nccr trade regulation
swiss national centre of competence in research

**Working Paper No 2011/26| May 2011**

# Methods for Program Evaluation

**Sebastian Galiani / Guido Porto**

Abstract

Throughout this chapter we will study the general framework for program evaluation, with the aim of assessing the causal effect of a determine program, such as the impact of job-training program on earnings. The objective of this chapter is to provide the research with enough tools to think of the problem of causality in a consistent way and offer the reader the menu of sources of causal inference and the most up-to-date techniques to estimate treatment effects. This Chapter is only introductory, containing many references to encourage further readings on the field and should be accessible for those readers with a basic knowledge of econometrics and statistics.



FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

Die Nationalen Forschungsschwerpunkte (NFS) sind ein Förderinstrument des Schweizerischen Nationalfonds.
Les Pôles de recherche nationaux (PRN) sont un instrument d'encouragement du Fonds national suisse.
The National Centres of Competence in Research (NCCR) are a research instrument of the Swiss National Science Fondation.

**Methods for Program Evaluation**

**Draft Chapter, Part I**[1]

**Sebastian Galiani & Guido Porto**

May 2010

Throughout this chapter we will study the general framework for program evaluation, with the aim of assessing the causal effect of a determine program, such as the impact of job-training program on earnings. The objective of this chapter is to provide the research with enough tools to think of the problem of causality in a consistent way and offer the reader the menu of sources of causal inference and the most up-to-date techniques to estimate treatment effects. This Chapter is only introductory, containing many references to encourage further readings on the field and should be accessible for those readers with a basic knowledge of econometrics and statistics.

We will start formulating the following problem, and throughout the Chapter we will develop different approaches to assess it. Consider the case of evaluating the effects of a job-training program on relevant outcomes, such as earnings of the participants. The question we are interested in, is quantifying the improvement on earnings that the training engendered. If the earnings of the workers who received the training increased after the treatment, could we affirm that it was only the training that caused the improvement in his income? In general we can relate a presumed effect with more than one cause. Many factors are usually necessary for an effect to occur, but we seldom know all of them and how they relate to each other. Therefore, causal relationships are not deterministic, but only increase the probability that an effect will take place. In social sciences confusion tends to occur especially when we face situations with many correlated variables. But correlation does not prove causation, because we do not know what event happened first. And also, there are often other variables affecting both presumed cause and effect, which tend to make the isolation of causal relations much more complicated. Furthermore, we must be aware that causal relations occur under some conditions but seldom universally in every time, space or populations. Consequently, we must be very careful when trying to generalize a causal relationship; which.

The starting point is the same for both analyses: we have a sample of observations drawn from an unknown distribution. The aim is to find parameters of that distribution, through different estimation techniques. In the standard analysis, once we find such parameters, we can infer association among variables, estimate the likelihood of past and future events, as well as update the likelihood of events in light of new evidence of new measurement. This is done under the assumption that experimental conditions remain the same, because there is no hint in a distribution

[1] PART II will explore applications of program evaluation to trade and trade policy.

function that tells us how that distribution would differ if external conditions were to change. In fact, the laws of probability theory do not dictate how one property of a distribution ought to change when another property is modified.

Why is causal analysis different? It goes further, thus its aim is to infer aspects of the data generation process. Then, we can be able to deduce not only the likelihood of events under static conditions, but also the dynamics of events under changing conditions. This capability includes: predicting the effects of interventions of spontaneous changes; identifying causes of reported events and assessing responsibility and attribution for occurrence of events (i.e. if an event was necessary or sufficient for the occurrence of another).

In conclusion, in the one hand, probability theory deals with static conditions; it allows us to decide whether two events are mutually correlated, or dependent. This association implies at most that when we find one of the events, we can expect to encounter the other. On the other hand, causal analysis deals with changing conditions. Generally, association is circumstantial evidence for causation. However, this evidence is not sufficient because there may be some hidden confounding factors which affect both the cause and the effect.

Therefore, we would ideally need to know the earnings that the participants would have had in the absence of the program, and compare this amount with the earnings that they have after participating. As a single individual cannot belong to the treatment group and the control group, one of the potential outcomes will be observable, whereas the other will be unobservable (counterfactual). In terms of our example of the training program, we will know what the outcome of a man who participated in the job training program is, but we will never know what his earnings would have been, hadn´t he participated. Because it is impossible to observe both outcomes for the same individual, the individual-level causal effect cannot be observed or calculated. This fact that we cannot infer the effect of treatment because we do not have the counterfactual evidence is known as the *Fundamental Problem of Causal Inference*. However, the existence of this problem does not mean that causal inference is impossible. Holland (1986) states that there are two solutions: the *scientific solution* and the *statistical solution*.

On the one hand, the *scientific solution* is the approach employed in a laboratory setting, where the scientist performs an experiment in which he can introduce a treatment, holding all the other factors at their previous states and levels, so that he can claim that the change in the outcome he measures is due to the treatment. This claim relies on the plausibility and reasonability of the homogeneity assumption. On the other hand, the starting point of the *statistical solution* is recognizing that although individual-unit effect is unobservable, looking at the whole population (or determined sub-populations) can allow us to gain knowledge about the average effect of the treatment.

What is key to understand is that dealing with causality involves asking ourselves "*what if*" questions, which require building a counterfactual. Throughout this chapter, we will go through the most common and up-to-date techniques that are employed to build such counterfactuals in order to identify causal parameters. This Chapter on program evaluation should be taken as an introduction to the main practices in the field.

We will emphasize the validity of the inferences provided by the different designs we study. The concept of validity is very important and refers to the extent in which the inference can be claimed to be true. Saying that something is valid imply a judgment about how supportive is the relevant evidence we have available, of that inference being true. Validity, as any human-made judgment, is subjective and relative. Validity has different degrees, it is not absolute. It is usually constructed from empirical findings, consistency with former accepted theories. We cannot be absolutely certain that inferences drawn from a single experiment are correct or that the other inferences are false. Validity is a property of inferences, not of the designs, techniques or estimation methods. A determined design draws some inferences that may be more or less valid according to the circumstances.

We can divide validity into two types: external and internal. External Validity makes reference to the extent that inference can be generalized from a sample of units with determined settings and conditions, to other units and contexts. The questions of external validity are about the generalization of the causal relationship, not only to different units or settings that participated in the experiments, but also to other persons, treatments or outcomes that were not involved in the experiment. For instance, the policymaker may be interested in the extent that the program for reducing poverty in a village located in a certain province will be effective in other poor locations; or whether the training program for random selected adults that proved to raise income in a certain percent will have the same effects on younger volunteers. We will see that some sources of causal inference provide estimates that are more generalizable than others.

It is clear that the threats to generalizing the causal relationship to various units, treatments, outcomes and settings arise when there is an interaction between the causal relationship and the varying factors. The interactions threatening external validity may be both those who prove to be statistically significant, but also the ones that are theoretical or practical.

Internal Validity refers to whether the observed correlation between two variables –the presumed treatment and the presumed outcome- is only the result of a causal relationship from the treatment variable to the outcome one. To support the inference, we should be able to show that the presumed caused preceded the effect chronologically, and the effect could not be the result of any other factor than the presumed cause. There are threats to this claim about causal knowledge: i) *Ambiguous Temporal Sequence:* we cannot be sure that the presumed caused occurred before the effect; ii) *Selection into Treatment and Control Groups*: this the problem of confounders, i.e. in the presence of another variable affecting both outcomes and selection to treatment, we cannot isolate the causal parameter; iii) *Attrition:* units belonging to the original sample in an experiment leave it before the measurement of outcomes; iv) *Other sources:* any other factor that affects outcomes and is not the treatment is a threat to internal validity. For example, if we change the measurement instruments, or if the way in which we test outcomes affects them.

As regards the sources of causal inference, we will first focus on social experiments, in which the researcher assigns units to a treatment group (which will receive an intervention) or to a control group (that will not be intervened) randomly or by chance (for example, by the toss of a coin or using a table of random numbers). When random assignment is correctly implemented, it generates

groups of units that are probabilistically similar to each other on average, except for the treatment status.

Another source of data is observational studies, in which the assignment to treatment is decided by the units studied. For the researcher selection to treatment is given and he only has to watch what happens and try to design a way of isolating causal relationships. In Rosenbaum (2002) words: an observational study is an empiric investigation of treatments, policies, or exposures and the effects they cause, but it differs from an experiment in that the investigator cannot control the assignment of treatments to subjects.

With observational (non-experimental) data, the main problem while making causal inferences is confounders. When we try to separate the cause of some effect, we find that in non-experimental studies there are so many variables interacting, that identifying a causal relationship is not straightforward.

Observational studies can sometimes be susceptible of a *quasi-experimental design,* when the researcher can manipulate the treatment, but the selection into groups is still given (by the units affected, laws, bureaucrats, teachers, nature, or whatever).

Apart from controlled experiments and observational studies, there is a third type of source of data for causal inference: natural experiments. They are observational studies where assignment to treatment or control is "*as if*" randomized by nature. That is to say, that the researcher does not manipulate the context to generate a "treatment" and a "control" group; but based on reasonable grounds he can claim that the assignment of subjects to the groups is random or "*as if*" random.

A natural experiment can be a naturally occurring event such as an earthquake, a flood; or a policy change (a new law) that has nothing to do with the outcome we evaluate but it can be used to contrast the event with a comparison condition. So, natural experiments describe a naturally-occurring contrast between a treatment and a comparison condition.

In conclusion, randomized controlled experiments are the ones that minimize the problem of confounding. However, though they seem ideal, they are not always capable of being held for different reasons and they could have other weaknesses we will study in the next sections. With observational studies, there should be a rigorous study of all the variables interacting in order to make sure that confounding variables does not impede the identification of the causal relation we look for. We will deal with the pros and cons of randomized experiments and different types of quasi-experiments throughout the Chapter.

*The Potential Outcomes Model*

The model commonly used in social sciences to study causal relationships was formalized by Donald Rubin (1974) after the original work of Neyman on experimental designs (1923). The Neyman-Rubin model is usually referred to as the potential outcome or the counterfactual model.

To illustrate the model, consider our example of the job-training program for a certain target of adults in a determined city (population of interest). Each person can volunteer to receive the program (constituting the treatment group) or not (control group). After the program is held, we are interested in the effect of it on earnings. The key issue is that any individual of the population can potentially belong to any of the two groups, as any individual can potentially be exposed to treatment.

Consequently, if we denote the treatment status $T_u$ (binary variable equal to 1 if the unit is treated and to 0 otherwise), $Y_u$ is the outcome of interest of the individual unit $u$, $\varepsilon$ is the unobservable error term and $\beta$ is a constant, after treatment takes place, there are two potential outcomes:

$Y_{u1} = \beta + \delta_u + \varepsilon_u$ if $T_u = 1$

$Y_{u0} = \beta + \varepsilon_u$ if $T_u = 0$

So, each unit $u$ in the population of interest has a theoretical potential outcome under each possible treatment status. The individual-level causal effect of the treatment is defined as the difference between the potential outcomes:

$\delta_u = Y_{u0} - Y_{u1}$

As a single individual cannot belong to the treatment group and the control group, one of the potential outcomes will be observable, whereas the other will be unobservable (counterfactual). In terms of our example of the training program, we will know what the outcome of a man who participated in the job training program is, but we will never know what his earnings would have been, hadn´t he participated. Then, the observable outcome can be expressed as:

$Y_u = (1 - T_u) Y_{u0} + T_u Y_{u1}$

So that $Y_u = \beta + \delta_u T_u + \varepsilon_u$

Notice that the effect of the treatment can be heterogeneous in the population, so that different groups of individuals, for example those who enrol in the job-training program (treated) and those who do not (untreated) may get different gains from the treatment. In fact, there is a whole distribution of the treatment effect. The methods we will develop in this Chapter, aim at estimating some feature of that distribution. For example, the mean of the population distribution of $\delta_u$, is the Average Treatment Effect (ATE):

$ATE = E[\delta_u]$

Another parameter of interest, in the frame of heterogeneous effects, is the Average Treatment Effect on the Treated. This parameter is sometimes more important than the effect of the treatment on an individual taken randomly from the whole population. The reason is that if we consider, for example, the job training program targeted to some group of the population, it does not really matter the effect of that program for people that are not likely to receive it (high skilled workers with stable job, for instance). Therefore, for policy evaluation, most of the times, it is more important to have an estimate of ATOT than of ATE.

$$ATOT = E[\delta_u | T_u = 1]$$

Notice that when we referred to the outcome of unit *u* under any of the treatment status, we did not even mention what the treatment situation of the other units was. Implicitly, we assumed that treatment of unit *u* affects only the outcome of unit *u*. This assumption is called the *Stable Unit Treatment Value Assumption* (SUTVA). In the words of Rubin (1961), SUTVA is simply the a priori assumption that the value of Y for unit *u* when exposed to treatment *T* will be the same no matter what mechanism is used to assign treatment *T* to unit *u* and no matter what treatments the other units receive.

SUTVA is violated when there are general equilibrium effects, externalities or spill over of treatments. For example, in the case of the effect on wages of a job training program in a small city, we may think that as more people receive the treatment if the number of jobs is fixed, there will be an upwards pressure on wages. If this is the case, wages of non-treated will also be affected by the treatment. In the case of a large city, SUTVA would be more reasonable because the wage structure will not be affected by a small program.

Another situation in which SUTVA does not hold is in the case of externalities. Miguel and Kremer (2004) study the case of the presence of externalities in medical treatments against intestinal helminths, which infect more than one-quarter of the world's population with severe consequences in the children's health and in turn affect their as school attendance. The authors claim that studies in which medical treatment –for illnesses with high re-infection rate- is randomized at the individual level potentially doubly underestimate the benefits of treatment, missing externality benefits to the comparison group from reduced disease transmission, and therefore also underestimating benefits for the treatment group. They evaluate the Primary School Deworming Project (PSDP) in the Busia district of rural Kenya, in which a mass deworming treatment was randomly phased into rural primary schools with high prevalence of worm infections. The treatment was randomly allocated to schools, rather than to individuals, allowing estimation of overall program effects. Schools were randomly assigned to one of three groups of 25 schools each. Group 1 was given treatment starting from 1998, while Group 2 was incorporated in 1999 and Group 3 in 2001. Schools with infection rates over 50 percent were mass treated with the appropriate drug regime.

The authors want to identify the total effect of the program on treated and untreated individuals. This means it is necessary to identify three effects: the effect of the treatment on treated children; the effect of the intervention on untreated children in treatment schools; and the effect of the intervention on untreated children in control groups. Thus, there are two external effects that need to be identified: within-school and cross-school externalities. As regards the first external effect, it

is measured by the difference in outcomes between untreated students in treatment schools and students in treated schools. Since treatment assignment within a school was not assigned randomly, there cannot be an experimental estimate. The claim made by the authors is that the group who did not take up treatment in Group 2 schools in 1999 provides a good counterpart to the subset of Group 1 pupils who did not take up treatment in 1998 (noncompliers in 1998 and 1999 are very similar in terms of their pre-treatment characteristics). In order to account for the second externality (cross-school), the authors exploit the variation in the density of children receiving treatment in nearby schools. If there are any schools close by, we would expect individual worm loads to be higher, because of higher possibilities of reinfection. This effect would depend on the number of students in those schools, and on their distance.

