IMPACT EVALUATION METHODS

Dina Pomeranz[*]

August 2011

Daily decisions made in tax administration can affect the economy of an entire country. How are these decisions made? Are they good or bad decisions? The objective of an impact evaluation is to provide information about the effects of current and potential policies. There are various evaluation methods, each with different degrees of validity. The quality of the evaluation is of utmost importance for obtaining correct results. This document offers a brief summary of the most common methods, their advantages and disadvantages as well as a description of the conditions under which each method produces reliable results.

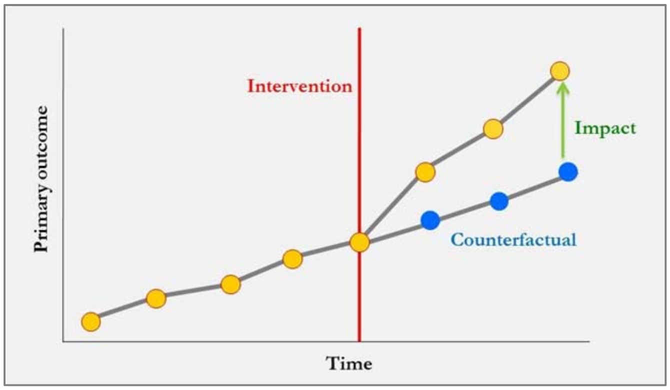
Before presenting the specific methods below, let's start with some basic concepts:

The objective of every impact evaluation is to demonstrate a causal effect: The goal is to measure the impact of a program or policy on some variable of interest. For example, what is the effect of a notification letter on tax amendments? There is a cause and effect -- the cause is the change in the policy or the implementation of a new program and the effect is the result that can be attributed directly to the new policy or program.

The difficulty of measuring the impact is that it is only possible to observe what happened, not what would have occurred without the program. We observe a tax payer that received the notification filing an amendment, but not what he/she would have done without the warning; that is, we don't know if the tax payer would have made the same modification. This imaginary situation, what would have happened without the program, is called the counterfactual. Understanding the ***counterfactual*** is key to understanding the impact of a program.

If an accurate representation of the counterfactual existed, then impact evaluation would be easy. The impact of the program or policy is the difference between the result we observe with the program and the result that would have prevailed without the program –the counterfactual.

FIGURE 1 - COUNTERFACTUAL



SOURCE: J-PAL (2010)

Given that the counterfactual does not exist in reality, since it's what would have happened under a different scenario, each evaluation tries – in an explicit or implicit manner- to construct an estimate of the counterfactual to compare it to what occurred. Normally, the counterfactual estimate is represented by a group called the ***control group or comparison group***. The control group consists of people or firms that didn't participate in the program, while the ***treatment group*** is the group that participated in the program. To measure the impact of the intervention, the treatment group is compared with the control group.

[*] Harvard Business School , Rock Center 213, Soldiers Field Road, Boston, MA 02163, dpomeranz@hbs.edu.

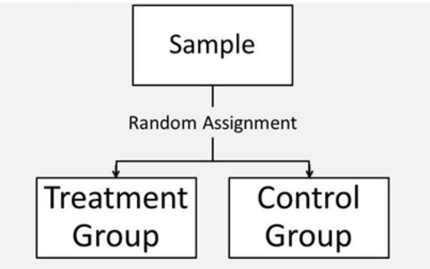
An evaluation will produce reliable results if the control group is identical to the treatment group in all its characteristics – observable or not – except one: their exposure to the treatment. In this case, any difference after the intervention can be attributed to the program, given that in its absence, both groups would be the same.

All methods used to construct the comparison group rely on certain assumptions under which the control and treatment group would be comparable. When the assumptions are realistic, the control group is a good representation of the counterfactual. Nevertheless, when the assumptions are not realistic, the impact evaluation will be **biased**. A biased evaluation may result in bad decisions and generates losses in terms of effort, time and public resources.

Therefore, it is important to use high quality methods and make explicit assumptions when using an evaluation technique. In the next section, the characteristics, strengths, and limitations of the different evaluation methods are presented.

1. Randomized Evaluation

FIGURE 2:
THE DESIGN OF A RANDOMIZED EVALUATION



Randomized evaluations (or experimental evaluations) create the ideal comparison group. The random assignment ensures there is no difference between the individuals in the treatment and control group, except the fact that one group has been randomly chosen to participate in the program and the other has not. Therefore, randomized evaluations are the ideal case for an impact evaluation. It is for this reason that in the evaluation of new medicines and in natural science research, this method is used almost exclusively.[1]

Nevertheless, randomized assignment requires that the evaluation is designed ***before*** the program has begun. For this reason, this method is also called prospective evaluation. In a random process, individuals (or other entities like firms) are assigned to the treatment group and those not selected are part of the control group. The randomization process can be something as simple as tossing a coin or a conducting a lottery. The random assignment is usually done using a simple process in Excel or Stata. It is not necessary for both groups to be the same size.

