker assumptions if compared to the ATE, which we will see in more detail in the next section. We now try to clarify the issue of interpretation with a simple example to explain that, in many situations, the ATT can indeed be the most interesting causal effect to consider.

Imagine we want to measure the causal effect of hospitalisation on citizens' health. That is, we want to understand whether, and how useful hospitals are[1]. To this end, we define the treatment $D_i=1$ if the individual i has been in the hospital and 0 otherwise. The potential outcomes are the health status of the individual in the case of treatment in the hospital, $Y_i(1)$, and in the absence of such treatment, $Y_i(0)$. Leave aside, for a moment, the fundamental problem of causal identification and let's assume to have information for the whole population on both the potential outcomes and on the treatment. We might then calculate the individual effect of the treatment on all citizens. It is reasonable to think that a substantial share of the population, probably the majority, are in good health and only a minority suffer from poor health. In other words, for the majority of healthy individuals, $Y_i(0)$ has high values whereas for the minority of the ill, $Y_i(0)$ has low values. It is then equally reasonable to assume the treatment effect to be almost null for the healthy. Indeed, if a healthy person went to the hospital the following would ensue: s/he would be sent back home by the attending doctor right after a short visit. Conversely, for the ill, going to the hospital would have with all probability a positive effect on their health because the doctor would recognise the disease affecting them, and would redirect them to the proper specialist who would intervene with a therapy to heal the diseases and/or manage the symptoms. Obviously, there could be cases in which the effect of the hospital on a healthy individual is negative – when, for instance, he is contaminated by the ill s/he is in contact with during the hospitalization – or cases in which the hospitalisation of an unhealthy individual is negative – because the doctor misdiagnoses or because, in the attempt of confining or curing the disease, it progressively worsens. However, generally speaking, the treatment effect is very likely to be null for the majority of healthy individuals and positive for the minority of the ill. The Average Treatment Effect (ATE) will then be the mean of many zeros for the majority of healthy individuals and of few positive values for the few ill, which will finally result in a

[1]This example is taken from chapter 2 in the econometrics textbook by Angrist and Pischke [2009].

rather small effect. We could reasonably assume the ill to be those exposed to the treatment, which is to say that, generally speaking, the healthy individuals do not go to the hospital and that the ill do. In this scenario, the Average Treatment Effect on treated individuals (ATT) will be calculated as the mean of the individual causal effect of only the ill and will then result in a relevantly higher number if compared to the ATE. In this case, it is easy to identify which statistic is the most indicative in order to understand whether hospitals have a positive effect on health. If we considered the ATE as the reference value for a cost-benefit analysis of hospitals, we could easily conclude that having hospitals is an inconvenience. The causal effect of a hospital on the total population is very small and, probably, does not compensate for the investment cost. However, hospitals are built to treat the ill, not healthy individuals, and thus it is better to evaluate their effectiveness solely on the sub-population of those who need them – the ill. The causal effect of hospital treatments on the ill is the ATT and not the ATE (assuming the ill – and not the healthy – are the main individuals going to the hospital).

Leaving aside the discussion on which statistic is the most indicative, neither ATE nor ATT allow us to make relevant advancements to solve the fundamental problem of causal identification. In both cases, it is necessary to observe both the counterfactual events, for the whole population in the case of the ATE, and for the treated individuals in the case of the ATT. In the next section, we will see different possible solutions to this problem.

BOX 3. A Numerical Example (2)

The numerical example in Box 2 allows us to clarify the definition of Average Treatment Effect (ATE) and Average Treatment Effect on the Treated (ATT) and how the fundamental problem of causal identification applies to them. Due to this problem, only the first 3 columns in Table B2.1 are effectively observable and useful to build a measure of the causal effect. Refer to the columns indicated in the darker section.

The Average Treatment Effect (ATE) is simply the mean of the individual causal effects reported in the last column, i.e. the *0.25* indicated in bold. The same values can also be computed as the difference between the means of the fourth and fifth column (*1.83 – 1.58 = 0.25*), which is to say the difference between the mean of the outcomes in the presence of treatment and of the outcomes in the absence of treatment:

[B3.1] $ATE = E(\Delta_i) = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$.

Unfortunately, neither of these two means ($E[Y_i(1)]$ and $E[Y_i(0)]$) can be computed with just the data in the first three columns.
The same applies for ATT, but the only difference is that in this case the mean is calculated on just the 3 observations for which $D_i=1$ (observations 1, 2, and 6). In the example in Table B2.1, ATT equals 0.4, i.e., the mean of Δ_i for the three treated individuals. Also in this case, the mean can be computed as the difference between the mean of $Y_i(1)$ and the mean of $Y_i(0)$:

[B3.2] $ATT = E(\Delta_i|D_i=1) = E[Y_i(1) - Y_i(0)|D_i=1] = E[Y_i(1)|D_i=1] - E[Y_i(0)|D_i=1] = 2.37 - 1.97 = 0.4$

The problem of identification refers to the second of these means. For the treated individuals, the observed outcome is $Y_i(1)$, thus the mean $E[Y_i(1)|D_i=1]$ can be estimated whereas $E[Y_i(0)|D_i=1]$ cannot because $Y_i(0)$ is never observable for the treated individuals.

4. Estimation methods of the average causal effect

In this section, we will survey some of the most used estimation methods for the average causal effect, for both the whole and the treated individuals. Up to now, we have omitted all the issues related to the sampling of data by assuming we had all the information concerning the observable variables, for all the individuals belonging to the population of interest. Let's now abandon this assumption and consider a representative sample of dimension N of the population of reference. Let's start by considering the simplest scenario and assume we know, for all individuals of the sample, only the values of the variables D_i and Y_i.

[4.1] $\{Di,Yi\}$ for all i = 1,...,N **Available data**

The Law of Large Numbers guarantees that, having available a representative sample of a certain population, the sample mean of any sample variable is a consistent estimator for the corresponding expected value in the population[2]. This implies, for instance, that the sample mean of Y_i, $\left[\frac{1}{N}\sum_{i=1}^{N} Y_i\right]$, accurately estimates the mean of the population, $E[Y_i]$.

Having only the simple data described in equation 4.1 available, there is, in reality, not much that can be done to assess the causal effect of the treatment. The most obvious statistic to be calculated is simply the difference between the mean of the observed outcome for the treated individuals and for the non-treated ones:

[4.2] $$\left[\frac{1}{\#T}\sum_{i\in T} Y_i\right] - \left[\frac{1}{\#C}\sum_{i\in C} Y_i\right]$$

where $\#T$ and $\#C$ indicate respectively the cardinality (i.e., the number of observations) of the treatment and control group. The Law of Large Numbers guarantees that equation 4.2 accurately estimates the following difference between the conditional means in the population:

[4.3] $$E[Y_i|Di=1] - E[Y_i|D_i=0]=E[Y_i(1)|D_i=1] - E[Y_i(0)|D_i=0]$$

where the last equality follows directly from the definition of Y_i (equation 2.6), according to which the observed outcome is $Y_i(1)$ when $D_i=1$ and $Y_i(0)$ when $D_i=0$. It is useful rewriting equation 4.3 adding and subtracting the same quantity, which is $E[Y_i(0)|D_i=1]$:

[4.4] $$E[Y_i|D_i=1] - E[Y_i|D_i=0]=E[Y_i(1)|D_i=1] - E[Y_i(0)|D_i=1]+E[Y_i(0)|D_i=1]-E[Y_i(0)|D_i=0]=ATT+bias$$

Equation 4.4 shows clearly that when we calculate the difference between the means of the outcome for the treated and non-treated individuals (equation 4.2) we are calculating the ATT estimator, which is biased. Such bias is equal to the difference $E[Y_i(0)|D_i=1]-E[Y_i(0)|D_i=0]$, i.e. the difference between the mean of the outcome in absence of treatment for both the treated and the non-treated individuals. In the example concerning hospitalisation, such bias is given by the difference between the average health status the ill would have had if they had not gone to the hospital and the health status of the healthy individuals, who did not go to the hospital. Hence, we can easily assume that the bias is negative in this specific

[2] A consistent estimator is an estimator whose probability of deviating from the parameter to be estimated tends to zero when the dimension of the sample N converges to infinite. It is one of the properties most often considered to evaluate the goodness of an estimator

example. Unfortunately, the fundamental problem of causal inference prevents us from observing and calculating such bias, because while it is possible to estimate the average outcome in absence of treatment for the non-treated individuals, ($E[Y_i(0)|D_i=0]$), $Y_i(0)$ is not observable for the treated individuals and thus $E[Y_i(0)|D_i=1]$ cannot be estimated.

Since the bias cannot be estimated, the only thing we can do is assume in which cases, and in which conditions, it equals zero. When this happens, the simple comparison between the sample mean of the outcome variable for the treated and non-treated individuals (equation 4.2) is a consistent estimation of the ATT.