The authors conclude that regression estimates show a large impact of the program on infection rates. The program has the direct effect of reducing the incidence of moderate-heavy infections by 25 percentage points (p.p.) The estimated externality is high as well, so SUTVA would not hold, with infection rates being 26 p.p. lower for every 1000 students in treatment schools within 3 km, and 14 p.p. for every 1000 students in treatment schools between 3 and 6 km away. When breaking down the impact in treatment schools, the direct effect of treatment is a 14 p.p. reduction, while the within-school externality accounts for a further reduction of 12 p.p. They find a large and statistically significant effect of the program on school participation especially for younger children (increase of 10 p.p.). Overall, the authors estimate that treating one child led to an increase in school participation of 0.14 years. The authors find no change in school performance associated with the program.

*Alternative Frameworks: Potential Outcomes Model and Structural Models*

As we mentioned, although we cannot recover the individual level treatment effect, we can estimate some features of the distribution of the treatment effect in the framework of the potential outcomes model. One main advantage of this setting is that it enables the definition of the causal effect without specifying the assignment mechanism and does not need any functional form or distributional assumption.

In contrast, there is another approach first developed by geneticists (Wright 1921) and economist (Haavelmo 1943), that consists on writing down structural models, which give functional forms to all the decisions involved in the problem we are analyzing and as a result allow us to identify all the relevant parameters of the model and the distribution of heterogeneity.

Structural equation models rely on the specification of systems of equations with parameters and variables that attempt to capture behavioural relationships and specify the causal links between variables.

Consider a model M of three equations:

$z = f_z(w)$

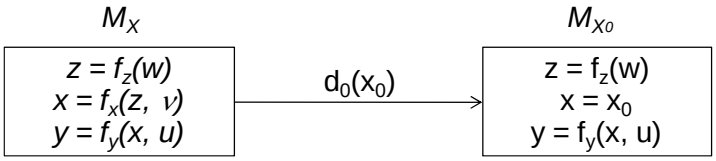
$x = f_x(z, \nu)$

$y = f_y(x, u)$

Notice that y is a function of *x* and *u*, but at the same time, x is a function of v and z which depends on w. Finally, outcome y depends on (u, v, w). We will try to find an expression of the effect of a certain intervention on y.

Interventions are represented through a mathematical operator denoted $d_0(x)$. Here $d_0(x)$ simulates physical interventions by deleting certain functions from the model, replacing them by a constant X = x, while keeping the rest of the model unchanged. In consequence, model M after the intervention is $M_x$, as shown in the following diagram:

*Introducing an intervention in a Structural Model M*




How can we link the introduction of an intervention in a Structural Equation Model with the Potential Outcomes Model?

Consider an individual with characteristics (u, v, w). We can analyze the potential outcomes of variable *y* for the two alternative specifications of X:

$$X \begin{cases} x = f_x(z, \nu) = f_x(f_z(w), \nu) \Rightarrow Y_x(u, \nu, w) = Y_x(u) = f_y(x, u) \\ x = x_0 \Rightarrow Y_{x0}(u, \nu, w) = Y_{x0}(u) = f_y(x_0, u) \end{cases}$$

This entity can be given a counterfactual interpretation, for it stands for the way an individual with characteristics (u, ν, w) would respond, had the treatment been $x_0$, rather than the $x = f_x(z, \nu)$ actually received by the individual.

This interpretation of counterfactuals, cast as solutions to modified systems of equations, provides the conceptual and formal link between structural equation modelling and the Rubin potential-outcome framework. It ensures us that the end results of the two approaches will be the same. Thus, the choice of model is strictly a matter of convenience or insight.

To choose between them, the researcher must take into account the estimation issues concerning each model. When researchers in political science approach data, the counterfactual model approach appears more attractive if they do not know how to specify precisely the sources of randomness in the data or the relationships that exist between these variables. That is, the structural approach

requires the researcher to be precise about the determinants of potential choices and treatment variables and the relationships between these in real datasets.

However, the structural modelling framework is richer because if we can assume that we know all the functional forms of the model, we can estimate all the structural parameters, whereas in the potential outcome model we would get a reduced-form model. To illustrate this point, consider a multi-equational model in which we can claim that there are direct and indirect effects of a variable *t* in our outcome variable Y.

For example, consider the following model in which outcome Y is a function of treatment t and variable x.

$Y = f(x, t)$

If we want to identify the effect of treatment *t*, we just have to find what happens to *Y* once we control x and change t (either using derivatives or finite changes). But, if simultaneously, we know that *X* is caused by the treatment variable *t*, we now have a system of two equations whose variables *x* and *y* are jointly determined.

$Y = f(x, t)$

$X = g(t)$

If *x* were determined outside the model or controlled in any way, we could hypothetically vary *t*, we will be in the same situation as before (the case of one equation). When we take the model in it structural form, we can see that variations in *t* have direct effects on *Y* and on *X*, which are called structural causal effects. But is we introduce changes in *t*, there are also indirect effects on *Y* through *X* (that is determined by *t*).

To obtain the total effect of variations in *t*, we can express the structural model in its reduced-form, which implies that every endogenous variable (those determined within the model) depends only on exogenous variables, parameters or constants (not on other endogenous as allowed in the structural form). The reduced-form of the previous model would be:

$Y = Y(t)$

$X = X(t)$

For example, if we want to determine the effect of going to pre-primary school (t) on earnings (Y), we may think that attending pre-primary school (*t*) and the labor supply of the mother (X) affects earnings (Y), which constitute the first structural equation. But attending pre-primary school also affects the labor supply of the mother of the individual (X), represented by the second structural equation. Thus, pre-primary education has a direct effect on earnings and an indirect one through its effect on the labor supply of the mother. In the structural model we can estimate all the direct and indirect effects, while in the reduced-form model we can only get the total effect.

*Estimation of the Treatment Effects*

Notice that the treatment effect can only be unbiased estimated if we assume that selection into treatment is independent on the potential outcomes of the unit. Or in other words, treatment assignment is "ignorable". Let´s start from the definition of ATE:

$$ATE = E[\delta_u] = E\left[Y_{u1} - Y_{u0}\right] = E\left[Y_{u1}\right] - E\left[Y_{u0}\right]$$

which can be unbiasedly estimated when we assume that treatment assignment is "ignorable". What is equivalent to assuming that *Y* and *T* are statistically independent for *u=1,…N*.

$$Y_{u1}, Y_{uo} \perp T_u$$

Under this assumption, a consistent estimator of the average treatment effect is simply the difference in the sample means of the observed outcome variable *Y* for the observed treatment and control groups.

$$\hat{\bar{\delta}}_{ATE} = \left[\hat{\bar{Y}}_{u1} \mid T = 1\right] - \left[\hat{\bar{Y}}_{u0} \mid T = 0\right]$$

The ignorability assumption is very strong. Imagine that we are assigned the task of evaluating the impact of a job-training program which is supposed to increase the future earnings of its participants. After the program, we will have information about the earnings of the participants and we could have a sample of non-participants earnings. Does the difference of the mean earnings of participants and non-participants estimates the effect of the program? What effect does that estimate captures? To answer the question, we must know how selection into the program was done, and how the control group was chosen. It is natural to think that those who applied for the program have on average, a lower income than those who did not. So there is a baseline difference between the pre-treatment incomes of both groups. This means that a simple mean difference estimator will not capture the effect of the treatment on the treated. We can also imagine that those who applied for the program had a poorer educational attainment. Therefore the factual future earnings of the control group will be far from the counterfactual post-treatment earnings of the participants in the absence of treatment. The post-program earnings of the (often more qualified) non-participants will be probably higher than the future income of the participants hadn´t they participated in the program. Then, the potential gains from the program will be higher for the applicants than for the more-qualified non-applicants, which mean that the effect of treatment is heterogeneous, so we are in the presence of selection on the potential gains. Then, ATE will not be consistently estimated by the previous expression (it will be underestimated), even if we consider an homogenous effect for treated and non-treated. In addition to this, if we assume that there is treatment heterogeneity, we may think that those who volunteer are the ones whose expected gains from the treatment are higher, so there will be a selection on potential gains that will also make the estimation of ATE inconsistent.

To clarify the selection problem underlying in the estimation of the effect of treatment, let's consider the linear model again:

$Y_u = \beta + \delta_u T_u + \varepsilon_u$

Note that the previous expression is slightly different from the linear regression model set-up, because our causal effect parameter has an individual subscript (we allow for heterogeneous treatment effect). We can re-write this expression in terms of the homogenous effect ATE:

$Y_u = \beta + \delta^{ATE} T_u + \varepsilon_u + \delta_u T_u - \delta^{ATE} T_u = \beta + \delta^{ATE} T_u + [\varepsilon_u + \delta_u T_u - \delta^{ATE} T_u] =$

$Y_u = \beta + \delta^{ATE} T_u + v_u$

$v_u$ contains all the unobserved terms. In this framework, the OLS estimation of $\delta^{ATE}$ will be consistent if the assignment rule determining $T_u$ is uncorrelated with the new error term. If we consider that the assignment mechanism is partially deterministic (i.e, Tu is determined by a set of observable variables W and by an error term ω), then, the treatment assignment can be endogenous due to a selection problem on observables (W) and/or on unobservable variables (ω). Depending on the assumption of the type of endogeneity, there are different ways of estimating the causal effect.

Usually, we tend to assume that selection occurs on unobservable variables, so in this Chapter we will provide different approaches to deal with the fundamental problem of causal inference in that case. On the other hand, if we could assume that selection was done on observable variables, we could apply Matching techniques that rely on the Conditional Independence Assumption (CIA), which means that although we cannot assume that treatment is ignorable, on reasonable grounds we do claim that conditional on a certain set of observable variables X, the treatment assignment mechanism is independent of the distribution of the potential outcomes:

Conditional Independence Assumption: $Y_{u1}, Y_{uo} \perp T_u / X_u$

However, claiming that there is only selection on observable variables is a very strong assumption and in many cases not realistic. Thus, we will formalize what OLS can identify in a setting where we could have both selection on observables and unobservables.

More formally, if we obtain the expression for $\hat{\delta}^{OLS}$ and take its probability limit, we get that the OLS estimator will identify:

$\delta^{OLS} = E[Y_{u1} | T_u = 1] - E[Y_{u0} | T_u = 0] =$

$\delta^{OLS} = \delta^{ATE} + E[\delta_u - \delta^{ATE} | T_u = 1] + E[\varepsilon_u | T_u = 1] - E[\varepsilon_u | T_u = 0] =$

$\delta^{OLS} = \delta^{ATOT} + E[\varepsilon_u | T_u = 1] - E[\varepsilon_u | T_u = 0] =$

$\delta^{OLS} = \delta^{ATOT} + E[Y_{u0} | T_u = 1] - E[Y_{u0} | T_u = 0]$

In the last expression we collected the first two terms of the previous one, to obtain ATOT. Then, we can see that the OLS estimator will only recover ATE when there are no selection on idiosyncratic gains (i.e., the treatment effect is homogeneous so ATOT = ATE) and there is no selection bias due to baseline differences (usually called selection on non-treated outcomes). Thus, OLS will only recover ATE if there is no endogenity of the treatment, and will recover ATT if there is selection on idiosyncratic gains but in the absence of correlation between ε and T.

*Randomized experiments*

Consider the hypothetical situation in which you, as a researcher, have some control over the program and the treatment assignment mechanism. How would you assign treatment so as to estimate unbiasedly the average treatment effect (ATE)? The answer is that there are good reasons why you would want to allocate the treatment T at random. This is what is known as a randomized controlled experiment. Random assignment is achieved by any procedure that assigns units to conditions based only on chance, in which each unit has a nonzero probability of being assigned to a condition. The use of a randomized experiment constitutes a possible solution to the problem of selection bias in an evaluation. Here we have the greatest assurance that the program participants and the control group of program eligible individuals are on the average alike in every important sense (including observable and unobservable characteristics) except that one group has had the program and the otherwise probabilistically "identical" group has not.

Social experiments are the most similar analogue in social sciences to the clinical trials in medicine. In clinical investigation, once the group of patients required for a certain trial is gathered (for example males of certain age suffering hypertension), the preferred way of assigning participants to control and intervention groups is randomization. It tends to generate groups comparable with respect to known and unknown risk or other prognostic factors and also removes the selection bias induced by the investigator if he assigns treatments. Participants should be aware of what their chances are of receiving either placebo or the alternative treatment, but they do not know their actual treatment status.

Even in a clinical trial, although random assignment to treatment groups removes selection bias, there are other sources of potential bias that can arise due to conscious, subconscious factors, or both. It can occur at several places of the trial: from the design to the data analysis and interpretations. For example, the investigator can treat differently those patients potentially benefited with the new drug from those getting placebo, perhaps complementing the treatment of those in the control group. Or he could report the results in a different way influenced by pre-concepts about the experimental treatment. A general solution for the problem of bias is to keep the participant and the investigator blinded or masked; meaning that neither knows which treatment the patient is actually receiving. The ideal clinical trials are randomized and double-blinded.

In social sciences randomized experiments are not always feasible, due to ethical or practical reasons, or for policy reasons because the government may not agree to exclude a certain group from a policy or may target the policy on purpose to a determined non-random sub-population.

Furthermore, there does not exist an analogue for the double clinical trial, because in a training program, for example, there is a point in time in which the participant knows if he was assigned to treatment or control group. In addition to this, take into consideration that proponents of randomization implicitly assume that randomization does not alter the program being studied, so we assume that the decision of the participants are not affected by the treatment assignment. Many programs have multiple stages (enrolment, assignment to treatment, promotion, review of performance, or placement stage), and the randomization stage should be the least disruptive, which is hard to determine.

To illustrate the use of randomized experiments to evaluate the effects of programmes, consider the paper by Banerjee et al (2007) in which two randomized experiments in India are analyzed to evaluate the impact of interventions to improve educational outcomes. The programs took place in two of the most important cities in Western India. In both cities children and teacher attendance was high, but the students' results in the tests were poor. Therefore these interventions aim at changing what children learn while in school and consist on providing supplementary inputs to children at the bottom of the class. The first of the remedying intervention consists on assigning a young woman from the community to work on basic skills with children who have reached grade three or four without having mastered them. These children are taken out of the regular classroom to work with this young woman for 2 hours per day (the school day is about 4 hours). The second intervention is a computer-assisted learning program, where children in grade four are offered two hours of shared computer time per week, during which they play games that involve solving math problems. Both programs were allocated using random assignment across the same set of schools and were run on a very large scale (over 15,000 students affected over 3 years).

For the first intervention, children in grade three in schools that received the program for grade four form the comparison group for children that receive the program for grade three, and vice versa. Whereas, the second program was implemented in half of the municipal primary schools focusing exclusively on children in grade four. The authors found that the programs were very successful. The remedial education program increased average test scores in the treatment schools by 0.14 standard deviations in the first year, and 0.28 in the second year. Moreover, the weaker students, who are the primary target of the program, gained the most. The results suggest that the untreated children in treated grades also improved slightly, but mainly because the weakest were taken out from the class and not because of the adding of new resources (reduce of class size). The computer-assisted learning increased math scores by 0.35 standard deviations the first year, and 0.47 the second year, and was equally effective for all students.