Randomization in Practice

It is important that the randomization process is truly random and not just a process that "seems" arbitrary. For example, assigning the treatment to people whose surnames start with the letters "A-L" and leaving those starting with "M-Z" as control may seem random, but it is not. Such assignment requires the assumption that the individuals whose surnames start with the letters "A-L" are the same as those that start with "M-Z". Nevertheless, it is possible that the families whose surnames start with the letters "A-L" are different from the families with a last name starting with the letters "M-Z". For example, the ethnic composition may vary. To avoid this situation, an automatic method like using a computer to generate random numbers that determine treatment assignment is recommended.

A computer also simplifies more complex randomization processes, like stratified randomization. Stratified randomization is recommended when the number of potential participants is small to ensure that both groups are balanced with respect to the most important variables. In stratifying, the sample is divided into sub groups of similar characteristics and each subgroup is then randomized. For example, if the population is divided by gender and a 30% of men and a 30% of women are assigned the treatment, this assignment will be perfectly balanced in terms of gender. The treatment group will have the same gender composition as the control group.

According to the law of large numbers, when there is a sufficient number of people in each group, random assignment will generate two groups that will be similar in all their observ-

[1] It is important to distinguish between a randomized evaluation and a random sample: Many studies use random samples to obtain representative information about a population. A random sample does not try to measure impact. The distinctive characteristic of a randomized evaluation is that the treatment is assigned randomly.

able (like education), and unobservable (such as motivation) characteristics. Therefore, any difference that arises later between the treatment and control groups can be attributed to the program and not to other factors. For this reason, if designed and applied adequately, a randomized evaluation is the most reliable method for measuring the impact of a program.

How do we determine the number of participants required in a randomized study? According to the law of large numbers, the greater the number of individuals included in a study, the more likely it is that both groups will be similar. This is one of the reasons why sample size is important. A larger sample is always better since it reduces the likelihood of having unbalanced groups. Nevertheless, a bigger study can be more costly and is not always feasible. Therefore, it is recommended that the statistical power be calculated to determine the sample size necessary for measuring the impact on the main outcome variables of interest.

Statistical power calculations incorporate the different factors that affect the number of required participants. Among the factors to be considered is the variance of the variable of interest and the minimum effect expected to be detected. The higher the variance of the dependent variable, the greater the number of observations necessary for being able to detect a statistically significant effect. Likewise, if the effect to be measured is small, the larger the number of required participants. Finally, the randomization design can affect the necessary group size. If the randomization is performed at the group level (***conglomerate design***), keeping all firms that use the same accountant together in the same group for example, more firms will be necessary than if the randomization is done at the individual level.

Randomized Evaluation: steps to follow

1) Choose a program, population and main outcome variables of interest.

2) Statistical power estimates: Determine the size of the treatment and control groups required for measuring the impact on outcome variables of interest.

3) Random assignment of treatment. Verify that the assignment is balanced with respect to the main variables of interest.

4) Pilot: Implement the program on a small scale to avoid unexpected problems (if possible)

5) Implementation: Make sure there is no difference between the treatment and control groups.

After determining the number of participants required, the treatment can be randomly assigned. It is important to verify that the groups are balanced with respect to the main outcome variables of interest. In the academic literature, experimental studies usually include a balance table that shows that the main characteristics are similar across the two groups.

Finally, the implementation of the program or policy to be evaluated is carried out. In many cases, a smaller scale pilot of the intervention is recommended to test all procedures and avoid unexpected problems in the implementation process. During this step, it is important to make sure that the random assignment of individuals to each group is respected and that no participant is moved from one group to another.[2] The most important thing in this process is to make sure that there is no difference between the treatment and control group except the application of the program. For instance, the validity of the study would be lost if other auditing activities in the control group end, but continue to be applied to the treatment group, or vice versa.

[2] In the case that the randomization is not respected in the implementation process, it is possible to use the "Intent-to-Treat" methodology, and use instrumental variables to observe the "Treatment-on-the-Treated" effect. For example, this method could be used if tax payers who were supposed to be audited as a result of being in the treatment group cannot be found when the audit is to be carried out, or if letters are sent to the tax payer as a treatment but are not received. It is very important that the original random assignment is used when working with the data to conduct the impact the evaluation; that is, those that were assigned the treatment are compared to those assigned to be the control. It is never valid to compare those who were in fact treated with those with those that were meant to be treated but that ultimately did not participate in the program because these two groups will no longer be identical ex ante.