Obviously, the bias is null when the two terms of the difference are equal, i.e., when the condition of mean-independence for $Y_i(0)$ and D_i holds:

[4.5] $E[Y_i(0)|D_i=0] = E[Y_i(0)|D_i=1] = E[Y_i(0)]$ **Mean-independence of $Y_i(0)$**

This condition indicates that the mean of $Y_i(0)$ is the same regardless of the individuals on which it is calculated, be them the sub-population of the treated (i.e., the individuals for which $D_i=1$), or the population of the non-treated ones (i.e., the individuals for which $D_i=0$). For this, the mean of $Y_i(0)$ is said to be independent of D_i. Moreover, if the mean calculated on the treatment and the control groups is the same, then it has to be equal to the mean calculated on the whole population. The mean dependence of $Y_i(0)$ allows us to use the mean of the observed outcomes for the individuals in the control group as a counterfactual for the treated ones, meaning it is used as a represantative of the outcome that would have been realised for the treated individuals if they had not been treated.

Evidently, we can also define the mean dependence for $Y_i(1)$, which holds true when the non-treated individuals, if not treated, experienced, on average, the same outcome of the treated individuals.

[4.6] $E[Y_i(1)|D_i=0] = E[Y_i(1)|D_i=1] = E[Y_i(1)]$ **Mean indipendence of $Y_i(1)$**

Once again, we are referring to an assumption which is, in reality, hardly credible. For instance, in the example of the hospitals, the mean-independence of $Y_i(1)$ would hold only if the cures were extremely effective on the ill and completely ineffective on the healthy. In other words, if the ill going to the hospital came back perfectly healthy and the healthy individuals going to the hospital remained healthy. In the most credible situation, in which the cured ill see their health status improving without necessarily recovering completely, the mean of $Y_i(1)$ would not be independent of D_i.

Equation 4.4 illustrates that the mean independence of $Y_i(0)$ is sufficient to consistently estimate the ATT through equation 4.2. Conversely, when the mean-independence of both $Y_i(0)$ and $Y_i(1)$ is assumed, then equation 4.2 also estimates the ATE:

[4.7] $E[Y_i(1)|D_i=1] - E[Y_i(0)|D_i=0] = E[Y_i(1)] - E[Y_i(0)] = ATE$

Equation 4.6 shows that if the mean-independence of both $Y_i(0)$ and $Y_i(1)$ holds, then the ATT equals the ATE. However, this is a rather infrequent situation in empirical work.

In the next sections, we will survey some of the most commonly used methods to estimate the ATE and ATT and we will see at which condition the assumption of mean-independence for one or both the potential outcomes can be justified.

BOX 4. A numerical example (3)

Referring once again to the numerical example of Boxes 2 and 3, it is possible to better clarify how the fundamental problem of causal inference also applies to the measure of the causal treatment effects, that is, the ATE and ATT. If the only observable data in reality are those in the first three columns of Table B2.1, then the corresponding estimation of equation 4.2 is the following:

$$\left[\frac{1}{\#T}\sum_{i\in T} Y_i\right] - \left[\frac{1}{\#C}\sum_{i\in C} Y_i\right] = \left[\frac{1}{3}(2.3 + 2.2 + 2.6)\right] - \left[\frac{1}{3}(1 + 0.9 + 1.7)\right] = 2.37 - 1.2 = 1.17$$

which corresponds to the ATT plus the bias:

ATT = $E[Y_i(1)|D_i=1] - E[Y_i(0)|D_i=1] = ⅓(2.3 + 2.2 + 2.6) - ⅓(2 + 1.6 + 2.3) = 2.37 - 1.17 = 0.4$

dist. = $E[Y_i(0)|D_i=1] - E[Y_i(0)|D_i=0] = ⅓(2 + 1.6 + 2.3) - ⅓(1 + 0.9 + 1.7) = 1.97 - 1.2 = 0.77$

ATT + dist. = 1.17.

4.1 Randomised studies

In several fields of medical and pharmaceutical sciences, it is an established practice to measure the causal effect of therapeutic treatments or medicines through clinical studies based on the random allocation of individuals to treatment and control groups. Such studies are commonly named *randomised experiments*. The procedure works as follows: to begin, the respondents of the study are recruited, be them animals or people. Then, they are divided into two groups, group 1 and group 2, in a completely random way – for instance, by flipping a coin or, more practically, via a computational generator of random numbers. At this point, only one of the groups is treated (e.g., with a therapy, or medicine) – let's say group 1; finally, the outcomes for all the individuals of both groups are registered. Group 1 is therefore the treatment group, and group 2 is the control group. The outcome of interest obviously depends on the specific experiment and could be, for example, body temperature, or the development of some specific antibodies. In these randomised experiments, the causal treatment effect can simply be estimated by computing the difference between the means of the outcome in the treatment group and in the control group, according to equation 4.2. Indeed, the causal allocation of individuals into the two groups guarantees that, at least theoretically, the condition of mean-independence (of both $Y_i(0)$ and $Y_i(1)$) is satisfied.

To understand this point, we can imagine that after having randomly created the two groups, one decides not to administer any treatment, but to measure nonetheless everyone's outcome. In this situation, we could observe the outcome in absence of treatment, i.e., $Y_i(0)$ for all individuals (both the treated and non-treated). Since the groups have been randomly created, there is no reason to expect them to have different average outcomes. In other words, the causal assignment to the treatment or the control group guarantees that $E[Y_i(0)|D_i=0]=E[Y_i(0)|D_i=1]$. As seen in the previous section, the mean-independence of $Y_i(0)$ is a sufficient condition for the simple comparison between the sample means in equation 4.2 to be a consistent estimator of the ATT.

Let's now imagine the scenario of a reversal of the experiment, meaning we still randomise the individuals into the two groups, but administer the treatment to group 2 – which was intended to be the control group – rather than to group 1 – which should have been the treatment group. In reality, up until the administration of the treatment, there is no element distinguishing the two groups and the decision of assigning a treatment to either group 1 or group 2 is irrelevant for the outcome of the experiment. Hence, there is no reason to expect the average outcome of the treated individuals to be different depending on the group the treatment is administered to, be it group 1 or group 2. In other words, the causal allocation

to the treatment and control groups also guarantees that $E[Y_i(1)|D_i=0]=E[Y_i(1)|D_i=1]$, which is the condition of mean-independence of $Y_i(1)$.

To summarise, in randomised studies, the conditions of mean-independence hold for both $Y_i(0)$ and $Y_i(1)$ and, as seen in the previous section, in this situation the simple difference between the average observed outcome for the treatment and the control groups identifies both the ATT and ATE, which are indeed equal.

As was aforementioned, the use of randomised experiments is a common practice in clinical and pharmaceutical studies. The application of the same principle to social sciences is obviously far more complicated, as it is not possible to oblige people to graduate or not to graduate, to go to the hospital or not to, to work or to be unemployed. Despite these obstacles, the use of randomised experiments has also spread within social sciences and particularly in labour economics. It is now worth discussing some of the methodological aspects directly related to the estimation of causal effects. Indeed, alongside the evident ethical issues, there also exists a series of technical issues that may significantly limit the validity of an experiment.

First, it is important to recall that the randomisation of individuals into the treatment and control groups guarantees the validity of the mean-independence assumption only in theory. Specifically, the foundations of statistical inference guarantee that if we could repeat the causal allocation an infinite number of times then the expected values of the potential outcomes in the two groups would be equal, on average, over the infinite randomisations. However, it is not possible in practice to repeat the experiment an infinite number of times, and the available data always refer to a limited number of individuals. If, theoretically, randomisation guarantees the validity of the mean-independence assumption, in practice this could not hold. One could not avoid, for instance, that by randomly allocating a group of people into two different groups, we end up with all women, or all short individuals, in the same group, and therefore none in the other group.

Another important matter which is often debated refers to the so-called *external validity* of the results obtained through randomised experiments. What really matters is obviously the general causal effect of the treatment, whereas experiments usually produce results relevant only in a specific context. It is then reasonable to investigate further whether, and how much, it is possible to generalise the results obtained in a single experiment. It is evidently a hard question to answer and there exist many reasons that could make the generalisation difficult or even impossible. For example, if the experiment was replicated on a larger scale there could be externalities influencing the conduct of the individuals in the control group, who would then become partially exposed to the treatment. Another situation in which the generalisability of the results of an experiment is threatened refers to potential *nonlinearities in the treatment effect*. Usually, the treatments tested in experiments are defined as a discrete change in some policy variable, such as an increase in the minimum salary of 2 dollars, or a decrease in the number of students in a class from 25 to 20. The problem is that these effects are not necessarily proportional to those that could have happened following a change of different size in the treatment, such as, for example, an increase in the minimum salary of 5 dollars, or a decrease of 3 dollars.

In reality, a solution to the problem of the external validity of randomised experiments does not exist. Perhaps the only and partial solution consists in replicating the same experiment several times and in different contexts. If the results obtained were to some extent similar, then it would be easier to make a valid claim regarding external validity.

BOX 5. Causal effects and regression analysis

In the main text we saw that the assumption of mean-independence of $Y_i(0)$ allows us to estimate the ATT simply through the difference between the sample mean of the outcomes of the treatment and control groups (equation 4.2). In the case in which the mean-independence of $Y_i(1)$ is verified too, then the same difference also estimates the ATE.