Now, let's study the implications of randomization more formally. Consider the potential outcome framework. If we randomly assign units to treatment and control groups then we should expect no difference between the potential outcomes of those who finally where assigned to each condition. Then:

$$E\left[Y_{u1}\middle|T_u=1\right]=E\left[Y_{u1}\middle|T_u=0\right]=E\left[Y_{u1}\right]$$
$$E\left[Y_{u0}\middle|T_u=1\right]=E\left[Y_{u0}\middle|T_u=0\right]=E\left[Y_{u0}\right]$$

So, if we return to the definition of ATE and ATOT, under randomization, we see that they are equivalent:

$ATE = E[\delta_u] = E[Y_{u0} - Y_{u1}] = E[Y_{u0}] - E[Y_{u1}]$

$ATOT = E[\delta_u | T_u=1] = E[Y_{u0} - Y_{u1} | T_u=1] = E[Y_{u0} | T_u=1] - E[Y_{u1} | T_u=1] = E[Y_{u0}] - E[Y_{u1}]$

That the estimation of ATE is equal to that of ATOT under randomization is quite intuitive. ATE differs from ATOT only when there is treatment heterogeneity. Randomization allows us to expect no treatment heterogeneity, so we should also expect that both effects were equal. Stated differently, since we randomized units to treatment and control, there is no reason why the treatment effect should differ between individuals that were assigned to each condition.

In the linear model specification, this is the same as assuming that:

$E[\varepsilon_u | T_u = 1] = E[\varepsilon_u | T_u = 0]$

$E[\delta_u] = E[\delta]$

So that $\delta^{OLS} = \delta^{ATE} = \delta^{ATOT}$

These can be easily estimated by their sample counterpart:

$$\hat{\delta} = [\hat{Y}_{u1} | T_u = 1] - [\hat{Y}_{u0} | T_u = 0]$$

Then, random assignment equates groups on the expectation of group means at pretest. That is, on the mean of the distribution of all possible sample means resulting from all possible random assignments of units to conditions. In principle, randomized trials ensure that outcomes in the control group really do capture the counterfactual for a treatment group. In practice we have only one drawn from the population, so we may or may not get ATE. So, what are the benefits of randomization? The advantage of randomization is that we can expect to obtain the true ATE. As we only have one draw from this distribution, either if we randomize or not, we might end up with a selected sample. But the main difference is that if we do not randomize, the assignment to treatment and control might be systematically correlated to the potential outcomes, and biases will be systematic. In the case of randomization, biases will result of bad luck, i.e, we obtained a random sample with an estimated ATE that has a low probability of materializing (estimates from samples placed in the tails of the distribution).

In most empirical studies, researchers are allowed to observed not only the treatment assignment and the outcome measures, but also some other variables that might be correlated or might explain to some extent the outcome variable. These variables are called covariates.

If these variables affect the potential outcome of the units, then they may act as confounders when we are trying to estimate the treatment effect. Additionally, using these variables we can sharpen our prediction about the baseline potential outcome of each unit and consequently reduce the residual error. When the covariates turn out to be closely correlated with the treatment assignment, then it would be hard to distinguish to what extent the differences in the observed outcomes between treated and control units are due to the treatment effect or due to the effect of the covariates. If the assignment to treatment and control was random, we have reason to expect that this assignment will not be correlated with the covariates.

However, if there are many covariates, there is a high chance that the assignment would end up being correlated with at least one covariate. If this happens, we say that the covariates are unbalanced between treatment and control. In this case, the precision of our estimator of the treatment effect would be low, and the power of our tests will also be low. On the other hand, if these covariates explain to an important extent the baseline potential outcomes, we will be able to construct a better counterfactual for treated observations, the variance of the estimator will fall and our tests will be more powerful.

At the experiment design test, we would like to take advantage of the existence of variables that sharpen our prediction about the counterfactuals, but we will be willing to avoid an unbalanced sample. We can add covariates, which are observable variables that affect the outcome measured. Including covariates the lineal model can be written as:

$$Y_u = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ju} + \delta T_u + v_u$$

Where $X_j$ denotes de j-th covariate and $\beta_j$ denotes the effect of this covariate on the outcome measure. For each surveyed unit we will observe the outcome measure $Y_u$, the value of the $k$ covariates and whether this unit was treated or not. With this information we can estimate consistently the parameters for the covariates and for the treatment by OLS, under the assumption that all the explanatory variables are exogenous.

Including the covariates in the analysis reduces the variance of the unexplained residual term. In fact, the regression estimator will be comparing the outcomes of treated and untreated individuals within each group. As a result, the estimator of the causal effect will be more precise and power will be higher. The inclusion of covariates helps to reduce the empirical variance of the estimator, sharpening predictions about the treatment effect.

Now, suppose that the random assignment of unit to treatment and control is such that the treatment ended up being highly correlated with the groups. For instance, too many individuals from the first group were assigned to treatment and few to control and the opposite occurred in the other group. Thus, when the treatment assignment ends up being correlated with the covariates, the result is a dampening in precision. If treatment is assigned randomly, we can expect no correlation with covariates in the same way that we expect no correlation with unobservables. An unbalanced sample is the result of bad luck in the same way that a selected sample is. However, there is an important difference. We will never know for sure whether our treatment assignment was or not

correlated with unobservables. At most, we know the probability that our estimated effect were the result of sheer chance, because of obtained a selected sample. On the other hand, we will observe whether we obtained an unbalanced sample or not.

Furthermore, we could design the experiment in order to avoid unbalanced assignment and increase precision. Suppose that we observed covariates before we randomized units to treatment and control. Moreover, assume that the only covariate we observe is belonging or not to a group, for example workers that belong to different firms. Then, we could randomize individuals within each group to treatment and control. We would ensure that the sample is balanced on observable covariates and at the same time we will be able to take advantage of the reduction in the residual error resulting in increased precision. This technique is called *stratification* and was first proposed by Fisher (1926). Stratification involves dividing the sample into groups sharing the same or similar values of covariates. Then randomize within each group, to ensure that treatment and control groups are comparable.

Under simple randomization there is a positive probability that we obtain an unbalanced sample. In this state of nature, the estimated variance of the OLS coefficients estimators is high since the sample does not allow us to disentangle the independent covariate and treatment effects. On the other hand, stratification avoids getting an unbalanced sample and; consequently, we can expect lower variance.

There are several ways to build the distinct groups and assign units to treatment and control. Each strata may contain the same proportion of treated and untreated units; in this case the average treatment effect estimator is equal to the difference between the outcomes of treated and untreated units in each strata averaged across strata. Alternatively, each strata may contain 2 units, one assigned to each condition. This strategy is called matching and the average treatment effect is estimated as the average across strata of the difference between the treated and the control observation. There are situations of matching with multiple controls, where each strata contains one treated units and several untreated. This strategy is useful when providing treatment to units is too costly.

So far, we have introduced randomization of units into treatment status, in which the randomization level was the single units we observed. In fact, we showed that under randomization, each unit has the same probability of being assigned to any treatment and we assumed that the observations were independent. We call this type of randomization at unit-level *Simple Randomized Trials* (SRT). Alternatively, we could gather units into groups so that each unit belongs to one and only one group, and randomize over groups of units. This is the case called *Group Randomized Trials* (GRT).

GRT have two distinct features: 1) each unit of observation is a member of an identifiable group; and 2) entire groups are randomly assigned to treatment and control conditions. For instance, in-job-training programs can be evaluated assigning firms to conditions and surveying the employees working at those firms. Entire groups are assigned to each condition but data is obtained from the members of those groups. Notice that GRT can arise on purpose and for other reasons. Sometimes the set up of the experiment requires that the units are gathered in certain groups of interest, or the social experiment is held so as to affect only particular locations, schools, firms, etc. Or GRT can

allow costs-reduction; for instance, when we have a large-scale household survey, it is common to sample groups of houses in the same neighbourhoods or blocks, and thus, this observations will no longer be independent within each neighbourhood or block; there will be spatial correlation.

What it is important to note is that while assignment to conditions is random, group forming is not. Groups were formed before the intervention and there is certainly some connection among their members. Due to these connections, the methods applied to calculate intervention effects on randomly assigned units should be modified to allow for this particular assignment procedure. In GRT there are three sources of variability in outcome measures:

1. Group effect: all members belonging to a group share a common error term that captures any group characteristic unobservable to the researchers.
2. Individual error term: captures individual unobservable characteristics.
3. Treatment effect: due to different assignment to conditions.

Our goal is to estimate treatment effect, so we would like to reduce to the maximum sources 1 and 2, so as to be able to isolate 3. The key difference between trials in which groups rather than units are randomly assigned to conditions is how the variance of the mean difference estimator behaves as sample size grows. In a simple trial, as the sample size grows, the first two sources of variance tend to vanish. Individual error terms cancel each other as there are more sampled individuals. Besides, since we are assigning into each condition members of all groups, group effects will cancel out across conditions. We would expect that for any given group, the same number of individuals were assigned to each condition. In our examples of firms randomized to a training program (instead of workers), if the sample size increases in a SRT, there will be a lot more of employees of all the firms in each of the two groups, therefore both samples –treatment and control- will be balanced in terms of group-effects; and in each group the increase in the amount of employees will lead to a neutralization of individual errors between each other (averaging zero).

However, in a GRT, as the number of members per group grows, we can sharpen our prediction about the group mean effect. We would get more and more observation of units belonging to a particular group. Within each group, individual error terms will tend to cancel each other. Consequently, the variance due to the individual error term tends to zero. But, as we are not increasing the number of groups per condition, the component of the variance attributable to groups is not vanishing.

To understand the effect of increasing the number of members per group in a GRT, we will consider this extreme case. Imagine the case that the variance attributable to individuals within groups is zero. Then, the outcome of every individual within each group will be exactly the same, and equal to the mean outcome in group. In that case, it would not matter how many individuals we sample from each group: once we observed one, we will know the outcome of every other member of the group. However, we cannot make a perfect estimation of the treatment effect because of the unobserved group effect. This same result will be obtained if we sampled an infinite number of members per group. We will not be able to increase precision unless we sample more groups. Increases in the number of members per group will not help to reduce the noise introduced by group effects.

For all the above mentioned, we see that in the presence of some kind of relation between the members of the same group, assigning groups rather than units to treatment, leads as to an estimator of the treatment effect with a greater variance, and thus lesser precision.

What are the implications of this result? There are two important implications. Firstly, since the variance introduced by the group-effect is reduced only if we assign units –rather than groups- to condition; for a given sample size, our estimator will be more precise when we perform SRT than when we assign groups (GRT). Secondly, if these group-effects were all zero, units belonging to the same group would not be correlated between each other. We could think of groups as being randomly formed, and we could apply the same methods we applied for SRT to estimate the treatment effect. However, in most GRT units belonging to the same group are correlated; the correlation among members within a group is called *intra-class correlation (ICC).*

Why do we expect ICC to be different from zero? There are many reasons. Mainly, consider that subjects frequently select the group to which they belong, so this is a hint that any pattern of auto-selection has an underlying reasons grounded on individual characteristics. In additions to this, all the individuals belonging to the same group or cluster (another name for group) share a common environment; so they are exposed to the influence of the same cluster-level variables. Finally, individuals in the same group are affected by personal interactions among cluster members.

Formally, we can think of the outcome measure of individual unit *u* belonging to group *k*, assigned to condition *l* as:

$$Y_{u:k:l} = \mu + \delta T_k + G_{k:l} + \varepsilon_{u:k:l}$$

Where $\mu$ is the population outcome mean, δ is the treatment effect, and $T_k$ is 1 when group k is assigned to treatment and 0 if it is assigned to control. All members belonging to a group are assumed to share some common characteristic unknown and unobservable to the researcher, $G_{k:l}$ stands for the total effect of such commonality on the outcome variable, it is assumed to have zero mean. Finally, $\varepsilon_{i:k:l}$ is the residual individual term which is also assumed to have zero mean. Zero mean assumptions are made without loss of generality since we are including a grand mean. Finally, μ and δ are parameters to be estimated while $G_{k:l}$ and $\varepsilon_{u:k:l}$ are random effects or unobservable error terms.

The variance of individual observations is:

$$V\left(Y_{u:k:l}\right) = \sigma_y^2 = V\left(G_{k:l}\right) + V\left(\varepsilon_{u:k:l}\right) = \sigma_{g:c}^2 + \sigma_e^2$$

Where $\sigma_y^2 = V\left(Y_{u:k:l}\right)$, $\sigma_{g:c}^2 = V\left(G_{k:l}\right)$ and $\sigma_e^2 = V\left(\varepsilon_{u:k:l}\right)$.

These are the two sources of variances that introduce noise in our treatment effect estimation: the group effect and the individual error term, respectively.

Consider first the group mean: that is, the mean of all observations belonging to group *k*.

$$\overline{Y}_{k:l} = \mu + \delta T_k + G_{k:l} + m^{-1}\sum_{i=1}^{m} \varepsilon_{i:k:l}$$

Where m is the number of members per-group which is assumed to be the same across groups to keep notation and analysis simpler. Then, the variance of the group k mean will have two components: the variance of the group effect, and the variance of the individual effect. If we sample more units from group k we will observe more realizations of the error term $\varepsilon_{i:k:l}$. However, no matter how many observations we get we will only observe one realization of $G_{k:l}$. As the sample size grows, the first component remains unaltered while the second tends to vanish.

$$V\left(\overline{Y}_{k:l}\right) = \sigma^2_{\overline{y}_g} = V\left(G_{k:l}\right) + m^{-1}V\left(\varepsilon_{i:k:l}\right) = \sigma^2_{g:c} + \frac{\sigma^2_e}{m}$$

The condition means are estimated as the weighted average of g-group-specific means.

$$\overline{Y}_l = \mu + \delta D_k + g^{-1}\sum_{k=1}^{g} G_{k:l} + g^{-1}m^{-1}\sum_{k=1}^{g}\sum_{u=1}^{m} \varepsilon_{u:k:l}$$

Where $g$ is the number of groups per condition. Since groups were assigned randomly to each condition, the variance of the condition mean is:

$$V\left(\overline{Y}_l\right) = \sigma^2_{\overline{y}_c} = \frac{m\sigma^2_{g:c} + \sigma^2_e}{mg}$$

Finally, the mean difference estimator is the difference between the estimated condition means.

$$\hat{\delta} = \delta + g^{-1}\left(\sum_{k=1}^{g} G_{k:T} - \sum_{k=1}^{g} G_{k:C}\right) + g^{-1}m^{-1}\left(\sum_{k=1}^{g}\sum_{i=1}^{m} \varepsilon_{i:k:T} - \sum_{k=1}^{g}\sum_{u=1}^{m} \varepsilon_{u:k:C}\right)$$

Note that the expected value of the estimator is the true treatment effect since all the remaining terms have zero expected values. Moreover the variance of this estimator is:

$$V\left(\hat{\delta}\right) = \sigma^2_{\hat{\delta}} = 2\frac{m\sigma^2_{g:c} + \sigma^2_e}{mg}$$

This variance tends to zero as $g$ tends to infinity but it tends to $2\sigma^2_{g:c}$ when $m$ tends to infinity. Consequently, the consistency of this estimator is based on the growth of the number of groups per conditions, not of members per group.

The correlation between members belonging to a group is responsible for the loss of the consistency property as *m* tends to infinity. If there were no such correlation, i.e. if $\sigma^2_{g:c} = 0$, the consistency property would hold anyway. Define the *intra-class correlation* (ICC) as the share of total observations' variability attributable to the group-effect:

$$ICC = \frac{\sigma^2_{g:c}}{\sigma^2_{g:c} + \sigma^2_e}$$

Then, the expression for the variance of the mean difference estimator can be re-written as:

$$\sigma^2_{\hat{\delta}} = \frac{2\sigma^2_y}{mg}\left(1 - (m-1)ICC\right)$$

Note that if we could randomly assign each individual to treatment or control, the variance of the estimator would be only:

$$\sigma^2_{\hat{\delta}} = \frac{2\sigma^2_y}{mg}$$

Then, the term $\left(1 - (m-1)ICC\right)$, represents the *variance inflation factor* (VIF) due to the group-level randomization under positive intra class correlation. In fact, when the ICC is equal to zero, the variance of the estimator is the same as under SRT (the last expression). And in this situation, increasing either the number of groups or the number of members per group causes the same effect: the variance tends to vanish.