This concludes the summary of randomized studies. However, in many cases, it is not possible to assign policies or programs randomly. In the following sections, we describe other evaluation methods that try to construct an approximation of the counterfactual under certain assumptions. The validity of each method will depend on how similar the treatment group is to the control group before the intervention.

Summary: Randomized Evaluation

Description: Experimental method that allows casual relationships between two variables to be measured by comparing those treated with those that are not when participation is randomly determined

Counterfactual representation: The comparison group is selected randomly before the start of the program within a group of potential participants.

Key assumptions: The randomization is valid. That is, both groups are statistically identical (in both observable and unobservable characteristics). No other treatment is applied to any of the groups.

Advantages: The estimate of the impact of the program is reliable when it has been designed and implemented correctly.

Disadvantages: Requires that the random assignment is done before the program. As a result, it is not possible to carry out retrospective randomized evaluations. The sample size must be sufficiently large to be able to detect a significant result.

2. SIMPLE DIFFERENCE (TREATED VS. UNTREATED)

The ***simple difference method*** is one of the most common methods. The methodology is straightforward: comparing the group that received the program with another that did not. Nevertheless, to have a good representation of the counterfactual, the control group must represent what would have happened to the treatment group without the program. Is this assumption real? Unfortunately, in many cases, the answer is no.

In many programs, there is a selection process that determines who receives the treatment. Sometimes the selection is explicit; for example, an audit program in which only tax payers identified as high risk are selected. The selection can also result in something not explicit or observable; for instance, if the auditors chose the contributors that behave in some irregular manner. In either case, this assignment is not random and introduces a **selection bias**. That is, the untreated group and the treated one are no longer the same prior to the implementation of the program. The difference that is observed between the groups could be the result of the program or of the initial differences between the two groups, or a combination of both.

For example, suppose there is a program that offers free tutoring for children who have difficulty in school and we want to measure its impact. If we simply compare the grades of those children that received help from a tutor with those that did not, it is possible that the children with tutors have grades that are lower than those without tutors. Concluding, based on this observation, that the tutors hurt the academic achievement of the kids would probably be erroneous. It is more likely that there was an initial selection bias in which the children with lower grades had a greater chance of receiving the help of a tutor. In this case, the selection bias introduces a strong underestimation of the impact such that the effect seems negative instead of positive.

Summary: Simple Difference

Description: Measures the differences between those that participated of a program with those that did not after it is applied.

Counterfactual representation: The comparison group corresponds to those individuals that did not participate in the program for any reason, and for which we have data after the program.

Key assumptions: Those that didn't participate are identical to the participants except for the intervention of the program. There is no selection in the type of person that entered the program.

Advantages: Administrative data can frequently be analyzed retrospectively. It does not require data on the situation prior to the program.

Disadvantages: Requires a group that was not affected by the program. If the treated group and untreated groups are different prior to the program, the method may under or over value the real impact of a policy; that is, a selection bias is introduced in the estimation.

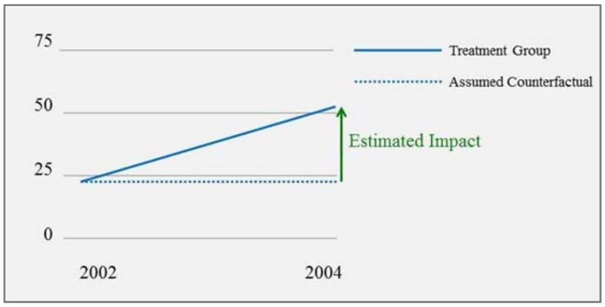
3. Pre-post (before vs. after)

A ***pre-post*** evaluation is a particular type of simple difference evaluation. Instead of using another group of persons as a control group, the same group of people is used *before the start of the program*.

Therefore, a pre-post evaluation measures change over time taking into account the initial state of the group. In this case, the impact is measured as the difference between outcome variables of interest before and after an intervention. The pre-post analysis is frequently used in evaluating programs. In many cases when there is data on outcomes prior to the intervention, this type of retrospective analysis seems convenient.

A pre-post evaluation allows us to take into account the initial education level of the students. But, is the group of persons before the start of the program a good representation of the counterfactual? In other words, is it correct to assume that without the program, during this period, there would not have been any change in the results of the treated group?