For those with some familiarity with econometrics, it is interesting to note that the same estimation can be obtained as a coefficient of a simple linear regression, using the realised outcome as a dependent variable, and the treatment indicator as the only explanatory variable.

To obtain this result, it is first of all useful to rearrange the potential outcomes as a sum of their means, and as one term of individual variation which we will call $u_i(0)$ and $u_i(1)$:

[B5.1] $Y_i(0) = E[Y_i(0)] + u_i(0)$

[B5.2] $Y_i(1) = E[Y_i(1)] + u_i(1)$

Such a transformation is valid for any statistical variable, which can indeed always be arranged as a sum of its mean, and of the deviation from it. By definition, the terms $u_i(1)$ and $u_i(0)$ must have a mean of zero.

By plugging the new definitions in B5.1 and B5.2 into equation 2.6, we obtain the following relationship:

[B5.3] $Y_i = E[Y_i(0)] + \{ATE\} D_i + \{u_i(0) + [u_i(1)-u_i(0)]D_i\}$

which can be seen as a univariate regression where $E[Y_i(0)]$ is the constant, $\{ATE\}$ is the coefficient of the only explanatory variable D_i, and $\{u_i(0) + [u_i(1)-u_i(0)]D_i\}$ is the error term. The error term is required to have a mean equal to zero, and to be uncorrelated with the explanatory variable D_i in order to use the simple ordinary least squares method to obtain consistent estimations of the coefficients of this regression, and therefore also the ATE. It is not difficult to prove that these conditions can be verified by assuming the conditions of mean-independence of $Y_i(1)$ and $Y_i(0)$ of D_i.[1]

If we now add and subtract to the RHS of equation B5.2 the term $E[u_i(1)-u_i(0)|D_i=1]D_i$, we obtain a new regression with a new error term and with a new slope coefficient which is equal to the ATT:

[B5.4] $Y_i = E[Y_i(0)]+\{ATT\} D_i + \{u_i(0) + [u_i(1)-u_i(0)]D_i - E[u_i(1)-u_i(0)|D_i=1]D_i \}$

It is not difficult to prove that the ordinary least squares method produces consistent estimations of the coefficients of this new equation by assuming the validity of the assumption of mean-independence for $Y_i(0)$.[2]

1. First, it is necessary to clarify that, once rearranged, the potential outcomes as in the definitions B5.1 and B5.2, the independence of their mean of D_i directly implies the independence of the means of $u_i(1)$ and $u_i(0)$ of D_i. The terms $u_i(1)$ and $u_i(0)$ are the only stochastic elements in equations B5.1 and B5.2, inasmuch as the means $E[Y_i(1)]$ and $E[Y_i(1)]$ are constants. Having said this, the zero-mean of the error term of the regression in B5.4 can be easily verified as follows:
   $E[u_i(0) + [u_i(1)-u_i(0)]D_i] = E[u_i(0)] + E[u_i(1)-u_i(0) \mid D_i=1] = E[D_i=1]\{ E[u_i(1)|D_i=1]-E[u_i(0)|D_i=1]\}$
   which evidently equals zero if both the mean of $u_i(1)$ and $u_i(0)$ are independent of D_i. It remains to verify the absence of correlation between the error of B5.3 and the explanatory variable D_i. To this end, it is suffice to show that the covariance is zero and, given that the error has mean zero, the covariance is simply the mean of the following product:
   $E[(u_i(0) + [u_i(1)-u_i(0)]D_i)D_i]=E[D_i=1]E[u_i(1)|D_i=1]$
   which is zero, given that the condition of mean-independence of $u_i(1)$ holds.
2. As for equation B5.3, also the zero-mean of the error of equation B5.4 can be easily verified:
   $E[u_i(0) + [u_i(1)-u_i(0)]D_i - E[u_i(1)-u_i(0) \mid D_i=1]D_i] = E[D_i=1]E[u_i(0)|D_i=1]$
   which evidently equals zero if the mean of $u_i(0)$ is independent of D_i. The absence of correlation between the error and D_i. is similarly proved:
   $E[D_i\{u_i(0) + [u_i(1)-u_i(0)]D_i - E[u_i(1)-u_i(0) \mid D_i=1]D_i\}]=E[D_i=1]E[u_i(0)|D_i=1]$.

4.2 Matching methods

In economics, as in social sciences in general, it is obviously not possible to always have data available from randomised experiments. Moreover, as we have seen, even if such data were available, their validity could be questioned.

For these reasons, various estimation methods have developed as alternatives to the simple comparison between the average outcome of the treatment and control groups. However, such methods require data which are slightly more detailed than those described at the beginning of this section. Let's imagine that we have information available not only on the observed outcome Y_i and the treatment indicator D_i, but also on the other characteristics of the sampled individuals. Let's define X_i as a generic vector of such characteristics. For example, X_i could be a vector of variables of the individual i, such as sex, age, and place of residence.

[4.8] $\{D_i, Y_i, X_i\}$ for all $i = 1,...,N$ **Observable data**

One of the most common methods to estimate the causal effects with data of this nature involves finding, for each individual in the treatment group, the individual (or individuals) similar to them in a control group, and using it (or them) as the counterfactual. The variables in X_i are indeed useful in order to determine what would be defined as the similarity (or similarities) between the treated individual and its counterfactual. This method is commonly called *matching*, of which there are several variants.

The simplest variant consists of considering, for all treated individuals, that their respective "match" in the control group shares the exact same values of all their respective variables in X_i. More formally, for all subjects i belonging to the treatment groups, a control group is defined as follows:

[4.9] $C_i = \{j : D_j = 0 \text{ and } X_j = X_i\}$

i.e., a control group composed of all those individuals j for which X_j and X_i are identical. Hence, for all the treated individuals the causal treatment effect is calculated as follows:

[4.10] $$\hat{\Delta}_i = Y_i - \frac{1}{\#C_i}\sum_{j \in C_i} Y_j$$

where $\#C_i$ indicates the cardinality of C_i, which is the number of individuals belonging to the set C_i. Following a notation widely used in econometrics, we use the symbol ^ to indicate an estimator. Once the control groups, C_i, for all the treated individuals is created, it is possible to calculate the ATT as the mean of all the individuals' causal effects:

[4.11] $$\widehat{ATT}_m = \frac{1}{\#T}\sum_{i \in T} \hat{\Delta}_i$$

where T is the treatment group, i.e., the set of all the treated individuals, and with $\widehat{ATT}_m$ we indicate the matching estimation of the ATT[3]. It is worth noticing that in this procedure the same control individual can be used to compute the counterfactual of other treated individuals, meaning that s/he can belong to the control group of different treated individuals.

To understand the difference between the empirical estimation of the ATT and the matching one, it is useful to clarify the assumptions underlying the two methods. As for the empirical estimation, we already noted that the necessary assumptions guaranteeing the consistency of the estimator proposed in equation [4.2] is the mean-independence of $Y_i(0)$. The consistency of the matching estimation relies on a very similar assumption, which is the conditional mean-independence of $Y_i(0)$:

[3]For completeness, we will define $\widehat{ATT}_e$ the expression in equation [4.2].

[4.12] $E[Y_i(0)|X_i, D_i=0] = E[Y_i(0)|X_i, D_i=1]$ **Conditional mean-independence of $Y_i(0)$**

In other words, in experimental studies, the individuals in the control group are simply assumed to represent a valid counterfactual for the treated individuals, and such an assumption is reasonable due to the randomised creation of the groups. In non-randomised studies which make use of the matching procedure, this same assumption is used, but only under the condition of comparing the treated and non-treated individuals who share the same values of the characteristics in X_i. The validity of this assumption obviously depends on the quality of the vector of the observable characteristics. The richer X_i is in terms of variables and information – and, in particular, information which is important to define the treatment – the greater the reliability of the matching procedure.

In the example on university education, one of the main reasons which the mere comparison between the wages of graduates and those of non-graduates is not a good estimator of the ATT is due to the fact that students who perform better are those pursuing further education and, consequentially, will also earn more. In this case, a matching estimator of the ATT would be reliable if among the variables of the vector X_i there was some measure of their propensity to do well, specifically an attitudinal test, a logic test, or their final GPA.

In principle, one would like to observe as many variables as possible in X_i and these variables to be very detailed. It may happen, however, that for some of the treated individuals there are in the control group, there are no individuals who have the exact same values for all the variables in X_i, and therefore the set C_i defined in equation 4.9, happens to be empty. Clearly, this problem becomes more probable when X_i is particularly rich in information.

For the treated individuals for whom there are no valid controls it is not possible to compute $\hat{\Delta}_i$ and, consequentially, the matching estimation of the ATT will result in the mean of the individual effects only for those individuals whose C_i is not empty. Technically, an estimation performed in this way is no longer the ATT, which should normally be computed as the mean of the Δ_i on the set of all the treated individuals.