*If group-level randomization results in higher variance and consequently lower precision, why should we ever rely on it?* There are several reasons to prefer a GRT over a simple randomized trial: it might be cheaper, it might be accurate when a simple trial is not or it might be more practical.

Firstly, it might be cheaper to randomize at group level. Remember that when performing program evaluation studies there is always a budget and time constraint; thus we must allocate the resources in order to assess estimations with the most possible statistical power. In general, providing treatment requires high fixed costs per-group and low marginal costs per additional member in the same group. Therefore, randomizing at the group level -paying the fixed costs only for those groups that are assigned to treatment- might be cheaper than randomizing at the individual level and providing treatment to some individuals in every group. This same argument applies when performing large-scale surveys, as mentioned at the beginning of this chapter.

Secondly, sometimes considering SRT when the outcomes of the units are not independent is not statistically accurate. Remember that we introduced the *Stable Unit Treatment Value Assumption*

(SUTVA) which implies that the potential outcomes of the units do not change when the assignment pattern is modified; or in other words, it means that the outcomes of the individual units are independent of what happens to the other units. But in many contexts the SUTVA could not hold if we randomized at the individual level; and it could hold if we used group-randomization.

For instance assume we want to study the effect of handwashing promotion on diarrhea incidence reduction in low income countries. Suppose that the treatment consists in explaining the treated individuals, the benefits of handwashing in reducing the risk of diarrhea and warning them of how dangerous the disease is for young children. If a household is assigned into treatment and thus, given sanitation instruction, it is probable that its members would tell what they learnt to their neighborhoods and friends. Some households assigned to control might receive the instructions via treated individuals. So, in this case, control units are contaminated; and, under widespread contamination it is hard to discriminate treatment effects. Conversely, if we randomized neighborhoods or villages instead of individuals, contamination of control units might not be as pervasive. Or recall the case of the worms infections we described earlier.

Thirdly, and for practical reasons, in some cases it might be impossible to separate members within a group so that some individuals receive treatment and the rest do not. For instance, in 1995, the National Lung and Blood Institute evaluated an intervention designed to reduce the time in seeking medical attention among patients experiencing a heart attack. The intervention included a mass-media campaign. It would be impossible to limit the delivery of the mass-media messages only to individuals assigned to treatment. Consequently, the only choice available was to allocate whole cities to study conditions.

Finally, for pure ethical reasons it may be unacceptable to provide treatment to some units in a given group or community while other are deprived of it. Sommer et al (1986) randomly allocated 229 villages in Sumatra to intervention with vitamin A supplements and 221 villages to serve as controls. "Political and administrative reasons" were mentioned as the rationale for GRT. Although in principle an individually randomized trial was conceivable, it was not feasible in that setting.

*Setting-up an Experiment*

There are some important elements of a social experiment that should be considered because will affect the estimations and inferences we can get from it. Firstly, we have to select the targeted population very carefully, according to the program we are evaluating, and decide the randomization level and stage. We also have to choose the design.

There is some variety in the designs under randomization that we can choose according to the program we are studying. The basic design involves random assignment of units to treatment and control groups (or to more different treatments and a control group, or only to different treatments which a differential effect is what we are interested in). In additions to this, the basic design requires the post-test assessment of units.

It is of great importance how the control group is defined. It depends on the definition of the treatment and the characteristics that the researcher considers that affect it. For example, in medical experiments about drugs, the control group should be probabilistically equal to the treatment group and must be provided with the same treatment but for the drug. Therefore, in medical experiments, the control group is given placebo instead of the drug. Notice that the control patients do go to hospital and do think that they are being treated. So as providing them with placebo, the *psychological* part of treatment is contemplated, and the only difference between the treatment and the control group is the pharmaceutically active ingredients of the drug. An example of an experiment like this was the Salk polio vaccine in 1954: more than 400,000 children were randomly assigned to receive either the vaccine or a placebo.

Although under random assignment the treatment and control group should be probabilistically similar in all observable and unobservable characteristics, it may be very useful to check that the sample we work with is balanced. For that purpose, one can compare a set of pre-test variables that on average should be similar for the two groups. Ideally the pre-test variables should be both of the post-test. This is what is called a pre-test post-test design which we will develop later. The pre-tests consist of gathering information about the units belonging to the control and treatment group before the treatment is implemented. Sometimes obtaining a pre-test is not feasible because it is either not possible in the context of the treatment studied, or it is too expensive or time demanding. Having a pre-test and a post-test leads us to think again on the fundamental problem of causal inference. If we could observe each treated unit twice, before and after treatment, could we use the first observation as a valid counterfactual for the second? The answer is yes, but only if the baseline outcome were time invariant. Otherwise, we would be confusing a trend effect with the true treatment effect.

In pre-test post-test we added an observation before the treatment. In this way, longitudinal design adds multiple observations before, during or after the treatment, depending on the effect we are trying to identify. In practice it is very rare to have various pre-test and also post-test. It depends on the type of treatment being studied (sometimes it has no sense to follow up units too long). And it is not totally useful to have a lot of post-tests, because the threat of attrition increases with the length of the period considered after treatment. Furthermore, it is probable that the control group will eventually be assigned to treatment (this or any other). In some cases it will not even be ethical to prevent those units from receiving a benefiting intervention.

Until now we considered a treatment as a defined discrete status. However, in some situations there can be treatments with different intensities. For example, instead of attending or not to extra lessons to improve performance in the school, students who take the program can choose between attending to 4-hour lessons or 2-hour lessons. Or treatments can be defined as having different doses of a drug (the control is the group that takes placebo).

Factorial designs provide the possibility of identifying the effect of two or more independent variables (factors), each with two or more levels of intensity. An example of a factorial design is the New Jersey Negative Income Tax. This experiment was held between 1968 and 1972 in New Jersey and then expanded to other locations. This experiment basically consisted of estimating the effect of giving a guarantee income and a negative tax, on the decline of effort. The experiment had a factorial design. There were two factors: the guarantee level that was money given to poor families

that  did not earn any private income; and the marginal tax rate which is the rate at which the benefit was reduced (or income is taxed) as the recipient earned more private income. Simultaneously, guarantee income level was 50%, 75%, 100% or 125% of the poverty level; and the tax rate was 30%, 50% or 70%. So there are 12 categories in which individuals can be included; it is a 4x3 factorial experiment.

There are some advantages of factorial design. Firstly, it often requires fewer units, because each unit will be used to determine the effect of two independent treatments. This happens unless there are special interactions which require a larger sample to attain certain statistical power in the tests. Secondly, it allows testing combinations of treatments easily. And finally, it gives the possibility of testing interactions between independent treatments with different levels. The drawback of this design is that when the number of level and factors grow, an experiment like this is really difficult to control in field settings. In laboratories it is more plausible to implement, so the range of topics where it is often applicable is reduced.

Finally, another common design is a crossover design, which would mean exposing the same individual to two alternative treatments successively, with enough time between the treatments so as the effect of the first dissipates. However, this cannot be generalized to any treatment, because the effect must be ephemeral.

A crucial element in the design of an evaluation study is the determination of the sample sizes needed to reject confidently the null hypothesis of no program impact or, equivalently, to detect an impact of some minimal magnitude with a given level of confidence.

The power of the design is the probability that, for a given effect size and a given statistical significance level, we will be able to reject the hypothesis of zero effect. Consequently, sample sizes, as well as other design choices will affect the power of an experiment. Remember that in hypothesis testing we can make two kinds of mistakes: i) Type I error: when we reject the null when it is true; ii) Type-II error: when we fail to reject the null when it is false which equals 1-power.

We will construct an hypothesis test with the null hypothesis of no treatment effect: $H_0$: $\delta = 0$

We are considering the mean difference estimator for ATE, which under random assignment of treatment (in if all the units comply), is equal to ATOT.

$$\hat{\bar{\delta}} = [\hat{\bar{Y}}_{u1}|T_u = 1] - [\hat{\bar{Y}}_{u0}|T_u = 0] = \hat{\bar{Y}}_{u1} - \hat{\bar{Y}}_{u0}$$

And the standard error of the estimated causal effect is:

$$se(\hat{\bar{\delta}}) = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}} = \sqrt{\frac{\sigma_1^2 + \sigma_0^2}{N}} = \sqrt{\frac{2\sigma_0^2}{N}}$$

Note that we are assuming that half of the population is assigned to treatment and half to the control group, then $N=N_1=N_0$ represents the amount of units assigned to each condition (the sample size in this case is twice N). And, we can assume that under the null hypothesis of no treatment effect, the variance of the treatment of the outcome of interest for the treatment and control group is the same.

Let's define the minimum difference we want to be able to detect as:

$$\Omega = \mu^A - \mu_0$$

If we want to detect a minimum effect of Ω, with a power of (1-$\beta$)%, we have to choose the sample size so that we reject the null hypothesis of no effect with probability (1-$\beta$)%, when the true effect is Ω and the significance level is set in α%.

$$p(-Z_{\alpha/2} < \frac{\hat{\delta}}{se(\hat{\delta})} < Z_{1-\alpha/2}) = \beta = 1 - power$$

$$p(-Z_{\alpha/2} - \frac{\Omega}{se(\hat{\delta})} < \frac{\hat{\delta} - \Omega}{se(\hat{\delta})} < Z_{1-\alpha/2} - \frac{\Omega}{se(\hat{\delta})}) = \beta$$

$$p\left(z < Z_{\alpha/2} - \frac{\Omega}{se(\hat{\delta})}\right) - p\left(z < -Z_{\alpha/2} - \frac{\Omega}{se(\hat{\delta})}\right) = \beta$$

Z is the standard normal distribution. The second term in the last expression is approximately zero, so we are left with the first one:

$$p\left(z < Z_{\alpha/2} - \frac{\Omega}{se(\hat{\delta})}\right) = \beta$$

This analysis shows that when designing an experiment; significance level, power of the test, minimum detectable difference and sample size are all subject to the choice of the researcher except one. The researcher can choose freely all but one of these parameters and the other will be determined by the following equation:

$$Z_{\alpha/2} + Z_{1-\beta} = \frac{\Omega}{\sqrt{\frac{2\sigma_0^2}{N}}} \Rightarrow (Z_{\alpha/2} + Z_{1-\beta})^2 \ \sigma(\hat{\delta}) = \Omega^2$$

$$(Z_{\alpha/2} + Z_{1-\beta})^2 \ \sigma(\hat{\delta}) = (\mu_0 - \mu_A)^2$$

We can assume that we know the baseline variance of the outcome variable, and can fix the significance level, the power of the test and choose a minimum detectable difference in the outcome; then, the expression above will be an equation for the sample size. As regards the variance of the outcome, it is usually information that is not available before we have a baseline survey, but in those cases, as a first guess we can look for these parameters in previously collected data, perhaps from another location that could have that variable distributed in a similar way.

Then, we have to choose the significance level, or probability of type-I error, which in general is set at 5% or 10%. The next step is to specify the effect size that one wishes to be able to detect. Perhaps we have a hint of the potential effect of the intervention, and choose a conservative estimate of it in case it results lower. If we consider a very subtle effect, we will have to increase the sample size so as to detect that effect at a certain significance level and with a determined power. Finally, we face the trade-off between sample size and power of the test. It is usually considered that a power of 80% to 90% is acceptable. Then, the sample size will be determined so as to meet all these requirements.

In most of the cases the sample size has also a budget constraint. Therefore, the researcher must consider given the maximum sample size he can have, which power he can attain, or what would be the minimum detectable effect at a certain power. Then, if the minimum detectable difference is to low or the test is not well powered, it is suggested that the experiment should not be performed if the sample size cannot be increased.

What about power calculations in the case of GRT? Now, N that represented the amount of units assigned to each treatment arm, should be replaced by M.G (number of members times number of groups), so we have to determine both parameters. And the variance of the mean difference estimator is now:

$$\sigma^2_{\hat{\delta}} = \frac{2\sigma^2_y}{mg}\left(1-(m-1)ICC\right).$$

We see that for reducing the variance, increasing M is not the same as increasing G. In fact, an increase of G, given the Intra-Cluster Correlation (ICC), will help reduce the standard error of the estimator at a faster rate for a fixed sample size. An increase in *M* will cause, that the estimator will be less affected by the noise coming from the individual error term. Increasing G will have the same effect on both, the individual and the group error term. If the cost depended only on the total number of units surveyed: *n=2MG*, then it is clear that we would choose to sample one unit from each group (M=*1*) and sample as many groups as necessary to achieve the desired power. Note that the larger the ICC, the grater the variance inflator factor, and thus, the higher the sample size needed.

In most real-life situations there will be additional fixed costs associated to sampling more groups. For instance, groups might be cities and sampling an additional city might require paying fixed costs such as administrative expenses, getting a permit to treat individuals in that city, train field

workers in a new region, etc. In those cases, the study design has to weigh the power benefits from large number of groups against its costs and might find it optimal to sample a few groups and try to reduce the estimator variance increasing the number of members per group (if possible).

*Some further considerations on Randomization*

If random assignment of units to treatment and control groups is correctly implemented, we can affirm that randomization minimizes the threat to internal validity coming from selection bias. However, randomized experiments do not guarantee internal validity, though they make valid inference about causal relationships, because we cannot be sure that the other threats to internal validity are controlled. There are three situations in which the internal validity of the causal inference is threatened, even under randomization: i) lack of two stages –related to external and internal validity-; ii) non compliance and iii) attrition.

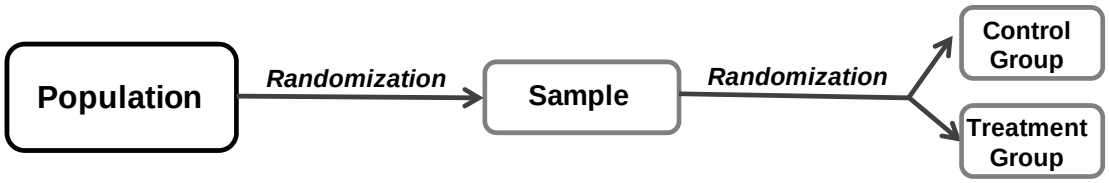
In addition to these, we must take into account that although randomization seems that does not require any assumptions, there are unstated assumptions about the problem of interest, the number of stages in a program, and the response of agents to randomization.

Randomization of units to conditions helps to solve the selection bias problem and disentangle the causal effect of a treatment on the sampled units. Ideally, we would like to generalize the effects that we found on the sampled units to the population where they belong. However, this generalization is not granted. For instance, we might have sampled just the units from a sub-population that were more responsive to the treatment, so that for the whole population the treatment effect has to be necessarily lower than our estimation. So the possibility or having sampled a particular group of units that are not representative of the whole population impedes us to generalize our conclusions.

This takes us back to our discussion about internal and external validity from the beginning of the Chapter. Although random assignment to treatment minimizes the threat to internal validity, it does not provide any information about the external. Therefore, we need a procedure to ensure the external validity of our results, which would allow us to generalize our results to the population where units belong.