FIGURE 3: COUNTERFACTUAL ASSUMPTION FOR PRE-POST



SOURCE: J-PAL (2010)

Let us look at this situation using the free tutoring program example. Is it believable to assume that in the 2 years of the program, the children would not have improved their grades without the tutors? In reality, it is likely that the students would have continued learning and improved their knowledge. If a pre-post evaluation is done, this learning, which is normal in the development of children, would be attributed to the tutoring program.

This natural evolution of outcomes over time is called ***secular tendency***. In addition to the secular tendency, there can be “shocks” that change outcomes, but are not related to the program. For example, if there is an economic crisis during the implementation period of an auditing policy, tax behavior may vary independent of the policy. It is not possible to know if the change over time is due to the crisis, the policy, or a combination of both.

Summary: Pre-post Evaluation

Description: Measures the change in outcomes over time for participants of a program. It is the difference between the situation before and after a treatment.

Counterfactual representation: The comparison group is represented by the same participants, but prior to the program.

Key assumptions: The program is the only factor that influenced a change in outcomes. Without the program, the outcomes would have remained the same.

Advantages: In many instances, there exists administrative data that can be analyzed retrospectively. It does not require information on people that did not participate in the program.

Disadvantages: Many factors that vary over time can affect an outcome, which contradicts the key assumption made above. In particular, the pre-post comparison does not control for secular tendencies or shocks that are unrelated to the program but that affect outcomes.

4. DIFFERENCES IN DIFFERENCES (DIFF-IN-DIFF)

A ***differences-in-differences*** evaluation combines the two previous methods to take into account both the differences between the two groups and the secular tendencies.

FIGURE 4 – ESTIMATING DIFFERENCES-IN-DIFFERENCES

	Result before the program	**Result after the program**	**Differences**
Treated group	24,80	51,22	**26,42** ①
Untreated group	36,67	56,27	**19,60** ②
Differences-in-differences estimation:			**6,82** ③

SOURCE: J-PAL (2010)

The difference in differences methodology uses both variations: The difference over time and the difference between the groups. To calculate the effect, the change over time for the treated group is calculated (1), then the change for the untreated group (2) and then the difference between these two differences is calculated (3).

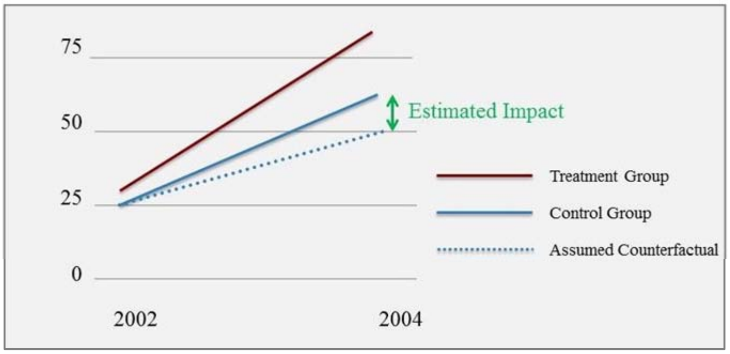
In a multivariable regression, the difference in differences is seen in the interaction term between the treatment group and the post treatment period:

$$Y_{it} = \alpha + \beta_1\, T_i + \beta_2\, post_t + \beta_3\, T_i * post_t + \epsilon_{it},$$

where Y_{it} represents the variable of interest for individual i in period t, T_i is a binary variable indicating whether or not individual i received the program, and $post_t$ is a binary variable indicating the period following the program. β_3 is the estimator of the difference-in-differences.

In essence, the differences in differences estimation uses both the change over time for the untreated group as well as the estimation of the counterfactual for the change over time for the treated group. The key assumption is that without the program, the tendency in both groups would have been the same. This is the ***assumption of common or parallel tendencies***. This assumption is not met if the treated group would have had a different tendency than the control group in the absence of the program.

FIGURE 5: COUNTERFACTUAL ASSUMPTION IN DIFFERENCES-IN-DIFFERENCES



SOURCE: J-PAL (2010)

In the case of the student tutoring program, the assumption implies that without the additional help, the children with a tutor and those without one would have improved their scholastic achievement at the same rate. Nevertheless, it is possible that even without the program, the slower kids may have improved more than the advance ones given that they had more room to improve, or vice versa. It is possible that without the tutors, the gap between the slower kids and the more advanced ones would have become larger. In both cases, we don't know if the difference in the difference is due to the characteristics of the groups, the tutoring program, or a combination of both.