Here, we are facing one of those cases economists like to call a trade-off: on one side, we would like vector X_i to contain many and very detailed variables, but the richer the vector, the higher the probability that many treated individuals will end up lacking an adequate control. In econometrics, this situation is called the curse of dimensionality: as the dimensionality of X_i increases (vector containing more variables, which are more detailed), the more difficult it is to find adequate controls.

How can we solve this shortcoming? The solution consists in synthesising all the relevant information of X_i in a single variable. Obviously, there are several ways to build such a synthetic variable, but it has been proved that only one particular function is able to do so without loss of information, which is important for the estimation of the causal treatment effect. This function predicts the probability of being treated on the basis of the variables in X_i, and is called a *Propensity Score* [Rosenbaum and Rubin 1983]:

[4.13] $P_i = Pr(D_i=1|X_i) = E[D_i=1|X_i]$ **Propensity Score**

It is possible to estimate the propensity score by simply calculating the share of treated individuals out of the treated and control groups who share the same values of the variables, X_i, or, for those a bit familiar with econometrics, as the prediction of the linear regression of D_i on all the X_i's.[4] Once the value of the propensity score for all individuals, both treated and non-treated, is obtained, it is possible to build the matching estimator for the ATT as described above, but replacing the entire vector X_i with just the propensity score. In other words, for each treated individual, his/her control group will be defined as the set of non-treated individuals who share the same value of P_i, i.e., those individuals who, given their X_i

[4] In reality, the prediction of *probit* models are preferred in practice to guarantee that the estimated probabilities belong to the interval [0,1].

characteristics, had the same probability of being treated but, for one reason or another, ended up in the control group.

Even though the dimensionality of the variables, based on the association of treated and control individuals, has been significantly reduced, it is still possible that the control group may not exist for some of treated individuals, with exactly the same value of P_i. It is then a common practice to define the control group as the non-treated individual(s) with the closest value to the propensity score:

$$[4.14] \qquad C_i = \{j : |P_i - P_j| = min_{k \in C} |P_i - P_k|\}$$

There are variants on how the propensity score can be used; for example, it is possible to consider not only the individual in the control group with the closest propensity score to the treated individual, but *all* the individuals whose propensity score is contained in a given neighbourhood of P_i.

The advantage of the propensity score with respect to the other manufactured measures of the variables in X_i lies in the fact that the estimation of the ATT performed by using P_i to match treated individuals and controls holds valid under the same assumption of the conditional mean-independence of $Y_i(0)$ necessary for the validity of the $\widehat{ATT}_m$ estimator (equation [4.11]). Using the propensity score allows us to resolve the problem of the dimensionality of X_i without additional assumptions.

Propensity score matching is used in many empirical studies in labour economics, as well as in other areas of applied economics and other disciplines, such as sociology and political science. As an example, it is worth citing the study by Ichino et al. [2008], which deals with an extremely important issue in the Italian context (but not exclusively), i.e., if and how much a fixed-term or a temporary job helps in finding a stable job. The results by Ichino et al. [2008] indicate that a individual working in a temporary job has a higher probability of finding a stable job compared to a similar individual who is still actively looking for a job, identified with the method of propensity score matching.

The matching approach to the estimation of causal effects, although interesting from several points of view, is also subject to critique. Firstly, one can prove that the same results can be obtained in a simpler way with a linear regression, with the observed outcome Y_i as the dependent variable and the treatment indicator and all the variables in X_i as explanatory variables. The differences refer to rather secondary aspects, such as the weight assigned to single observations in the regression. The matching method, therefore, would not illustrate any new results, but would merely show a different and more intuitive way of presenting the results. Another criticised element is the lack of explanation in matching strategies, for the fact that individuals who share the same observable characteristics end up in different groups, some in the treatment group and some in the control group. As there is not a sufficient answer for this question, some individuals are unavoidably sceptical about the fact that, given the same values of X_i,, the control can be considered as the counterfactual of treated individuals. Still, there remains some doubt that there are important factors determining who is treated and who is not, which are not observed in the vector X_i. In this case, the assumption of conditional mean-independence would not be valid and, along with it, nor would the matching method.

BOX 6. Matching and regression analysis

The fundamental principle of the matching procedure consists in comparing the outcomes of the treatment and the control groups, given that the values of their individual characteristics X_i are the same. Those familiar with the basic notions of econometrics would recognise that this is the same principle of multivariate regression. Estimating the effect of a given variable – for instance D_i – on an outcome variable – for instance Y_i – with the condition of X_i being held constant is a standard problem in econometrics and it is solved by estimating the coefficients of a regression with the outcome variable as dependent and including the explanatory variables both in the treatment and the vector X_i:

[B6.1] $$Y_i = \alpha + \beta D_i + \gamma X_i + \varepsilon_i$$

This equation corresponds exactly to equation B5.4 (or to B5.3, depending on whether only the mean-independence of $Y_i(0)$ is assumed or also that of $Yi(1)$), where $\alpha = E[Yi(0)]$ and $\beta=ATT)$. Moreover, it can easily be proved that when the mean of $Y_i(0)$ is independent of D_i then the error ε_i has mean zero and is uncorrelated with D_i.

In Box 5 we proved that the assumption of simple independence (i.e., not conditional with respect to X_i) of the mean of $Y_i(0)$ allows us to estimate the ATT from a univariate regression of the outcome Y_i over the treatment indicator D_i. If, instead, as usually occurs in the absence of randomised experiments, the mean-independence of $Y_i(0)$ is assumed to hold only conditionally to a vector of additional variables X_i, the *ATT* can still be estimated through a simple regression, but it will be necessary to include the vector X_i among the explanatory variables, as in equation B6.1.

In reality, the estimator of the ATT calculated as such does not exactly correspond to the estimator obtained through the matching procedure based on the propensity score as described in the main text. However, the differences are minimal and refer exclusively to the weight assigned to each observation *i* in the sample.

1. For a detailed (but advanced) proof, see Angrist and Pischke [2009].

4.3 Differences in differences (DiD) method

Let's consider the case in which we have information on the outcome for the treatment group and for the control group, both before and after the implementation of the treatment. In other words, let's imagine that we have the following data available:

[4.15] $\{D_{it}, Y_{it}\}$ for all $i = 1,...,N$ and $t = 0,1$ **Observable data**

where $D_{i1}=0$ for all the individuals. That is, the outcome variable is observed twice for all subjects – one observation before the treatment, and the other one after. At time $t=0$ no individual has been treated, whereas at $t=1$ some individuals have been treated and some have not.

This situation often occurs, especially when the treatment consists of state interventions which involve policy reforms that only apply to a portion of the population. For instance, this is the case of the famous study assessing the effect of an increase in the minimum federal wage in the United States on employment [Card and Krueger 1994]. This increase was introduced in 1992 and only effected workers in Pennsylvania. There are countless examples of selective reforms such as this: increased subsidies for senior workers, training programmes for young people, subsidies for small firms and not big ones, and so on.

Data as described in equation 4.15 allow us to calculate an estimator of the ATT by using an

alternative assumption to the mean-independence of $Y_i(0)$. To clarify this point, let us rearrange ATT at time $t=1$ as follows:[5]

[4.16] $$ATT = E[Y_{i1}(1) - Y_{i1}(0)|D_{i1}=1] = E[Y_{i1}(1) - Y_{i0}(0)|D_{i1}=1] - E[Y_{i1}(0) - Y_{i0}(0)|D_{i1}=1]$$

where the last equality is obtained by adding and subtracting $Y_{i0}(0)$ to the argument of the expected value. Now recall that the individuals treated at time 1 were not treated at time 0 (as no one was treated at time 0); thus the observed outcome is $Y_{i1}(1)$ at time 1 and $Y_{i0}(0)$ at time 0. Therefore, it is possible to estimate the first of the expected values in equation 4.16, i.e., $E[Y_{i1}(1) - Y_{i0}(0)|D_{i1}=1]$, by comparing the mean of the difference between the observed values of the variable Y at time 1 and time 0 for the treated individuals. Estimating the second expected value, i.e., $E[Y_{i1}(0) - Y_{i0}(0)|D_{i1}=1]$, is more complicated because it is not possible for the treated individuals at time 1 to observe the outcome in absence of the treatment at time 1 ($Y_{i1}(0)$). However, in many applications it is reasonable to assume that the following condition holds:

[4.17] $$E[Y_{i1}(0) - Y_{i0}(0)|D_{i1}=1] = E[Y_{i1}(0) - Y_{i0}(0)|D_{i1}=0]$$ **Common trend assumption**

Assumption 4.17 is commonly referred to as the *common trend assumption*, and it requires that the average change in the outcome between period 0 and period 1 that *would have* been observed for the treated *if* they had not been treated is equal to the change observed for the controls. Obviously, the validity of this assumption needs to be evaluated on a case-by-case basis, but it is nonetheless important to note that this assumption is weaker than the simple mean-independence of $Y_{it}(0)$. Indeed, it can be easily seen that if the mean-independence of $Y_{it}(0)$ is assumed as valid (both at time 1 and at time 0), then the common trend assumption must also be valid. The contrary does not apply, i.e., the common trend assumption does not imply mean-independence of $Y_{it}(0)$.