Attaining both internal and external validity requires a two-stage randomization procedure, as shown in the next figure:

*Two-Stage Randomization*




-The purpose of the first-stage is to ensure that the results in the sample will represent the results in the population within a defined level of sampling error (external validity).

-The purpose of the second-stage is to ensure that the observed effect on the dependent variable is due to some aspect of the treatment rather than other confounding factors (internal validity).

Therefore, an ideal experiment would involve a two-stage randomization. In the first stage, the sample is chosen selecting randomly units from the population. This stage will allow the researcher to generalize the results of his study. In the second stage, sampled units are assigned to treatment and control. This stage will allow the researcher to infer causal relationship from the sample.

We have to introduce a subtle but not trivial distinction between *random assignment* and *random sampling*. We draw random samples of units from a population by chance when we get information from random samples of people or units. Random sampling ensures that answers from the sample approximate what we would have gotten had we asked everyone in the population. Random assignment, by contrast, facilitates causal inference by making samples randomly similar to each other, whereas random sampling makes a sample similar to a population. The two procedures share the idea of randomness, but the purposes of this randomness are quite different.

However, in many practical situations first stage randomization is not possible. For instance we might be allowed to randomize workers into conditions, but only some workers that are employed by a particular firm or group of firms, or that have an income under determined threshold. These workers might not be representative of the whole population of workers. In these cases, even if causal relationships can be correctly inferred from data, generalization to a wider population set is not granted.

Consequently, if we could get Two-Stage Randomized Trials, in large samples, we ensure that ATE is consistently estimated. However, if randomization takes place on a selected subpopulation the mean difference estimator only estimates ATOT consistently for that population.

If the first stage is missing, ATE will be biased due to two sources: i) *non-representative sample*: the distribution of participant characteristics in the sample does not reflect the populations'; ii) *non-constant treatment effect*: treatment conditions interact with participant characteristics. This is a main concern when due to the positive results of a randomized experiment, then the policy-makers consider the scaling up of the pilot.

There are also some issues that arise even in randomized experiments that can seriously threaten internal validity. Randomized social experiments solve the problem of selection bias by generating an experimental control group composed of person who could have participated but who were randomly denied access to the program. However, there are some problems that can arise because the units we are dealing with are human beings that cannot be forced to follow the experiment protocol.

For example, random assignment does not prevent participants from changing their behavior once they are assigned to treatment or control group. When participants are aware of the group they belong to, they can re-optimize and change their behaviour. For example, a worker not assigned to the job training program he volunteered to, may look for another program; or a worker assigned may have also found another alternative to improve his future earnings and may not participate.

These type of behaviours are typically referred to non-compliance. Another problem that can arise even after random assignment is attrition, which denotes the situation in which units belonging to the original sample in an experiment, leave it before the measurement of outcomes. Enrolling in a program, or being assigned a treatment, but then dropping-out, is another usual behavior among human beings. The consequences of imperfect compliance and attrition in randomized experiments, as well as the possible solution to these problems will be exposed in the next sections.

In general, Heckman and Smith (1995) state that there are two assumptions that must hold for outcomes of an experimental control group to correspond to the outcomes that participants would have experienced had they not participated in the program (counterfactual outcomes of participants):

i) Absence of Randomization Bias: randomization should not alter the process of selection into the program, so those who participate during an experiment do not differ from those who would have participated in the absence of an experiment. If we suppose that the effect of treatment would be the same for everyone, this assumption is unnecessary.

ii) Absence of Substitution Bias: participants assigned to control group cannot find close substitutes for the treatment elsewhere. It may happen that those note benefitted from the program will look for another similar program and thus drop out the study.

In the presence of these biases, the mean outcome of the experimental control group no longer represents the outcome that would have obtained the beneficiaries on average, had they not received treatment. When these sources of bias are relevant, the researcher must think of the potential drawbacks of performing social experiments and compare them with those of alternative non-experimental evaluations. In non-experimental techniques the counterfactual outcome is obtained econometrically using models that explain participation and outcomes; which naturally also require assumptions.

Furthermore, randomization itself could be a source of bias. Imagine that applicants to a training program know that they will be randomized in or out. It is probable that those risk-averse would seek other programs. In consequence, the only internal validity threat that randomization prevents from occurring is selection bias, which it rules out by definition.

Finally, random assignment is not the only way of eliminating selection bias. Uncorrelated errors can be created with other forms of controlled assignment to treatment status, as with non-experimental techniques such as regression discontinuity design that we will study further in this Chapter.

*Imperfect Compliance and Intention to Treat*

Consider that we can have an experimental design in which individuals who qualify for the job training program are randomly assigned to participate or not. Then we can compare outcomes of those participating and not to identify the causal effect. Note that if we considered randomly individuals from the population, and from this sample we randomly assign some to treatment and

some to control group, we would be identifying an average treatment effect. However, if the randomization is within a sub-group not randomly selected (volunteers, for example), we would be able to identify the average treatment effect on the treated.

In social sciences, where the experimental units are human beings, we can assign units to different treatment status but generally treatment is not perfectly enforceable. Thus, we must be careful, because for example in the job training program, assignment to the program does not necessarily mean participation. There are some participants which are assigned to the treatment group, but refuse to take it; and also it could happened that some controls are finally accepted in another program: this implies partial or imperfect compliance. Consequently, although assignment is random, participation may depend on certain factors not contemplated by the researcher.

In these cases, we can think that the assignment only affects the probability that the individual receives treatment but not actually rules out the treatment status. There are multiple causes of partial compliance; for example in many cases it is not possible to enforce compliance of control units when treatment is profiting and can be obtained from another source different from the program (a new drug that can be acquired, or even job training provided by another program).

In the former example of the training program, imagine we randomly assigned treatment to a randomly chosen sample. Now the variable indicating treatment assignment will be $Z_u$. If we allow for imperfect compliance, we must use another variable to refer to treatment actually taken, which will be $T_u$. But, how do we deal with the identification of the causal effect in the case of imperfect compliance?

So, if we calculated the difference between the outcomes of those assigned to treatment (Z=1) and those in control group (Z=0), would we be able to identify ATE? This is clearly not the effect of treatment (participation in training program) on outcome (income), because variable Z does not necessarily indicate participation. This difference will only identify ATE in the case of perfect compliance in which Z would coincide with T. Nevertheless, even in the presence of non-compliers, it is an interesting parameter. Which one? It represents the overall effect of the intervention to the entire sample, which is important for evaluating policies. We call it Intention-to-Treat (ITT), and can be thought of as the causal effect of treatment assigned but not necessarily the causal effect of treatment received.

$$ITT = \delta^{ITT} = E\left[Y_u \mid Z_u = 1\right] - E\left[Y_u \mid Z_u = 0\right]$$

Expanding each term by the law of iterated expectations, we can see that under perfect compliance, ITT is equal to ATE.

$$\begin{aligned}\delta^{ITT} &= E\left[Y_u \mid Z_u = 1\right] - E\left[Y_u \mid Z_u = 0\right] \\ &= E\left[Y_u \mid Z_u = 1, T_u = 1\right]\Pr(T_u = 1 \mid Z_u = 1) + E\left[Y_u \mid Z_u = 1, T_u = 0\right]\Pr(T_u = 0 \mid Z_u = 1) - \\ &\quad - E\left[Y_u \mid Z_u = 0, T_u = 1\right]\Pr(T_u = 1 \mid Z_u = 0) + E\left[Y_u \mid Z_u = 0, T_u = 0\right]\Pr(T_u = 0 \mid Z_u = 0)\end{aligned}$$

Because of perfect compliance, we can replace the probabilities by 0 if $T_u \neq Z_u$, and by 1 if they are equal, getting:

$$\delta^{ITT} = E[Y_u \mid Z_u = 1] - E[Y_u \mid Z_u = 0] = E[Y_u \mid T_u = 1] - E[Y_u \mid T_u = 0] = \delta^{ATE}$$

If treatment effect is homogenous among treated and untreated individuals, the fact that people from the treatment group do not take it (thus reducing the average outcome of those assigned), and people in the control group may take another treatment (thus increasing the average outcome of those not assigned to treatment), then ITT will be less than ATE. It is said to be a conservative estimate of ATE. However, when treatment is heterogeneous, ITT is still a biased estimate of ATE, but it could under or over-estimate it; we have no clue of the direction of the bias.

Remember that it was very important to assign treatment in a random way so as to balance the treatment and control groups in both observables and unobservables. But, when there is imperfect compliance, and we estimate the mean difference between those assigned to treatment and control groups, we can no longer ensure that the groups are balanced. This is because in general non compliance has diverse underlying reasons (for instance smaller expected gain from treatment or lesser motivation from participants), which unbalances groups and makes ITT a biased estimate of ATE.

Another estimator of the causal effect we could try, is the mean difference between those taking treatment and not. This does clearly not identify the causal effect of the training program (treatment), because even though assignment to the program was random, there was self-selection of participants. Thus making on average those actually treated different in some way from those not treated. Perhaps treated are more motivated or expect a greater gain than those that prefer not to take the program in spite of the fact of having been offered a place.

In this last case, we can claim that treatment is endogenous because the participation decision that makes the individual to take or not the treatment that is correlated to the potential outcomes of the individual. If selection is on unobservables, we can apply the standard instrumental variables framework, were we able to find a suitable instrument. Remember that the conditions that a valid instrument should satisfy are: i) the instrument must be correlated to the endogenous variable (T in this case); ii) the instrument must be uncorrelated to the error term, thus it must affect the outcome variable only through its effect on the endogenous variable T.

Notice that in the case we are analyzing of a randomized experiment with imperfect compliance, we have the variable denoting assignment Z, which is a perfect instrument for T because the treatment assignment is supposed to influence the participation into treatment at least to some extent; and its random nature makes Z uncorrelated to other variables affecting Y. So we can instrument T with Z and get the instrumental variable estimator. In the next line we will derive the estimator and then try to disentangle which assumptions we need to make it a meaningful causal parameter.

Recall the lineal regression model we studied earlier in this Chapter:

$$Y_u = \beta + \delta^{ATE} T_u + [\varepsilon_u + (\delta_u - \delta^{ATE}) T_u] = \beta + \delta^{ATE} T_u + \nu_u$$

We are looking for the causal effect of T (treatment; training program) on the response variable Y (earnings). Thus, we will have the following outcome regression;

$$\text{Outcome Regression: } Y_u = \beta_0 + \delta T_u + v_u$$

$T_u$ is not random, because in the end, there is no enforcement of treatment assignment, so the choice of participating or not in a certain program/treatment is done by each individual; there is self-selection into treatment. The regressor T is contemporaneously correlated with the error term; because there are unobservable factor influencing the decision to take the treatment (participate in job-training programs) that at the same time affect the outcome Y.

More formally, we can derivate the IV estimator applied to the case of an endogenous binary variable (treatment) as follows:

There are N units (u). Which can be assigned to treatment or not: $Z_u = 1$ or $Z_u = 0$.

As we allow for the possibility of imperfect compliance, each unit can actually receive treatment or not: $T_u = 1$ or $T_u = 0$.

Participation in treatment depends on the treatment assignment: $T_u = T_u(Z)$

The final outcome, $Y_u$, would depend on the assignment and treatment: $Y_u = Y_u(T,Z)$

We have two different regressions:

Outcome Regression: $Y_u = \beta_0 + \delta T_u + v_u$

Treatment Regression: $T_u = \lambda_0 + \gamma Z_u + w_u$

There are three parameters in which we may be interested in: i) The effect of Z on T; ii) The effect of Z on Y; and iii) The effect of T on Y

Under which assumptions could we identify those parameters? Under SUTVA (potential outcomes and treatments of each unit *u* are independent of the potential assignments, treatments and outcomes of the other units) and Ignorability (applied to assignment, meaning that each individual has the same probability of being assigned to treatment. $\Pr(Z_i=1) = \Pr(Z_j=1)$), we could dentify:

A. The causal effect of Z on D:
   $E[D_u|Z_u=1] - E[D_u|Z_u=0]$

B. The causal effect of Y on D:
   $E[Y_u|Z_u=1] - E[Y_u|Z_u=0]$

The instrumental variables (IV) estimator is defined as:

$$\delta^{IV} = \frac{\text{cov}(Y_u, Z_u)}{\text{cov}(T_u, Z_u)}$$

In case of a binary instrument, as in this case, this is equal to the Wald estimator:

$$\delta^{IV} = \delta_{Wald} = \frac{E\left[Y_u \mid Z=1\right] - E\left[Y_u \mid Z=0\right]}{E\left[T_u \mid Z=1\right] - E\left[T_u \mid Z=0\right]}$$

The IV estimator can be estimated in two steps: in the first one, the treatment equation is estimated by OLS and calculated the predicted $\hat{T}$. Then in the outcome regression, T could be replaced by $\hat{T}$ estimating the parameters again by OLS. This procedure is called Two Steps Least Squares (2SLS).

The IV estimator does not recover ATE, because note that we can re-write the IV estimator as:

$$\delta^{IV} = \delta + \frac{E[Z_u v_u \mid T_u = 1]\Pr(T_u = 1 \mid Z_u)}{\operatorname{cov}(T_u, Z_u)}$$

So the IV estimator only identifies ATE under the unlikely assumption that covariance between the return to treatment and the instrument is zero among the treated units.

Therefore, without further assumptions and general imperfect compliance $\delta^{IV}$ does not have a causal interpretation. There is still a special case when although there is imperfect compliance among the ones assigned to treatment, we can rule out the possibility that those randomized out could get the treatment. For instance, if we take a small village and randomize workers in and out a job training program, there is no such a program available in the village apart from the one we provide. $P(T_u=1|Z_u=0) = 0$

Expanding the conditional expectations from the expression for $\delta^{IV}$ and rearranging terms, we can get that under this special patter of imperfect compliance we recover the average treatment effect for the treated (ATOT):

$$\delta^{IV} = \frac{E\left[Y_u \mid Z=1\right] - E\left[Y_u \mid Z=0\right]}{E\left[T_u \mid Z=1\right] - E\left[T_u \mid Z=0\right]} = \frac{E\left[Y_u \mid Z=1\right] - E\left[Y_u \mid Z=0\right]}{P\left[T_u = 1 \mid Z=1\right]} = E\left[Y_{u1} - Y_{u0} \mid T=1\right] = \delta^{ATOT}$$

The IV estimator can also be related to the ITT parameter we defined before, because the numerator of the IV estimator is exactly the ITT.

$$\delta^{IV} = \frac{E\left[Y_u \mid Z=1\right] - E\left[Y_u \mid Z=0\right]}{E\left[T_u \mid Z=1\right] - E\left[T_u \mid Z=0\right]} = \frac{ITT}{E\left[T_u \mid Z=1\right] - E\left[T_u \mid Z=0\right]}$$

However, $\delta^{IV}$ in the general case of imperfect compliance does not have a meaningful causal interpretation. Angrist, Imbens (1994), state some assumptions to provide such intepretation

For the causal effect of T on Y, we need some further assumptions:

1. Non-zero average causal effect of Z on T. This means that treatment assignment influences the individuals' choice of participating in the program/ taking treatment.

This means that the probability of treatment must be different in the two assignment groups:

$Pr(T_u(1) = 1) \neq Pr(T_u(0) = 1)$

This is equivalent to stating that in the treatment regression, γ ιs significativly different from zero.

2. Exclusion Restriction: assignment only affects the final outcome through treatment. This cannot be tested.

3. Monotonicity: No one does the opposite of his/her assignment, no matter what the assignment is. In the next table we show all the alternatives that can happen combining Z and T.

*Classifications of Units according to Assignment and Treatment Status*

| | | $Z_u = 0$ | |
|---|---|---|---|
| | | $T_u(0) = 0$ | $T_u(0) = 1$ |
| $Z_u = 1$ | $T_u(1) = 0$ | *Never-Taker* | *Defier* |
| | $T_u(1) = 1$ | *Complier* | *Always-Taker* |

Monotonicity means that there are no defiers. They are represented as those for whom:

$Y_u(1,0) - Y_u(0,1) = -(Y_u(1) - Y_u(0))$

This assumption also implies that $T_u(1) \geq T_u(0)$.

Combining 1 and 3, we can claim that there are no defiers, and there is at least a complier. If from 1 to 5 hold, the following expression holds:

$E[Y_u(1,T_u(1)) - Y_u(0,T_u(0)] = E[Y_u(1) - Y_u(0) \mid T_u(1) - T_u(0)] \cdot Pr\{(T_u(1) - T_u(0)\}$

And the causal relation of T on Y would be:

$$\delta^{LATE} = E[Y_u(1) - Y_u(0) \mid T_u(1) - T_u(0)] = \frac{E[Yu(1,Tu(1)) - Yu(0,Tu(0)]}{Pr\{(Tu(1) - Tu(0)\}} = \frac{E[Y_u \mid Z = 1] - E[Y_u \mid Z = 0]}{E[T_u \mid Z = 1] - E[T_u \mid Z = 0]}$$

This is what they call the Local Average Treatment Effect (LATE), representing the average effect of treatment for those who change treatment status because of a change of the instrument; i.e. the average effect of treatment for compliers.

Notice that then, if we assume that the causal effect of treatment is the same for all the treated individuals independently of assignment (which would imply a strong not-verifiable assumption), then the instrumental variables estimator would estimate the causal effect of the treatment. However, if treatment is heterogeneous, we would only identify (under some further assumptions) only a local average treatment effect (LATE). This causal effect represents the effect of treatments only for those who are compliers; i.e., they participate or not in the program, according to what they have been assigned.

To illustrate the use of ITT and IV in the literature, let's consider the paper by Kling et al about the effects of neighborhood on crime. A majority of theoretical models suggest that returns to engaging in criminal activity are larger in communities with high levels of crime, and low economic opportunity. The Moving to Opportunity (MTO) experiment has been in operation in Baltimore, Boston, Chicago, Los Angeles, and New York since 1994. Participant families had to be living within public or section 8 project-based housing in certain high poverty census tracts. The families who signed up for the program were randomly divided into three groups: the *experimental* group, which received housing vouchers which could only be used in census tracts with a 1990 poverty rate below 10%; the *section 8* group, which received vouchers with no constraint on the area where they could be used; and the *control* group, which kept its pre-existing social services, but received no further assistance under MTO. So, the design consists of two alternative treatments and a control group.

Although voucher *availability* was random for each family, not every household actually used their subsidy. Thus, *take-up* was not random, which means that comparing outcomes across groups will not give an unbiased estimate of the effects of moving to a better neighborhood. However, initial assignment can be used as an instrument for actual treatment in the estimation of the effects of MTO on those who actually use it. This paper uses initial assignment directly to estimate the *Intention to Treat* effect, and as an instrument to obtain the *Treatment on the Treated* effect of the MTO program. The authors want to study the effect of program participation on youth delinquency and criminal behavior. Initial characteristics are similar for all treatment groups, in both samples, indicating a successful randomization. A year after randomization, however, those who moved lived in strikingly different areas from those who didn't, and the difference persisted even four years after that. This effect persisted three years later – with experimental and section 8 households living in richer neighborhoods with lower violent and property crime. Moreover, in 2002 (about eight years after randomization), these households lived in neighborhoods with higher adult involvement and police activity. The paper attempts to estimate what part of these differences should be attributed to the MTO program, given that actually using the vouchers was not random. The authors conclude that overall, violent crime went down – the IV estimate for the experimental group indicates that on average, youths whose households decided to move as a consequence of voucher availability had 0.147 less lifetime arrests for violent crime. Given that the sample mean is 0.388 lifetime arrests for this crime

category, this is a substantial effect size. Though not significant, there is an increase in property crime for this group.

### *Attrition*

Attrition is the last problem that remains even after random assignment that we will deal with in this chapter. As stated before, attrition is the name for the situation in which units belonging to the original sample in an experiment, leave it before the measurement of outcomes. In which situations can attrition appear? There is a wide range of ways in which attrition can evidence: simply failing to ask a single question to a participant, dropping people from the originally assigned sample because they do not meet the necessary criteria (no longer have the illness required for the clinical trial); or because the participant moved. It is very important to determine whether attrition was influenced by treatment or not. For example, in the first cases attrition may be caused by a random failure of the researcher, but in the last case (participants who move and thus drop out of the experiment) it is complicated to identify if the cause of its move was treatment assignment or if that person would have moved anyway.

Even though any kind of attrition reduces the statistical power, if attrition is not correlated with the conditions of the unit in the experiment or the future outcomes, it would not be a threat to internal validity because it will not bias the estimate of the causal parameter. However, if attrition is correlated with the assignment or with potential outcomes it introduces biases in the estimates that are hard to deal with.

Remember that random assignment had the advantage of equating groups on expectations in the pre-test, and assuming that that the equivalence would hold in the post-test. However, when there is non-random attrition, the correlation between attrition and outcomes act as a confounder to detecting the causal effect of treatment. In those cases changes in outcomes could be attributed both to treatment and the unknown correlation between attrition and treatment. Randomization deals with selection bias, but cannot prevent attrition bias to occur.

Usually, a first step to deal with attrition, is planning in advance a larger sample. This is effective in the case where attrition is balanced in the two groups, not only in the rate but also in the observable and unobservable characteristics of those quitting (the simplest case when we say that attrition is random). Another important issue when designing the set-up of an experiment is the contemplation of retention and tracking strategies so as to prevent attrition. For example, there is a list of procedures that can be set so that the researcher can find the individual after treatment and not lose its information (collecting data about its location and its relatives, offer economic incentives to keep people in the sample, provide adequate contention during interviews to prevent answers' refusal, minimizing time and obstacles between randomization and treatment, etc.).

However, if attrition is differential, there will be a complicated problem. For instance, assume that individuals who get the lower benefit from a program tend to quit. Then, if we don't take this into account, we will overestimate the effect of the program. Consider the example of students taking

extra classes, if the worst students leave the treatment group before the test, the causal estimate will be biased upwards. So, what does constitute a threat to internal validity is differential attrition.

Besides, differential attrition may be present even when the rates of attrition are the same in the control and treatment group, because they may drop out due to different reasons. Therefore, our estimates might be biased despite both groups exhibit the same percentage of drop-outs. For instance, in a clinical trial evaluating some procedure or medication, attrition might be due to health status being extremely bad, death; or extremely good, making people to lose interest in the program. If the treatment is beneficial, attrition in the treatment group will be among those with the best potential outcome, while in the comparison group will be among those with the lowest. Even if attrition rates are similar, our estimates of the treatment effect will be biased downward.

Statistical procedures to deal with attrition ultimately depend on assumptions made on how attrition occurs. Therefore, planning the data collection process in order to minimize its occurrence is crucial. When attrition is unavoidable and following up attritors is too expensive, a random sample of them can be selected for intensive follow-up. The problem of attrition has received diverse approaches and treatments, from assuming that nonresponse is ignorable, to modelling the process determining which data are missing (Heckman 1979), or bounding the parameter of interest (Lee 2005, Horowitz and Manski 1998).

*Descriptive Analysis and Identification of Attrition Patterns*

If attrition occurs, the researcher must report attrition rates in both condition groups. If baseline data is available he should compare attritors with non-attritors to detect systematic patterns along any observable dimensions. The existence of a pre-test contributes very much when there is differential attrition, as we mentioned in the designs section.

As part of the descriptive analysis, the overall attrition rate and the differential ones for the treatment and control group should be reported. Also, there should be analyzed whether those who completed the study differed from those who did not on important observables; and the same study should be performed for those who completed the treatment and who did not; and for those who remained in the control and treatment group. Furthermore, it should be assessed not only whether different groups have different patterns of attrition, but also if different measures show different patterns. Also, it should be determined whether certain subsets of respondents or sites have complete data that could be useful to identify the causal effect.

The appropriate statistic procedure to deal with the problem depends on the diagnosis about the attrition pattern.

*Replacing Drop-outs*

One way to deal with attrition would be to set the number of units which quit the experiment before measurement in each group, and in order to regain the original sample size, randomly select new participants (and randomly assign them treatment) from the same pool of applicants where the original group was selected.

The advantage of this approach is that it keeps the original sample size, which is important for statistical power sake. However, it does not necessarily solve the bias that attrition may generate. There are some assumptions under which this procedure would solve attrition. Following Shadish et al (2002), these assumptions are: i) both attrition and replacement are at random (which is unlikely for the case of attrition and very difficult to verify); or ii) former and new participants have the same latent characteristics (again unlikely and difficult to verify).

*Sample Selection Modelling*

This approach starts from the fact that the outcome variable is not observable for every unit of the sample. The reason may be diverse, for example, we may be interested in the effect of a training program on earnings, and we will only have the post-treatment income of those who work; or merely because of drop-out units.

The key issue of this type of analysis is that although we cannot see all the observations for the response variables, we could model the sample selection model. For that, we should find some instrument that could affect the probability of observing certain units' responses in the sample, but at the same time must not affect that response or outcome.

If we could find suitable variables to model the selection process, then, following Heckman (1979), we should include another term (corresponding to the inverse of the Mill ratio) as a regressor in the outcome equation, to correct the estimation for sample selection.

In more detail, the Heckman Sample Selection Model, starts considering the response equation that has a response variable *y*, observed only when *d* is positive. Otherwise, the response variable is not observed, but the variables involved in the probability of observing that variables are observed.

$$y_i^* = \mathbf{x_i'\beta} + \varepsilon_i$$
$$d_i^* = \mathbf{z_i'\gamma} + u_i;\ d_i = 1[d_i^* > 0] \text{ (probit)}$$
$$y_i = y_i^* \text{ if } d_i = 1; \text{ not observed otherwise}$$

The main assumptions about these variables are:

$$[\varepsilon_i, u_i] \sim \text{Bivariate Normal}[0,0,\sigma^2,\rho,1]$$

Then Heckman proposes to correct the estimation for sample selection as follows:

$$\begin{aligned} E[y_i^* | x_i, d_i = 1] &= \mathbf{x_i'\beta} + E[\varepsilon_i \mid x_i, d_i = 1] \\ &= \mathbf{x_i'\beta} + E[\varepsilon_i \mid x_i, u_i > -\mathbf{z_i'\gamma}] \\ &= \mathbf{x_i'\beta} + \rho\sigma\left(\frac{\phi(\mathbf{z_i'\gamma})}{\phi(\mathbf{z_i'\gamma})}\right) \\ &= \mathbf{x_i'\beta} + \rho\sigma\lambda_i \end{aligned}$$

From the last expression we can see that performing OLS provides us with inconsistent estimates, because it does not include the last term.  is called the inverse of the Mills, and represents a monotone decreasing function of the probability that an observation is selected into the sample.

He have already mentioned the difficulty of finding credible instruments for endogenous regressors, and evaluate programs where intention-to-treat is randomized. However, devising instruments for the probability of observation is a more challenging problem. One way to deal with this problem is partial randomization of the probability of observation. An alternative to perform it, is to alter data collection procedures, such as by randomizing half of sampled units to more intensive follow–up.

A common critique to this parametric approach to correcting sample selection is that its validity relies on distributional and functional form assumptions, as well as exclusion restrictions. A number of less restrictive semi-parametric approaches have been recently introduced in the literature. A common thread in all these formulations, however, is that they need to characterize the mean of the regression error term conditional on the regression covariates and the sample selection rule. An alternate approach is discussed in Ahn and Powell (1993). They propose to eliminate the selection bias by differencing observations with similar probabilities of selection sidestepping the problem of estimating the unknown conditional mean function.  Angrist (1995) propose using propensity scores to condition on the probability of selection given the instrument chosen, as long as the relation between the instrument and selection status, satisfies monotonicity.

*Setting Bounds to the Estimation*

When using instruments to estimate the probability of retaining the observation in the sample, we must trust on exclusion assumption which may not be credible because they are very difficult to satisfy. Remember that instruments must not appear in the outcome regression but must influence de probability of selecting that observation into the sample.

Lee (2005) set bounds that allow for treatment effect heterogeneity and require no valid exclusion restriction. However, they do involve additional assumptions. Assume that we know or assume that those assigned to treatment are at least as likely to remain in the sample as those who are not, and that treatment is never harmful. Then, the mean difference estimator will provide a lower bound on the average treatment effect on the non-atrittors. To obtain an upper bound one can choose a quantile, q, such that q not higher than the attrition rate in the control group, and then drop the lower $q$ percent of the treatment distribution to calculate the mean difference. There is one additional assumption behind this procedure, that the counterfactual observation of those treated units that are not dropped (i.e. they are in a quantile higher than $q$ in the distribution of outcomes under treatment) belong to quantile higher than $q$ in the distribution of outcomes under absence of treatment.

Horowitz and Manski (2000) provide another approach using bounds; but their bounds are wider, and are constructed considering the worst case assumption about missing data, and the higher assumptions for the upper bound.

***Quasi-Experimental Designs***

We have been discussing throughout the previous pages that to estimate causal effects we would ideally like to know the counterfactual outcome, which is impossible. We introduced the mean difference estimator, that consisted on comparing the outcome of the treated units to that of the control group. If the intervention is randomly assigned, the treatment and control group are probabilistically similar in every observable and unobservable characteristic, except for treatment status. When randomization is not possible, but we can claim that selection occurs on observable variables, we could try to find an appropriate control group through matching for example. When selection into treatment is due to differences in unobservables (like the case of imperfect compliance), we could look for a suitable instrument and use the IV estimator that under the monotonicity assumption can provide the LATE (effect of the treatment on compliers). However, it is really complicated, and sometimes not even possible to find a valid instrument. When we have random assignment, we can have the perfect instrument, but when it is not the case, it is more complicated.

In the next part of the Chapter, we will introduce some designs that under certain assumptions allow the researcher to identify causal effects, but they do not imply random assignment of units to treatment status. That is why these designs are often called Quasi-Experimental Designs; they look like experiments, but lack the main characteristic: randomization.

The first type of quasi-experiments that we will study are the *natural experiments*. Sometimes, we can find that natural events, (such as floods, earthquakes) or administrative regulations (changes in laws that does not have to do with the outcomes we study), can be a source of exogenous variation that can allow us to identify causal effects. A natural experiment can be though of as an instrument in some sense, but broadly speaking is another evaluation method very popular in practice, though, finding natural experiments is as difficult as finding valid instruments.

To illustrate how natural experiments can be used to identify causal relationships, consider the difficult task of evaluating the causal effects of property rights. In most of the cases their allocation is endogenous (based on wealth, family characteristics, individual effort, previous investment levels, or other selective mechanisms), thus impeding the identification of the causal parameters. Galiani and Schargrodsky (2008) exploit a natural experiment in the allocation of land titles to overcome this identification problem.