In situations in which it is reasonable to consider assumption 4.17 as valid, it is possible to replace in equation 4.16 the term $E[Y_{i1}(0) - Y_{i0}(0)|D_{i1}=1]$, which cannot be estimated since $Y_{i1}(0)$ is not observed for the treated individuals, with $E[Y_{i1}(0) - Y_{i0}(0)|D_{i1}=0]$, which can instead be easily estimated because, for the non-treated individuals, the observed outcomes are exactly $Y_{i1}(0)$ at time 1 and $Y_{i0}(0)$ at time 0.

[4.18] $$ATT = E[Y_{i1}(1) - Y_{i0}(0)|D_{i1}=1] - E[Y_{i1}(0) - Y_{i0}(0)|D_{i1}=0]$$

The ATT can thus be estimated as follows:

[4.19] $$\widehat{ATT}_{DiD} = \left[\frac{1}{\#T}\sum_{i\in T}(Y_{i1} - Y_{i0})\right] - \left[\frac{1}{\#C}\sum_{i\in C}(Y_{i1} - Y_{i0})\right]$$

where T and C are the treatment and control groups respectively, defined on the basis of the treatment indicator at time 1:

[4.20] $$T = \{i : D_{i0}=0 \text{ and } D_{i1}=1\};\ C = \{i : D_{i0} = D_{i1}=0\}$$

You may often hear rather often causal statements which are based exclusively on the comparison between outcome variables before and after treatment of the treatment group. Imagine, for instance, that yesterday you had a fever and that last night, before going to bed, you took a paracetamol pill. The following day, you take your temperature and see that it has gone. From this outcome, can you therefore

[5]Obviously, the ATT at time $t=1$ is the only causal effect that can be estimated in this case because at time $t=0$ no one has been treated.

conclude that the paracetamol lowered your temperature? Not necessarily, because you cannot know what your temperature would have been if you had not taken paracetamol. The fever could have left in any case. Many things have happened between yesterday and today that may have lowered your temperature: you slept, you got some rest, and you kept warm. The difference between your temperature yesterday and today is most likely due to a combination of factors, and not just to the paracetamol. Imagine now that a relative, a friend, or a housemate of yours had the fever yesterday afternoon but they did not take any pill. You could then compare the change of your temperature with theirs to understand what is the causal effect of paracetamol. Obviously, this comparison can rule out the effects of all the other factors that may have influenced your temperature only if your mate had been subject to the exact same factors you had been subject to, except for paracetamol. The comparison will be reliable, for instance, if the person had slept the same number of hours, in the same conditions, and so on.

The fever example may seem trivial, but it is likely that you have heard people on TV insisting that a certain measure of the government has had an effect simply because things were one way before the measure, and another way after. For example, the government may lower taxes at time 1 and unemployment decreases between *t=0* and *t=1*. But between *t=0* and *t=1* many other things may have happened impacting unemployment; namely, an economic recovery, an increase in exports, a decrease in wages, and other probable factors. How can we be certain that the observed change is the effect of the tax cuts? We would need to have information on the observed change in a similar country in which the same taxes remained unchanged.

In the notation of equation 4.19, estimating the causal effect only using the change before and after the treatment is equivalent to considering only the first term of the difference $\left[\frac{1}{\#T}\sum_{i\in T}(Y_{i1}-Y_{i0})\right]$ and, as we have seen, the advantage of having data available to calculate the second difference $\left[\frac{1}{\#C}\sum_{i\in C}(Y_{i1}-Y_{i0})\right]$ allows us to have the possibility of "clearing" the former of the potential causal effect of other factors outside of the treatment. This estimation method is called Differences in Differences (DiD) as it is based on the difference of the means of two differences.

BOX 7. DiD and regression analysis

In Boxes 5 and 6, we saw that it is possible to replicate, with simple regression methods, the ATT estimators obtained by comparing simple sample means through the assumption of mean-independence of $Y_i(0)$, be it simple or conditional.
The same result applies for the DiD estimator. Assume we have the data described in equation 4.15 available and consider the following linear regression equation:

[B7.1] $$Y_{it} = \alpha + \beta D_i + \gamma t + \delta [D_i \cdot t] + \varepsilon_{it}$$

where D_i (with no further subscript t) is a variable taking value 1 (both at time 0 and at time 1) for the individuals treated at time 1:

[B7.2] $$D_i = \begin{cases} 1 \; if \; D_{i0} = 0 \; and \; D_{i1} = 1 \\ 0 \; if \; D_{i0} = 0 \; and \; D_{i1} = 0 \end{cases}$$

The variable t indicates the period and takes value 0 in the period prior to the treatment and 1 in the following one. ε_{it} is a zero-mean error term. The coefficients α, β, γ, and δ of regression B7.1 can be easily estimated through the ordinary least squares method.
Starting with equation B7.1, we now try to calculate the mean of Y_{i1} for the treated individuals, i.e., the mean of the outcome for the treated individuals at time 1:

[B7.3] $$E[Y_{i1}|D_i=1] = \alpha + \beta + \gamma + \delta$$

The same mean is calculated for the non-treated individuals:

[B7.4] $$E[Y_{i1}|D_i=0] = \alpha + \gamma$$

By repeating the same calculations for Y_{i0}, that is at time 0, when no one has been treated, we obtain:

[B7.5] $$E[Y_{io}|D_i=1] = \alpha + \beta$$
[B7.6] $$E[Y_{io}|D_i=0] = \alpha$$

These equations allow us to establish a direct link between the ATT defined in equation 4.18 and the parameters of regression B7.1:

[B7.7] $$ATT = E[Y_{i1}(1) - Y_{i0}(0)|D_{i1}=1] - E[Y_{i1}(0) - Y_{i0}(0)|D_{i1}=0] = [\gamma + \delta] - [\delta] = \gamma$$

In other words, the same estimator of the ATT of equation 4.19 can be obtained as the coefficient γ of the regression in B7.1.

An important element contributing to the validity of the DiD estimator is the possibility to provide empirical evidence supporting assumption 4.17. Despite the impossibility to directly verify the common trend assumption between $t=0$ and $t=1$, there is often data available for, more or less, long series of periods before treatment ($t=-1,-2,-3,...$). In all the periods preceding $t=1$ no one is treated, and it would then be possible to observe the change in $Y_{it}(0)$ for both the controls and the treated individuals. In other words, it is possible to verify if the changes in the outcomes of the control and the treated individuals are

similar in the periods before treatment. If this was the case, it would then be easier to argue the similarity between the groups for *t=0* and *t=1*. If the temporal evolution of the outcome for the two groups was instead very different before the treatment, it would be rather complicated to persuasively claim that the treated individuals would have experienced changes similar to those of the control-group individuals if the treatment had not been administered.

Finally, it is also useful to notice that in order to calculate the DiD estimator in equation 4.17 it is not necessary to observe the same individuals at time *t=0* and at time *t=1*. Imagine observing two separate samples which are representative for the same population of the treated individuals and of the controls at time *t=0* and at time *t=1*. We could then calculate the ATT as:

$$[4.21] \qquad \widehat{ATT}_{DiD} = \left[\left(\frac{1}{\#T1}\sum_{i\in T1} Y_{i1}\right) - \left(\frac{1}{\#T0}\sum_{i\in T0} Y_{i0}\right)\right] - \left[\left(\frac{1}{\#C1}\sum_{i\in C1} Y_{i1}\right) - \left(\frac{1}{\#C0}\sum_{i\in C0} Y_{i0}\right)\right]$$

where *T0, T1, C1, C0* are the treatment and control groups at time 0 and at time 1.

4.4 Instrumental variables

Those familiar with the basic notions of econometrics will surely already know the method of instrumental variables. This estimation method allows us to solve a more general problem, though similar to the issue of the presence of bias in equation [4.4.].

As with matching and DiD estimations, the method of instrumental variables also requires availability of data which have to be slightly more detailed than the outcomes Y_i and treatment indicator D_i for a sample of individuals observed just once over time. In exchange for this enriched information, it is possible to estimate the ATT without resorting to the assumption of mean-independence of $Y_i(0)$ [Angrist, Imbens and Rubin 1996].