In 1981, about 1,800 families occupied a piece of wasteland in the suburbs of the Province of Buenos Aires, Argentina. The occupants were groups of landless citizens organized through a Catholic chapel. As they wanted to avoid creating a shantytown, they partitioned the occupied land into small urban-shaped parcels. In 1984, the Congress passed a law expropriating the land from the former owners with the purpose of entitling it to the occupants, some of the original owners

accepted the government compensation, while others are still disputing the compensation payment. These different decisions by the former owners generated an allocation of property rights that is exogenous in equations describing the behavior of the squatters. They found that entitled families increased housing investment, reduced household size, and improved the education of their children relative to the control group. The authors could benefit from the change in a property rule and different responses to assess the effects of property rights in outcomes of interest.

In most of the quasi-experimental designs, we will need to have a pre and a post test of each unit. Generally speaking, when a cross-section or group of units (people, students, schools, firms, etc.) are surveyed periodically over a given time span, we refer to the pool of data as a Panel. Panel analysis can provide a powerful study of a set of units, if one is willing to consider both the space and time dimension of the data. We will exploit the special and temporal dimensions to estimate causal effects, in situations in which we can assume that the counfounders are unobserved but fixed omitted variables.

The pre-tests consist of gathering information about the units belonging to the control and treatment group before the treatment is implemented. Sometimes obtaining a pre-test is not feasible because it is either not possible in the context of the treatment studied, or it is too expensive or time demanding. For example, when we assign workers to a training program, the researcher might add multiple pretests to reveal if motivation or cognitive development trend before the treatment is analogue between the groups, and then compare it with pos-tests measurements.

In pre-test post-test we added an observation before the treatment. In this way, longitudinal design adds multiple observations before, during or after the treatment, depending on the effect we are trying to identify. Although such a design would be useful because it would provide very valuable information, in practice it is very rare to have various pre-test and also post-test. It depends on the type of treatment being studied (sometimes it has no sense to follow up units too long). And it is not totally useful to have a lot of post-tests, because the threat of attrition increases with the length of the period considered after treatment. Furthermore, it is probable that the control group will eventually be assigned to treatment (this or any other). In some cases it will not even be ethical to prevent those units from receiving a benefiting intervention.

Having a pre-test and a post-test leads us to think again on the fundamental problem of causal inference. If we could observe each treated unit twice, before and after treatment, could we use the first observation as a valid counterfactual for the second? The answer is yes, but only if the baseline outcome were time invariant. Otherwise, we would be confusing a trend effect with the true treatment effect. Notice that we perfectly controlled for any time invariant factor, but time varying factors will be confounders (time trend in the formula above). Any secular trend increasing the outcome will inflate our estimation of the treatment effect. So, we need a counterfactual to see what would have been the true trend if units had not been treated. Now the problem of causal inference reappears with all its strength! However, the problem is different; we are not concerned about counterfactual levels but about counterfactual trends.

For example, imagine we return to our example of the job training program. From a population of unqualified workers, a group is assigned to a training program and another to a control group. We

observe their earnings before the program and after it. If we compare the earnings of the participants before and after the program, we may find a positive difference that could be attributed either to the intervention or to economic growth in that place at that time which spilled over all the sectors of the economy, in particular to that where the workers of our experiment belong.

Let's look at the problem employing some simple expressions. In a pre-test post-test design we observe for each individual two observations, both defined by:

$$Y_{it} = \alpha + \chi t + \delta T_{it} + u_i + \varepsilon_{it} .$$

Where $\alpha$ is the grand mean in period 0, $\chi$ is a trending term, $\beta$ is the treatment effect, $D$ is a dummy that is 1 when unit $i$ is treated in period $t$ and 0 if not. Besides $u$ and $\varepsilon$ are error terms, $u$ is an error term that affects the level of each unit and $\varepsilon$ is a trending error term. Consider that treatment takes place in period t=1 for those treated, and never for those units assigned to the control group

How does this general individual outcome expression look like for every period for a treated individual?

For t=1: $Y_{i1} = \alpha + \chi + \beta + u_1 + \varepsilon_{i1}$

For t=0: $Y_{i0} = \alpha + u_1 + \varepsilon_{i0}$

Subtracting observation in period 1 minus that of period 0 for those treated, we obtain:

$$\Delta Y_i = \chi + \beta + \Delta\varepsilon_i$$

This average before and after comparison is not an unbiased estimate of the treatment on the treated, because even though we could set $E(\Delta\varepsilon_i \mid \Delta D = 1) = 0$, we have a trending term $\chi$, which need not be zero.

Let's do the same exercise for those in the control group:

For t=1: $Y_{i1} = \alpha + \chi + u_1 + \varepsilon_{i1}$

For t=0: $Y_{i0} = \alpha + u_1 + \varepsilon_{i0}$

Subtracting observation in period 1 minus that of period 0 for those in the control group, we obtain:

$$\Delta Y_i = \chi + \Delta\varepsilon_i$$

Finally, if we substract the difference in outcome from those in the control group from the difference in outcomes of those in the treatment group, we would be able to identify the causal effect:

$$E[\Delta Y_{it} \mid \Delta D_{it} = 1] - E[\Delta Y_{it} \mid \Delta D_{it} = 0] = \beta$$

So, if we compare the increase in treated units with the increase in control units we will obtain an estimator of the true treatment effect: the Difference-in-Difference (DiD) estimator, most popular as Dif-in-Dif.

We can also express it as:

$$DiD = \left[\bar{Y}_1 \mid D = 1 - \bar{Y}_0 \mid D = 1\right] - \left[\bar{Y}_1 \mid D = 0 - \bar{Y}_0 \mid D = 0\right]$$

What is really important to note is that the DiD transforms our original problem into a new one. In our original problem we were concerned about making a counterfactual such that the level of baseline outcome of units assigned to treatment is comparable to that of units assigned to control. In our new problem we focus on making a counterfactual such that the trend of baseline outcome is comparable across conditions.

If we want to explicitly compare the panel data estimation with the cross section one, assume again that there are two time periods: *t=0* and *t=1*; and that the outcome is determined by the following expression:

$$Y_{it} = \alpha + \chi t + \beta D_{it} + u_i + \varepsilon_i t$$

In the cross section analysis, we only observe one time period, say *t=0*, and we compare the outcomes for treated and untreated units. The outcomes are determined by:

$$Y_{i0} = \alpha + \beta D_i + u_i$$

The variance of the mean difference estimator will depend on the variance of the error term *u*. In particular, if we assigned $n_0$ units to each condition this variance will be:

$$V\left(\hat{\bar{\delta}}\right) = 2\frac{\hat{V}(u)}{n_0}$$

And the mean difference estimator will be consistent if the error term *u* is uncorrelated with the treatment assignment.

In a pre-test post-test design we observe for each individual two observations and subtracting observation in period 1 minus that of period 0 we obtain:

$$\Delta Y_i = \chi + \beta D_i + \varepsilon_i$$

Now we are facing a similar problem as before. Notice that this last expression is a re-labeling of equation $Y_{i0} = \alpha + \beta D_i + u_i$.

| Cross Section Design (Basic) | Pre-Test Post-Test Design |
|---|---|
| $Y_{i0} = \alpha + \beta D_i + u_i$ | $\Delta Y_i = \chi + \beta D_i + \varepsilon_i$ |

In a pre-test post test design, the DiD estimator is the analogue to the mean difference estimator in a cross section; the population mean trend is the analogue to the grand mean in levels and the error term is now an idiosyncratic variation across time.

The DiD estimator compares the mean increase in treated units against that of control units. This estimator will be consistent if $\varepsilon$ is uncorrelated with the treatment assignment and its variance will be:

$$V\left(\hat{\bar{\delta}}\right) = 2\frac{\hat{V}(\varepsilon)}{n_0}$$

From this expression it is clear that a pre-test post-test design can improve the power of the analysis if the variance of the trending error term $\varepsilon$ is lower than that of the error term $u$.

If we had to choose between a pre-test-post-test and a cross section design taking into consideration this concern, we should take into account which heterogeneity do we prefer to face. If we think that baseline heterogeneity will be lower, we might prefer a cross section design, since the estimator variance will be lower in this design.

The *post-test-pre-test* design is also possible when we assign entire groups to the treatments (such as in the Group Randomized Trial). Each group will be surveyed twice, once before and once after treatment is administered. This design will perfectly control for every time-invariant group effect, but now trending group effects will confound with the treatment effect. Now, the DiD estimator will compare group mean trend between treated and untreated units.

Relative to the *post-test-only* design using a pre-test post-test design can greatly increase power if the variance of the trending group error is smaller than the variance of the time-unvarying group error.

When we assign groups to conditions, there are two different flavors of this design: the cohort design and cross-section design. In the cohort design we obtain a pre-test and a post-

test observation for each member in each group. In the cross-section design, we obtain a pre-test and a post-test observation for each group but within each group from different members. In both cases the DiD estimator is valid, the only difference is that the estimator variance is lower in the cohort design since we can control for individual time-unvarying effects.

In both, cross-section and cohort design, the model can be written as:

$$Y_{i:kt:l} = \mu + \chi t + \delta D_k + G_{k:l} + tQ_{k:l} + e_{i:kt:l}$$

Where $\mu$ is the population outcome mean, $\chi$ is a trending term, $\delta$ is the treatment effect, and $D_k$ is 1 when group $k$ is assigned to treatment and 0 if it is assigned to control. $G_{k:l}$ stands for the time invariant group effect, $Q_{k:l}$ is a trending error term that is shared by individuals belonging to the same group, both errors are assumed to have zero mean. Finally, $\varepsilon_{i:kt:l}$ is the residual term which is also assumed to have zero mean.

Taking means across groups will lead to:

$$\overline{Y}_{kt:l} = \mu + \chi t + \delta D_k + G_{k:l} + tQ_{k:l} + m^{-1}\sum_i e_{i:kt:l}$$

Subtracting group means after and before treatment lead to the following group mean variation

$$\Delta\overline{Y}_{kt:l} = \chi + \delta D_k + Q_{k:l} + m^{-1}\left(\sum_i e_{i:k1:l} - \sum_i e_{i:k0:l}\right)$$

Taking means across conditions:

$$\overline{\Delta\overline{Y}}_{l} = \chi + \delta D_k + g^{-1}\sum_k Q_{k:l} + (gm)^{-1}\left(\sum_i e_{i:k1:l} - \sum_i e_{i:k0:l}\right)$$

Now the DiD estimator is:

$$\hat{\delta} = \delta + g^{-1}\left(\sum_k Q_{k:T} - \sum_k Q_{k:C}\right) + (gm)^{-1}\left(\sum_i e_{i:k1:T} - \sum_i e_{i:k0:T} - \sum_i e_{i:k:C} + \sum_i e_{i:k0:C}\right)$$

Now the relevant group effect is the time varying $Q_{k:l}$ and its variance will appear on the estimator variance formula.

In the cross-section design we can assume that individual errors terms are uncorrelated and the DiD variance will be:

$$V\left(\hat{\delta}\right) = \sigma^2_{\hat{\delta}} = 2\frac{m\sigma^2_{Qg:c} + 2\sigma^2_e}{mg}$$

Where $\sigma^2_{Qg:c}$ and $\sigma^2_e$ are the variances of the time varying error term $Q_{k:l}$ and the individual error term $e_{i:kt:l}$, respectively.

On the other hand, in a cohort design the individual error term can be decomposed in an individual error term plus an individual trending error term: $e_{i:kt:l} = u_{i:k:l} + t\varepsilon_{i:k:l}$. In this case the individual error term $u$ will cancel out and the variance of the DiD estimator becomes:

$$V(\hat{\delta}) = \sigma^2_{\hat{\delta}} = 2\frac{m\sigma^2_{Qg:c} + \sigma^2_{\varepsilon}}{mg},$$

where $V(\varepsilon_{i:k:l}) = \sigma^2_{\varepsilon}$.

Note that as the expression for the variance of the estimator is different in panel data than in the cross-section setting, when we estimated the mean difference estimator, the formula for power calculations will also differ. The main difference is that in the case of the DiD estimator, the time invariant group effect will cancel out, whereas the effect that is important now is the group-time varying effect. So we will redefine the ICC as the share of the variance that corresponds to the group-time effect.

Having a pre-test might be also relevant even when we could assign units randomly to the control and treatment groups. This is because, although under random assignment the treatment and control group should be probabilistically similar in all observable and unobservable characteristics, it may be very useful to check that the sample we work with is balanced. For that purpose, one can compare a set of pre-test variables that on average should be similar for the two groups. Ideally the pre-test variables should be also of the post-test.

On the other hand, having a pre-test is also very useful in case we have attrition, because it would reveal if there is differential attrition (if those who dropped are different from those who stayed). So having a pre-test helps us find a threat to internal validity in this cases.

What problems can we find in applying the DiD estimator? There are some underlying assumptions that must be met in order to claim that the DiD estimator really captures the causal effect of an intervention. One main condition is that the error term is not correlated with treatment assignment; another one is that both the treatment and control group share the same trend. Finally, treatment assignment must not depend on past or future outcomes (it must be exogenous).

For example, is we implement the training program to workers that have in the last year an income lower than a certain threshold, we will have participating in the program workers

that have structural characteristics that make their income low and also, workers that last year were affected by a negative shock. If the shocks are independent, then this year those workers may probably not be affected by such a negative shock and thus, the estimated causal effect will be overestimated. This is called the "Ashenfelter's dip".

One common problem when trying to identify causal effects is simultaneity. For example, if we look for the effect of police on crime, we could regress crime on police and may find a positive correlation. This is because although more policemen in the streets tend to reduce crime, but the fact that cities with higher crime's rates have more policemen can make the relation have the opposite sign one expects. In the presence of simultaneity, we cannot identify the DiD estimator. Nevertheless, in the context of program evaluation, simultaneity is seldom a concern because interventions are often exogenous. If they are not, as we will see later, we will face problems identifying the causal effect.

The assumption that is a real concern to obtain a consistent DiD estimator in program evaluation, refers to trends. When we use a pre and post-test for control and treatment groups, we are assuming that the control group would have the same trend that the treatment group would have had in absence of the intervention. This assumption allowed us to drop out the trend term in the expressions previously studied.

But assuming that both groups have the same trend is not trivial at all. In fact, the trend heterogeneity plays exactly the same role that baseline heterogeneity played in a cross section design. If by sheer luck we assign units with higher trends to treatment we will be overestimating the true ATE; and if we will underestimate ATE if we assigned units with lower trends. Therefore, the variance of our DiD estimate will increase with the trend heterogeneity. Consequently, the selection bias must be additive and time-invariant, so that the trend is the same in both groups. In practice, matching methods can be useful to control for time varying selection bias. The researcher must have a measure of the variable of interest of participants and non-participants based on observed characteristics in baseline conditions, and prior trends.

To illustrate the use of panel data to estimate the causal effect, consider the paper by Berlinski et al (2009) in which the authors examine the returns to pre-primary education by taking advantage of a large infrastructure program aimed at increasing school attendance for children between the ages of 3 to 5. Between 1993 and 1999, Argentina constructed enough classrooms for approximately 175,000 additional children to attend preschool. The government used a non-linear allocation rule based on an index of unsatisfied basic needs constructed with data from the 1991 Census in order to target the construction to poor areas with low pre-primary enrolment rates.