Specifically, in order to apply the method of instrumental variables it is necessary to observe not only D_i and Y_i but also (at least) another variable, which we will call Z_i:

$$[4.22] \qquad \{D_i, Y_i, Z_i\} \text{ for all } i = 1,...,N \qquad \textbf{Observable data}$$

Z_i is the instrumental variable and, for the sake of simplicity, we assume it can take only two values, zero and one, even though the method described in this section can be applied also to more general cases. In order to be used as an instrument, the variable Z_i must have two important properties. First, it must be *relevant*, meaning it must be correlated with the treatment. In other words, the percentage of treated individuals with a value of Z_i equal to 1 must be different from the percentage of treated individuals with Z_i equal to zero. Moreover, both percentages should be between 0 and 1 (excluded), otherwise the result would be D_i and Z_i being the same variable (or one the contrary of the other):[6]

$$[4.23] \qquad 0 < E[D_i|Z_i=1] - E[D_i|Z_i=0] < 1 \qquad \textbf{Instrument relevance}$$

The second property that Z_i must satisfy is *exogeneity*, which requires that Z_i does not have a direct causal effect on the outcome:

$$[4.24] \qquad E[Y_i(k)|D_i=h,Z_i=1]=E[Y_i(k)|D_i=h,Z_i=0] \text{ for all } k=0,1 \text{ e for all } h=0,1 \qquad \textbf{Instrument exogeneity}$$

To fully understand the condition of exogeneity, let's set *h=1*. In this case, both terms of equation

[6]Note that for variables taking only values 0 and 1, the expected value corresponds to the probability that the variable equals 1: $E[D_i] = Pr(D_i=1)$.

4.24 are conditioning on $D_i=1$, i.e., we are considering only the treated individuals. Now, let's divide the treated individuals into two sub-groups: on one side, those for whom the instrument Z_i equals 1; on the other side, those with Z_i equal to 0. The two means of equation 4.24 are the means of the potential outcome calculated on these two sub-groups and, due to Z_i's exogeneity, they are required to be equal. The same condition must hold for both the potential outcomes and for both the treatment and the control group. A variable Z_i satisfying the condition of exogeneity has no influence (on average) on the potential outcomes and, hence, no causal effect on the outcome variable.

Without going into details, it is possible to use condition 4.24 to prove that the difference between the mean of the observable outcome Y_i among the individuals with $Z_i=1$ and those with $Z_i=0$ is proportional to the ATT (see Box 8 for a more detailed proof):

[4.25] $$E[Y_i|Z_i=1] - E[Y_i|Z_i=0] = ATT \cdot (E[D_i|Z_i=1] - E[D_i|Z_i=0])$$

We can then rearrange equation 4.25 to obtain the following expression for the ATT:

[4.26] $$ATT = \frac{E[Y|Z_i = 1] - E[Y|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

The condition of instrument relevance guarantees that the denominator of this expression is not equal to zero.

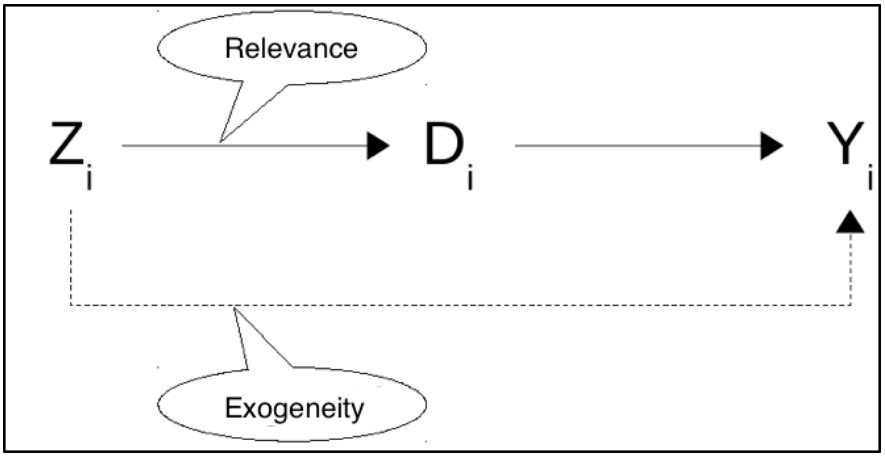
Note that both terms in the numerator and denominator are expected values of observable variables; it is thus possible to estimate the ATT by replacing all the expected values on the RHS in equation 2.46 with their corresponding pairs in the sample:

[4.27] $$\widehat{ATT}_{IV} = \frac{\left[\frac{1}{\#I_1}\sum_{i\in I_1} Y_i\right] - \left[\frac{1}{\#I_0}\sum_{i\in I_0} Y_i\right]}{\left[\frac{1}{\#I_1}\sum_{i\in I_1} D_i\right] - \left[\frac{1}{\#I_0}\sum_{i\in I_0} D_i\right]}$$

where with I_1 and I_0 we indicate the sets of individuals whose Z_i equal 1 and 0, respectively.

The description just given of how the method of instrumental variables is applied may not clarify what an instrument is, and how it can be found in practice. Less formally, an instrument is an observable variable that has a direct effect on the probability of being treated, but no direct effect on the outcome variable. Figure 4.1 graphically shows this situation. The arrow connecting Z_i and D_i indicates a causal effect of the former variable on the latter; likewise the arrow linking D_i and Y_i indicates the potential presence of a similar causal link. The dashed line indicates the absence of a causal link between Z_i and Y_i.

Figure 4.1



The instrument relevance condition (condition 4.23) states that there exists a causal link between Z_i and D_i, graphically represented from the arrow going from Z_i to D_i. The instrument exogeneity condition (condition 4.24) states instead that a direct causal effect between Z_i and Y_i does not exist. The dashed arrow between Z_i and Y_i in figure 4.1 indicates exactly this – in order for the instrument to be valid, and for equation 4.24 to hold, it is necessary that no causal effect between Z_i and Y_i exists.

Let's look at an example. In a famous study on the causal effect of university education on wages, economist David Card used the information concerning the place of residence of individuals in his dataset and the location of universities in the United States to calculate, for each person, the distance between their residence and the closest university [Card 1995]. In order to use the distance from a university as an instrument, it is necessary to assume that it satisfies both the relevance and the exogeneity conditions. The relevance condition can be justified on the basis that those who live close to a university incur lower costs to attend, and will then be more likely to go to university. Therefore, there exists a direct link between the instrument Zi, the distance from a university, the treatment D_i, and the likelihood of being graduated. It is however necessary that the instrument also satisfies the condition of exogeneity. One then needs to assume, besides the fact that those living closer to a university have a higher probability of enrolling, that there is no other reason for proximity to a university to have a direct effect on wages. Obviously, these assumptions, as always in an econometric analysis, can be criticised and it is thus important that they are always made clearly explicit to allow the evaluation of the credibility of the analysis.

BOX 8: Instrumental variables and regression analysis

Econometric textbooks usually present the method of instrumental variables within the framework of regression analysis. In particular, let's consider the simple univariate regression:

[B8.1] $$Y_i = \alpha + \beta D_i + \varepsilon_i$$

where ε_i is a zero-mean error term. The most common method to estimate the β parameter is ordinary least squares, which is based on the assumption that the explanatory variable D_i is uncorrelated with the error term, i.e., $Cov[D_i,\varepsilon_i]=0$. In Box 5, we saw that the condition of mean-independence of $Y_i(0)$ implies that there is no correlation between D_i and ε_i and, consequentially, that β is equal to the ATT.

The method of instrumental variables allows us to estimate β without using the assumption of mean-independence of $Y_i(0)$, and it is thus useful in all cases in which we are unsure of its validity. As discussed in the main text, this is almost always the case for non-experimental contexts.

Having an instrument we deem as valid available, we can exploit the condition of exogeneity to rearrange β as a function of moments of observable variables. To this end, it is useful to arrange the instrument exogeneity in terms of correlation with the error term ε_i, i.e., as $Cov[Z_i,\varepsilon_i]=0$. We now substitute $\varepsilon_i = Y_i - \alpha - \beta D_i$ in this covariance:

[B8.2] $$Cov[Z_i,\varepsilon_i] = Cov[Z_i,Y_i - \alpha - \beta D_i] = Cov[Z_i,Y_i] - \beta Cov[Z_i,D_i] = 0$$

Rearranging terms we obtain:

[B8.3] $$\beta = \frac{Cov[Z_i,Y_i]}{Cov[Z_i,D_i]}$$

Equation B8.3 suggests that β can be estimated by simply substituting the covariance in the numerator and at the denominator with the corresponding pairs in the sample:

[B8.4] $$\hat{\beta} = \frac{\sum_{i=1}^{N}(Z_i-\bar{Z})(Y_i-\bar{Y})}{\sum_{i=1}^{N}(Z_i-\bar{Z})(D_i-\bar{D})}$$

where $\bar{Y}$ and $\bar{D}$ are the sample mean of the respective variables. In econometric terms, one would say that D_i is instrumented with Z_i.

Consider now the numerator of expression B8.3 and rearrange it as follows:

[B8.5] $$Cov[Z_i,Y_i] = E[Z_i,Y_i] - E[Z_i]E[Y_i] = E[Z_i](1-E[Z_i])(E[Y_i|Z_i=1] - E[Y_i|Z_i=0])$$

Similarly, we can rearrange the denominator in B8.4 as follows:

[B8.6] $$Cov[Z_i,D_i] = E[Z_i,D_i] - E[Z_i]E[D_i] = E[Z_i](1-E[Z_i])(E[D_i|Z_i=1] - E[D_i|Z_i=0])$$

and, consequentially, B8.3 is identical to expression 4.16:

[B8.7] $$\beta = \frac{E[Y|Z_i=1]-E[Y|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]}$$

In other words, the ATT estimator can be obtained by using either an instrumental variable or both formulas 4.27 and B8.4.

There is a very important difference between the condition of relevance and that of exogeneity of the instrument. Relevance can be verified empirically in data, whereas exogeneity cannot. To understand this point, let's consider the condition of relevance described in equation 4.23. Here, the expected values can be directly estimated with the available data. Indeed, the denominator in equation 4.26 is exactly an estimation of the difference between the expected values $E[D_i|Z_i=1]$ and $E[D_i|Z_i=0]$ included in the condition of relevance. In other words, it is possible to directly verify in the data if and how much the instrument has an effect on the treatment. In the work by David Card, for instance, the author calculated the share of graduates that at the age of 18 lived close to a university and the percentage of graduates who lived far, and showed that the former is significantly higher than the latter. The expected values included in the condition of exogeneity in 4.24 cannot be estimated, because they refer to potential outcomes that are, by their nature, not observable, at least not for the entire population.

The most complicated issue of the method of instrumental variables is the search for a new instrument and, unfortunately, the econometric theory is not of great help (however, it indicates the conditions that a good instrument must satisfy). Besides this, the research of a valid and credible instrument is a process that requires a lot of creativity. For example, in a famous study, Angris and Krueger [1991] use the trimester of birth as an instrument to estimate academic performance. It is difficult to imagine that the exact moment of birth can be chosen in order to understand future income, and therefore instrument exogeneity is evident. The relevance relates instead to the laws on compulsory education, that allow individuals to legally leave school only once a certain age is reached, for instance 16 years old. Among all the students born in the same year, those born in the first months of the year will have the possibility to leave school before others. Angrist and Krueger [1991] show that those born between January and March leave school before those born between October and December.

4.5 Regression Discontinuity

The last method of estimation we review is called *Regression Discontinuity*, and it is viable in situations in which the treatment is assigned on the basis of on a deterministic rule, characterised by a point of discontinuity in an observable variable [Imbens and Lemieux 2008]. These situations are very common, as public policies often target a group of beneficiaries whom are identified through a set of deterministic and discontinuous rules. It Italy, the most remarkable case is probably the well-known "*Article 18*" of the Workers' Statute, which applies only to firms with more than 15 employees; that being said, policies aimed at subsidising young, elderly, or poor people are promoted all over the world, all characterised by their accessibility only to those who match certain criteria, e.g., age or income.

To better understand how this estimation method works, it is useful to formally identify the context in which it can be applied. First of all, we define the observable data as follows:

[4.28] $\{D_i, Y_i, W_i\}$ for all $i = 1,\ldots,N$ **Observable data**

where D_i and Y_i are, as usual, the treatment indicator and the observed outcome, whereas W_i is the variable which indicates the basis on which access to the treatment is determined. For instance, imagine a treatment which is offered only to subjects whose W_i value is over a specific threshold c:

[4.29] $$D_i = \begin{cases} 1 \; if \; W_i \geq c \\ o \; if \; W_i < c \end{cases}$$ **Treatment indicator**

As an example, the treatment could consist of the opportunity to participate in a professional retraining programme which is only accessible to people who are at least 50 years old. In this case, W_i will be the age of the individual and c will be 50 years.

In these situations, two things occur. On one hand, the condition of conditional mean-independence is valid – here the conditioning is no longer on a vector of variables X_i as in 4.12, but just on the single

variable W_i. Indeed, according to 4.29, D_i is a deterministic function of W_i and consequentially once having conditioned the expected value on a specific value W_i, it is irrelevant also conditioning it to the value D_i. In other words, if the value of W_i for a certain individual i is known, the treatment indicator D_i can then be determined.

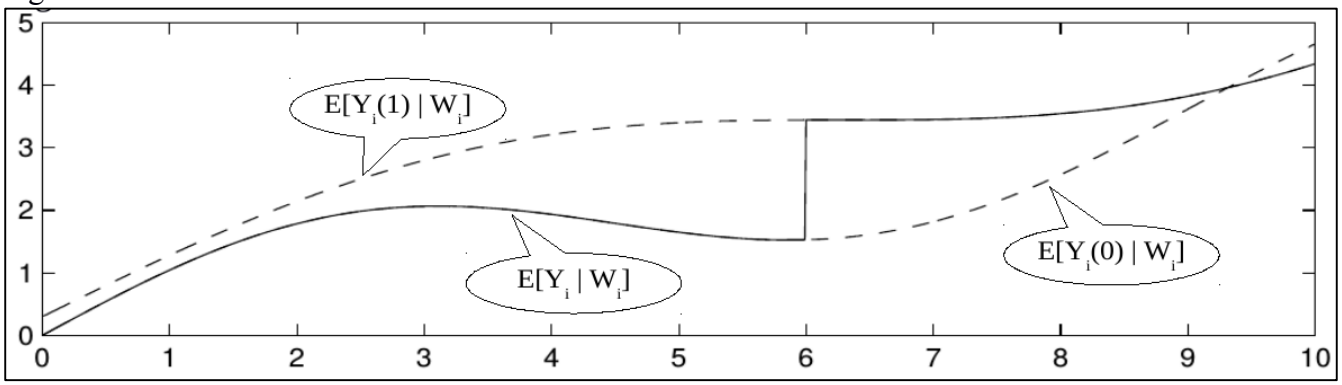
The validity of the condition of independence of the conditional means may seem like good news because, as we saw in section 4.2, we could exploit this condition to estimate the ATT. Unfortunately, the case in reality is slightly different. In section 4.2 we also saw that the possibility of estimating the ATT on the basis of conditional mean-independence can be significantly limited by the so-called *curse of dimensionality*, i.e., the fact that for certain values of the control variables there exist either only treated individuals or only controls. Normally, this happens when the dimensionality of the vector of control variables is high, which does not seem to be the case for in this instance, as we are conditioning on one variable (W_i). Nonetheless, the definition of the treatment in 4.29 overstates the curse of dimensionality, making the use of conditional mean-independence impossible for the estimation of the ATT. Indeed, for all values of W_i there exists either only treated individuals or only controls. For values of W_i which are strictly less than c only controls exist, and for values equal or greater than, only treated individuals exist. In the previous example, all the individuals younger than 50 are controls, and individuals who are 50 or more are treated.

The method of *regression discontinuity* replaces the condition of mean-independence, which becomes trivial in this context, with an assumption about the continuity of the conditional mean:

[4.30] $E[Y_i(k)|W_i]$ is a continuous function of W_i $(k=0,1)$ **Continuity of the conditional mean**

On the basis of the condition of continuity of the conditional mean, we can thus think of discontinuity – a jump in the conditional mean of the observed outcome $E[Y_i|W_i]$ – as the causal effect of the treatment on the outcome. Figure 4.2 further clarifies this point.

Figure 4.2



This figure is adapted almost identically from Imbens and Lemieux [2008], figure 2, p. 617.

The figure shows two continuous functions that represent the means of $Y_i(0)$ and $Y_i(1)$ for different values of W_i. The distance between the two curves is the casual effect of treatment and the figure clearly highlights that such effect may vary according to the value of W_i. In the figure, for example, the causal effect is almost null for very small or very big values of W_i, and it peaks around the centre of the support.

Unfortunately, data do not allow us to observe the conditional means of the potential outcomes $Y_i(0)$ and $Y_i(1)$, but only the mean of Y_i for different values of W_i. In other words, for all the values of W_i that are less than c, we observe the conditional mean of $Y_i(0)$ and for all values that are equal to or greater than c, we observe the conditional mean of $Y_i(1)$. This is, in this context, the fundamental problem of

causal identification.

However, we observe, as in figure 4.2, when $W_i=c$, the mean of the observed outcome Y_i varies discontinuously; we can think of that jump as the effect of treatment on the individuals with $W_i=c$. Figure 4.2 also clarifies why the continuity of the functions $E[Y_i(0)|W_i]$ and $E[Y_i(1)|W_i]$ is necessary in order to interpret the discontinuity in c as a causal effect. Let's try to imagine the same figure but with many other points of discontinuity in the mean of Y_i for different values of W_i. In this kind of situation, how can we determine whether the particular discontinuities observed in c were the causal effects of the treatment or, merely, one of the many jumps in the function?

More formally, the causal effect that can be estimated with this method is the ATT conditioned on $W_i=c$:

[4.31] $$ATT(c) = E[Y_i(1) - Y_i(0)|W_i=c\ ] = E[Y_i(1)|W_i=c] - E[Y_i(0)|W_i=c\ ].$$

The continuity of the conditional means for the potential outcomes allows us to replace the two expected values $E[Y_i(1)|W_i=c]$ and $E[Y_i(0)|W_i=c]$ in equation 4.31, which are not observable, with the following limits:

[4.32] $$ATT(c) = \left[\lim_{w\to c^+} E(Y_i|W_i=w)\right] - \left[\lim_{w\to c^-} E(Y_i|W_i=w)\right]$$

both of which can be estimated, since Y_i is the observable outcome, and it is equal to $Y_i(1)$ as it gets closer to c from above, and to $Y_i(0)$ as it gets closer to c from below.