The growth in enrolment between 1991 and 2001 is noticeable, as the average enrolment rate increased to 64 percent and the number of children attending pre-primary school

climbed by 330,845. Comparing 1991 to 2001, all provinces increased enrolment in pre-primary education by at least 10 percentage points. In contrast, primary school enrolment increased negligibly from 97 percent in 1991 to 98 percent in 2001. How much of the increase in enrolment was caused by the construction program? To answer this question, they exploit data on preschool enrolment of children aged 3 to 5 from the Argentine household survey, estimating the pre-primary school enrolment as a function of the new places created by the rule, and time and province fixed effects. The authors find that one place constructed per child in preschool age increases the likelihood of preprimary school attendance by 0.813. Moreover, they cannot reject the null hypothesis that the coefficient is one and therefore that there was full take-up of newly constructed places (this is an indicator that the ITT estimator may not differ from ATOT). Given that the average number of places constructed per child over the period was 0.09, the average increase in pre-primary school attendance as a consequence of the program is estimated to be approximately 7.317 percentage points. Hence, the program explains about half of the 15 percentage point increase in the gross enrolment rate experienced from 1991 to 2001.

As regards the effect of pre-primary on subsequent school outcomes, the authors would, in principle, like to compare test scores of students who were offered a pre-primary school place to the counterfactual—i.e. test scores for the same students if they were not offered a place. Since the counterfactual is never observed and we do not have a controlled randomized experiment, we turn to non-experimental methods. Specifically, they exploit the variation introduced by the program's expansion over time that generated differences in exposure by cohort and municipality. Authors add interactions between pretreatment (1991) preschool enrolment at the municipality level and cohort dummies to the previous model. The idea is to allow for idiosyncratic trends in municipality enrolment levels in pre-primary education. Municipalities start with different enrolment rates and therefore school performance may naturally grow at different rates, which could be systematically correlated with the construction program. However, the data reject this hypothesis as the point estimates do not significantly change. They find that an increase of one preschool place per child increases Mathematics test scores by 4.69 points and Spanish test scores by 4.76 points. They also find that preprimary school attendance positively affected student's behavioral skills such as attention, effort, class participation, and discipline. This positive effect on behavioral skills provides evidence of possible pathways by which pre-primary might affect subsequent primary school test performance as preschool education facilitates the process of socialization and selfcontrol necessary to make the most of classroom learning. Moreover, behavioral skills are as important as cognitive skills to future success in life.

To test the robustness of their estimates they estimate alternative explanations for their findings. They have already controlled for time invariant differences between municipalities, schools, and cohorts, and for time varying differences at the provincial level

such as school policy or changes in the economic environment. They also consider an additional robustness test that relies on a false experiment that tests for the presence of time varying factors at the municipality or school level that could have affected primary school outcomes, while the second investigates whether the effects can be explained by the migration of students from private to public schools. In order to test the causal interpretation of the results against possible omitted time-varying municipality-level factors, they test whether the expansion of preschool places is correlated with the performance of sixth and seventh grade students. During the period studied (1995-1999), none of the cohorts of students in sixth and seventh grade could have been affected by the construction program. The authors find very small point estimates, which are not statistically different than zero, for both Mathematics and Spanish. Similarly they cannot reject the null that the effects on the behavioral measures are statistically significantly different than zero. These results suggest that underlying municipality or school trends in test scores and classroom behavior that are systematically correlated with the program are not driving the findings.

*Regression Discontinuity Design*

Even in the absence of random assignment and without using the previous Quasi-Experimental Design, there is still another alternative researches can use to identify treatment effects. It could be the case that units are not randomly assigned to treatment, but on the score each unit gets in an assignment variable. This variable should be measured at the pre-test, and the researcher must stipulate a cut-off value, so that all the units with score above that cut-off will receive treatment and the ones below will belong to the control group (or vice versa). This is called Regression Discontinuity Design (RDD).

For examples, units willing to participate are divided into two groups according to whether or not a pre-intervention measure exceeds a known threshold, but only units scoring above that threshold are assigned to the program.

For identification at the cut-off point to hold it must be the case that any discontinuity in the relationship between the outcome of interest and the variable determining the treatment status is fully attributable to the treatment itself. Then, in the neighborhood of the cut-off point or threshold, this technique provides some characteristics analogue to an experiment. The comparison of mean outcomes for participants and non-participants at the margin allows one to control for confounding factors and identifies the mean impact of the intervention, but only locally at the cut-off for selection into treatment.

The assignment to treatment must be only done on the bases of the assignment variable, which plays a crucial role in RDD. The selection criteria must be strictly followed as if it were a random rule like the toss of a coin. It is necessary that the assignment variable is not caused by treatment (that is why it should be measured before the intervention takes place) and must not be correlated

with outcome. The assignment variable can even be a pre-test, but not necessarily; in fact it can be not related at all with the post-test variable (it could be year of birth, order of application to a program, etc.). Apart from these characteristics, it must be ordinal and preferably continuous because it makes it more plausible to find a regression line between the assignment and outcome variables at both sides of the cut-off point. If we used a binary variable to select units into treatment, we will not be able to estimate a regression line for either group; furthermore, there will be a high correlation between the assignment variable and the treatment dummy.

The choice of the cut-off point is also very relevant to the identification of the causal effect in the margins of it. Sometimes the researcher has the chance to choose it, and other times it has to do with exogenous factors not controlled by the researcher. We have to take into consideration that our aim is to estimate a regression line at both sides of the cut-off, so it should not be too low or to high, in order to make sure that there are enough observations at both sides. Ideally it could be place at the mean of the assignment variable. In addition to these, it is important to specify correctly the type of relationship that there is between the assignment and outcome variable at both sides; it could be a line or any functional form, such as a polynomial of a higher level.

RDD can be easily compared to randomized experiments in which assignment to treatment is done by chance instead of using a fixed rule. To estimate the causal effect in the case of randomized experiments, as we have seen throughout the previous chapters, we compare the post-test means of the treatment and control group, under the underlying assumption that both groups are probabilistically similar so there is no selection bias or treatment heterogeneity. In RRD, we do something analogue, but instead of comparing mean outcomes, we compare regression lines at both sides of the cut-off point. This is a Quasi-Experiment, in fact! When there is no treatment, we assume that the regression lines are equivalent instead of claiming that the mean of the post-test is the same. However, note that we need more observations in RDD than in a randomized experiment, to attain the same statistical power.

There are two kinds of RDD: sharp (which is the case described above) and fuzzy (when there is imperfect compliance with the assignment rule at the cut-off point). More formally, in a RDD we need an observable assignment variable, that will be S. This variable will have a cut-off point in which we know that the probability of being assigned to treatment changes from 0 to 1 (or from 1 to 0), in the case of a Sharp RDD; and changes discontinuously in a Fuzzy RDD.

The condition to identify the causal effect is that $Y_0$ conditional on S is a continuous function of S at $\overline{s}$, which means stating that the outcomes of those that not receive treatment show no discontinuity at the cut-off point. Or, in other words, that the discontinuity is only the consequence of treatment.

Let's deduce the identification condition. Remember that S is the observable assignment variable, and let $\overline{s}$ be the cut-off point. So units marginally below and above $\overline{s}$ will be denoted as $\overline{s}^-$ and $\overline{s}^+$ respectively. Then, for $\overline{s}$ be the cut-off point:

$$\Pr(D=1 \mid \overline{s}^+) \neq \Pr(D=1 \mid \overline{s}^-)$$

Imagine that above the cut-off point we assign treatment, so:

$$\Pr(D=1 \mid \overline{s}^{+}) - \Pr(D=1 \mid \overline{s}^{-}) > 0$$

And especially for a Sharp RDD:

$$\Pr(D=1 \mid \overline{s}^{+}) - \Pr(D=1 \mid \overline{s}^{-}) = 1$$

Returning to our expression of the outcome:

$$Y_i(u) = Y_0(u) + \delta D(u)$$

We can re-express it as:

$$\begin{aligned} &\mathrm{E}\left[\mathrm{Y} \mid \overline{s}^{+}\right] - \mathrm{E}\left[\mathrm{Y} \mid \overline{s}^{-}\right] = \mathrm{E}\left[\mathrm{Y}_0 \mid \overline{s}^{+}\right] - \mathrm{E}\left[\mathrm{Y}_0 \mid \overline{s}^{-}\right] + \mathrm{E}\left[\mathrm{D(s)} \mid \overline{s}^{+}\right] - \mathrm{E}\left[\mathrm{D(s)} \mid \overline{s}^{-}\right] \\ &= \mathrm{E}\left[\mathrm{Y}_0 \mid \overline{s}^{+}\right] - \mathrm{E}\left[\mathrm{Y}_0 \mid \overline{s}^{-}\right] + \mathrm{E}\left[\delta \mid \overline{s}^{+}\right] \end{aligned}$$

The condition to identify the causal effect in a Sharp RDD is that $Y_0$ conditional on S is a continuous function of S at $\overline{s}$, which means stating that the outcomes of those that not receive treatment show no discontinuity at the cut-off point. This would identify the causal effect on those in the right hand-side of the cut-off. To extend the condition to those in the left-hand side, we should state that $Y_1$ is continuous. Those conditions would be:

$$\begin{aligned} &\mathrm{E}\left[\mathrm{Y}_0 \mid \overline{s}^{+}\right] = \mathrm{E}\left[\mathrm{Y}_0 \mid \overline{s}^{-}\right] \\ &\mathrm{E}\left[\mathrm{Y}_1 \mid \overline{s}^{+}\right] = \mathrm{E}\left[\mathrm{Y}_1 \mid \overline{s}^{-}\right] \end{aligned}$$

These implies that if Y is continuous on S in the cut-off, then the outcomes and the treatment assignment are orthogonal given S:

$$(Y_0, Y_1) \perp D \mid S = \overline{s}$$

Then, we could identify the average treatment effect on the treated only in the neighborhood of the cut-off point:

$$\mathrm{E}\left[\delta \mid \overline{s}\right] = \mathrm{E}\left[\mathrm{Y} \mid \overline{s}^{+}\right] - \mathrm{E}\left[\mathrm{Y} \mid \overline{s}^{-}\right]$$

The treatment effect can be estimated by the outcomes of the units on the neighborhood of the cut-off point if the sample size is large enough. If it is not, a function of Y on S have to be estimated on each side of the threshold, using all the observations lying on both sides of it.

The Fuzzy RDD appears when compliance in the threshold is not complete. Thus, some units in the upper side although must belong to the treatment (control) group, are observed in the control (treated) group, and the same happens in the other side of the cut-off point. In this design the identification conditions for the causal effect stated for the Sharp RDD are no longer enough. This is because the treatment status does not only depend on the assignment variable (which we called S)

but on unobservable variables. Then, we need to state that at least in the neighborhood of the cut-off point, the assignment variable is independent of the potential outcomes, and only affects them through the treatment D. This situation brings us again to our familiar IV framework, where S will be the instrument, D is the endogenous treatment status because it affects the outcomes and also is affected by unobservables (not ruled out by the cut-off in S as in the Sharp Design) and Y is the observed outcome.

**References**


Berlinski, S., S. Galiani and P. Gertler ( 2009): "The effect of pre-primary education on primary school performance," *Journal of Public Economics, Elsevier, vol. 93(1-2), pages 219-234, February.*

Morgan, S. and C. Winship (2007): "Counterfactuals and Causal Inference", Cambridge University Press. Chapter 1

Wooldridge, J. (2002): "Econometric Analysis of Cross Section and Panel Data", The MIT Press. Chapter 18.

Shadish, W, T. Cook and Campbell D. (2002): "Experimental and Quasi-Experimental Designs for Generalized Causal Inference", *Houghton Mifflin: Boston*. Chapters 1-3.

Ahn, H. and J. Powell (1993, July). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics 58 (1-2), 3{29.

Angrist, J (1995): "Condition on the Probability of Selection to Control Selection Bias", *NBER Technical Working Paper 181.*

Ashenfelter, O. (1977), "Estimating the effect of training programs on earnings", *The Review of Economics and Statistics, Vol. 60 No.1, pp.47-57.*

Heckman, J. (1979): "Sample Selection Bias as a Specification Error", *Econometrica 47, 153-161*

Imbens, G, J. Angrist and D. Rubin (1996): "Identification of Causal effects Using Instrumental Variables", *Journal of Econometrics, Vol. 71,No. 1-2, 145-160*

Lee, D. (2005): "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *October 2005. Unpublished manuscript, University of California, Berkeley.*

Rosenbaum, P. (2002): Observational Studies, *Springer*. Chapter 2.

Shadish, W, T. Cook and Campbell D. (2002): “Experimental and Quasi-Experimental Designs for Generalized Causal Inference”, *Houghton Mifflin: Boston*. Chapters 1-3.

Campbell, D. and Stanley, J.(1963): "Experimental and Quasi-Experimental Designs for Research on Teaching." *In N. L. Gage (ed.), Handbook of Research on Teaching. Chicago: Rand McNally*.

Galiani, S. and E. Schargrodsky (2007): “Property Rights for the Poor: Effects of Land Titling”, *Mimeo*.

Haavelmo, T. (1944): “The probability approach in econometrics”, *Econometrica 12, pp. iii-vi+1-115.*

Holland, P. (1986): “Statistics and Causal Inference”, *Journal of the American Statistical Association 81, pp. 945-70.*

Rubin, D. (1974): “Estimating causal effects of treatments in randomized and nonrandomized experiments”, *Journal of Educational Psychology* 66, pp. 688-701.

Goldberger, A. (1972): “Structural Equations Methods in the Social Sciences”, *Econometrica* 40, pp. 979-1002.

Pearl, J. (2000): Causality: Models, Reasoning and Inference, *CUP*. Chapters 1, 5 and 7.

Friedman, M., C. Furberg and D. Mets (1998): “Fundamentals of Clinical Trials”, *3rd ed. New York, NY: Springer.*

Heckman, J. and J. Smith (1995): “Assessing the Case for Social Experiments”, *The Journal of Economic Perspectives, Vol. 9, No. 2 (Spring, 1995), pp. 85-110*

Kling, J., J. Ludwig and L.Katz (2005): “Neighborhood Effects on Crime for Female and Male Youth: Evidence from a Randomized Housing Voucher Experiment”, *Quarterly Journal of Economics.*

Duflo, Esther and Kremer, Michael (2006), "Using Randomization in Development Economics Research: A Toolkit". *NBER Working Paper No. T0333* Available at SSRN: http://ssrn.com/abstract=953351

Rosenberger W.F., Lachin J.M.. (2002): “Randomization in clinical trials”. *New York, NY Wiley*.

Cameron, C., Gelbach, J.B. and Miller, D.L. (2007), “*Bootstrap-Based Improvements for inference with Clustered Errors*”, mimeo, Florida State University. Available at SSRN: http://ssrn.com/abstract=956890

Casella, G. and Berger, R.L (1990). “*Statistical Inference”*. Belmont, CA: Duxbury Press.

Horowitz, J.L. (2001). “*The bootstrap*” in: J.J. Heckman and E.E. Leamer, Editors, Handbook of Econometrics vol. 5, Elsevier Science, North Holland.

Murray, D.M. (1998) "*Design and Analysis of Group Randomized Trials"*. New York, NY: Oxford University Press.

Rubin, D.B. (1974), *"Estimating causal effects on treatments in randomized and nonrandomized experiments"*, Journal of Educational Psychology Vol. 66; 688-701.

Miguel, E. And M. Kremer (2004), "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities", Econometrica, 2004

Banerjee, A., C, Shawn, E. Duflo and L. Linden (2007), "Remedying Education: Evidence from Two Randomized Experiments in India", The Quarterly Journal of Economics, MIT Press, vol. 122(3), pages 1235-1264, 08