There are several ways to estimate the limits of equation 4.32. If the cardinality of the sample allows us to, ATT(c) could be estimated by considering only the observations with W_i in a very small neighbourhood of c and by comparing the means of the observed outcomes for both the treated and the control individuals. Alternatively, one could estimate two simple regressions of Y_i on W_i – one using only the observations with W_i less than c, and the other using only the observations with W_i equal to or greater than c (this is the most common method used). The difference between predictions of these regressions, obtained by setting W_i equal to c are consistent estimations of the two limits in equation 4.32 (see Box 9).

BOX 9 Regression discontinuity and regression analysis

The most common way to estimate the ATT(c) defined in equation 4.32 is based on the comparison between the constants of two regression lines. The procedure is the following:

1. Only the observations in a reasonably small range around the threshold c are selected, for instance, the observations with $W_i \in [c-h,c+h]$, in which h determines the wideness of the range;
2. The parameters of the two following regressions are estimated (through the least squares method):

i. $Y_i = \alpha_1 + \beta_1(W_i-c) + e_i$ using only the observations on the left of the threshold, i.e., those with $W_i \in [c-h,c)$;

ii. $Y_i = \alpha_2 + \beta_2(W_i-c) + u_i$ using only the observations on the right of the threshold, i.e., those with $W_i \in [c,c+h]$;

3. The ATT(c) is estimated as the difference between the estimations of α_2 and α_1.

As for many other assumptions encountered in this chapter, the continuity of the conditional means of the potential outcomes cannot be verified in the data, at least not globally, because in point c the

observation of $E[Y_i(0)|W_i]$ is stopped, and the observation of $E[Y_i(1)|W_i]$ takes over. However, thinking through expression 4.31, one realises that it is technically sufficient to assume continuity at c to estimate the ATT(c). However, the validity of this assumption would hardly be convincing if the means of $Y_i(0)$ and $Y_i(1)$ were discontinuous at other points. In reality, when the method of regression discontinuity is used, it is common practice to verify whether the mean-trend of the observable outcome Y_i is continuous after and before the discontinuity. Obviously, if we happen to find that the function $E[Y_i|W_i]$ presents different discontinuities both after and before c, it would be difficult to believe it to be continuous at c.

The method of regression discontinuity is quite intuitive and efficient; in fact, it is widely used in economic studies and, particularly, in labour economics. For instance, Battistin et al. [2009] applies a discontinuity to the rules granting the entitlement to retire, to estimate the scale of the decrease in consumption, due to the ending of activity in the labour market.

The main issue ascribed to this method is the fact that it does not allow us to estimate the ATT, that is, the average causal effect on the treated individuals, but rather the causal effect on a very specific subpopulation of the treated ones, i.e., the individuals with W_i equal to c.

Figure 4.2 clearly shows that the causal effect measured at the discontinuity could be quite different from the effect measured at other values of c. To better understand the importance of this problem, we could relate this to the matter of *Article 18*, and imagine having a sample of Italian firms (each i is a firm), with W_i being the number of employees within a firm, and D_i being equal to 1 for those with more than 15 employees. We may be interested in studying the effect of *Article 18* on different outcomes, but for now, let's assume Y_i represents the investment in research and development. With the method of regression discontinuity, we could estimate the causal effect for firms with more than 15 employees, but it would be reasonable to conclude that the effect is higher for smaller and bigger firms. The method of regression discontinuity would tell us nothing about the effect of the treatment for values of W_i that are different from the threshold[7].

5. What one can (and cannot) learn from causal analysis

In this chapter, we first defined what is meant by a causal effect, and we then surveyed a series of estimation methods that can be used to measure it. The emphasis on the very concept of causality has allowed considerable advancements in empirical analysis in labour economics, as well as in numerous other fields in the social sciences.

It is nonetheless important to emphasise that causal analysis does not provide all the information needed to understand an empirical phenomenon. More specifically, a causal estimation of the effect of a treatment on an outcome does not allow one to fully understand the behavioural mechanism that triggers such an effect. This limitation is of utmost importance, because it can be very difficult to design effective interventions without fully understanding the mechanism(s) underlying the causal links between the phenomena.

An example may help clarify this point. Consider the causal effect the size of classes have on students' learning attitudes. Most studies highlight a negative effect: in classes with many students, learning is less effective. This result can be obtained through two mechanisms, which are quite different from each other. The first concerns the teacher's behaviour – in a smaller class, a teacher may be able to more smoothly adapt his/her teaching methods towards student attitudes. In other words, the teacher may be able to support students with more difficulties, and, at the same time, be able to go more in-depth on certain

[7] Obviously, the particular definition of the tratment in expression 4.29 and the particular form of the functions in figure 4.2 were chosen randomly and for the sake of clarity. The method of regression discontinuity can be applied also to cases in which the treatment has been assigned to individuals with values of W_i less than the threshold and nothing forbids the conditional mean of $Y_i(0)$ to be globally greater than that of $Y_i(1)$ or the two to intersect.

arguments, benefitting the most capable students. Adapting one's teaching attitude becomes quite difficult – if not impossible – in larger classes, e.g., those over 20 students. The second mechanism instead concerns students' behaviour, that is naturally prone – more or less frequently – to distractions, which could be provoked, for example, by a message on their phone, or by a joke from a classmate. The bigger the classes, the bigger the chances of distractions, which could effect the efficiency of the class, and hence, limit the learning opportunities for everyone. All the countermeasures we can adopt in this context depends heavily on the mechanism. If the issue points towards adapting teaching methods, we could, for instance, imagine exploiting new technologies (videos, forums, etc.) to offer more gifted students the possibility of independently delving further into some of their subjects, allowing the teacher to dedicate more time towards clarifying any doubts and queries of students with difficulties learning. If the issue has to do with distractions, we could imagine forbidding the use of mobile phones in classes, or changing the placement and arrangement of students' desks. If we are to intervene to enhance students' learning with these kind of measures when the problem lies instead in the challenge of adapting didactics, we would risk having zero positive results, or even worsening the situation.

Understanding the mechanism(s) that triggers a causal effect is as important as estimating it correctly; above all, it is crucial in order to accurately determine how to intervene.

References

Angrist, J.D. and A.B. Krueger [1991], *Does Compulsory School Attendance Affect Schooling and Earnings?*, in «The Quarterly Journal of Economics», 1975, n. 4, pp. 979-1014.

Angrist, J.D. and J-S. Pischke [2009], *Mostly Harmless Econometrics. An Empricist's Companion*, Princeton University Press.

Angrist, J.D., G.W. Imbens and D.B. Rubin [1996], *Identification of Causal Effects Using Instrumental Variables*, in «Journal of the American Statistical Association», 1996, n. 91, pp. 444–472.

Battistin E., A. Brugiavini, E. Rettore and G. Weber [2009], The Retirement Consumption Puzzle: Evidence from a Regression Discontinuity Approach, in «American Economic Review», 2009, n. 99, pp. 2209-2226.

Brucchi, L. [2001], *Manuale di Economia del Lavoro*, Bologna, Il Mulino.

Card D. [1995], *Using Geographic Variation in College Proximity to Estimate the Return to Schooling,* in *Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp*, a cura di L.N. Christofides, E.K. Grant, e R. Swidinsky, Toronto, University of Toronto Press.

Card, D. E A.B. Krueger [1994], *Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania*, in «American Economic Review», 1994, n. 84, pp. 772-93.

Ichino, A., F. Mealli and T. Nannicini [2008], From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity?, in «Journal of Applied Econometrics», 2008, n. 23, pp. 305-327.

Imbens, G.W. and T. Lemieux [2008], *Regression discontinuity designs: A guide to practice*, in «Journal of Econometrics», 2008, n. 142, pp. 615-635.

Rosenbaum, P.R. and D.B. Rubin [1983], The Central Role of the Propensity Score in Observational Studies for Causal Effects, in «Biometrika», 1983, n. 70, pp. 41-5.

Rubin, D. [2005], Causal Inference Using Potential Outcomes, in «Journal of the American Statististical Association», 2005, n. 100, pp. 322–331.

Summary of key identification assumptions

Strategy/Estimator	Assumption
Randomized experiments	Mean Independence $E[Y_i(0)\|D_i = 1] = E[Y_i(0)\|D_i = 0]$
Regression&Matching	Conditional Mean Independence $E[Y_i(0)\|X_i, D_i = 1] = E[Y_i(0)\|X_i, D_i = 0]$
Instrumental Variables	Exogeneity&Relevance of the instrument(s) $Cov(Z_i, u_i) = 0 \quad ; \; Cov(Z_i, D_i) \neq 0$
Differences-in-Differences	Common Trends $E[Y_{i1}(0) - Y_{i0}(0)\|D_{i1} = 1] = E[Y_{i1}(0) - Y_{i0}(0)\|D_{i1} = 0]$
Regression Discontinuity	Continuity of Conditional Means $E[Y_k\|W_i]$ is a continuous function of $W_i \quad \forall k = \{0, 1\}$