Summary: Differences-in-Differences

Description: Compares the change in participant outcomes with the change in the outcomes for those that didn't participate in the program.

Counterfactual representation: The change for those who do not participate in the program represents of the counterfactual of the change for those that did participate in the program.

Key assumptions: Assumption of common tendencies: assumes that without the program, both groups would have had identical trajectories throughout the period.

Advantages: Controls for all the characteristics that do not change over time (both observable and unobservable) and for all the changes over time that affect the treated and untreated group in the same manner.

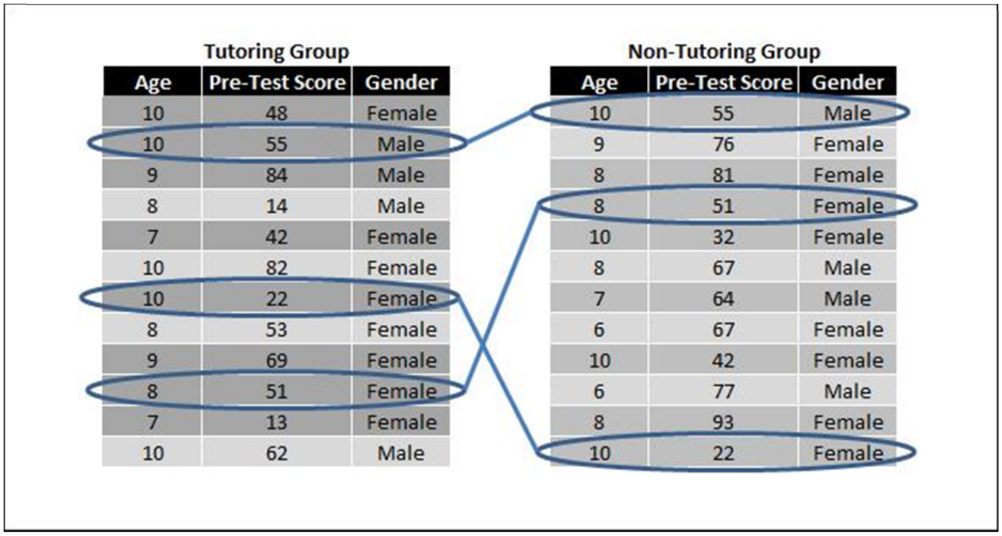
Disadvantage: If both groups would have developed in a different way in the absence of the program, there will be a selection bias. This method requires a group that was not affected by the program as well as information about this group prior to the intervention.

5. Matching and Propensity Scores

With ***matching*** we come back to the original objective of constructing a counterfactual representation and creating a group that is the same as the treated group. Matching constructs an identical group in terms of the observable characteristics prior to the program. There are several matching methods; the basic case, described below, is one in which each individual in the treated group is matched to an individual with the same observable characteristics in the untreated group. To estimate the impact of a program, the method compares the outcomes between the treatment group and the control group, which is composed of individuals with characteristics identical to the treated individuals. Given that both groups have the same observable characteristics before the program, it is expected that the only difference after the program will be having been exposed to it.

In the tutoring program example, for instance, we could find children that did not sign up for the program, but that had the same grades as a kid that received the help of a tutor before the intervention. This way we create a group of all those that were treated and a group with identical treated peers; that is, individuals that are not treated but have the same observable characteristics. Figure 6 shows the peer selection process with three characteristics: ages, pre-test score and gender.

FIGURE 6 - MATCHING PROCESS IN THE TUTOR EXAMPLE



Tutoring Group

Age	Pre-Test Score	Gender
10	48	Female
10	55	Male
9	84	Male
8	14	Male
7	42	Female
10	82	Female
10	22	Female
8	53	Female
9	69	Female
8	51	Female
7	13	Female
10	62	Male

Non-Tutoring Group

Age	Pre-Test Score	Gender
10	55	Male
9	76	Female
8	81	Female
8	51	Female
10	32	Female
8	67	Male
7	64	Male
6	67	Female
10	42	Female
6	77	Male
8	93	Female
10	22	Female

SOURCE: J-PAL (2010)

In certain cases, matching can be better method than differences in differences because the process of finding peers ensures that the two groups are identical in the observable characteristics that we consider important. But is it reasonable to assume that the treated group is identical to the group that is similar according to observable characteristics?

The problem with this method is that matching can never control for unobservable variables. In the tutoring program example, there is a non-random reason that two children with the same grades receive a different treatment. Does the teacher think that one student has more potential than the other? Does one student have more supportive parents who looked for a tutor? If there is something that is not reflected in our data or that is difficult to measure (for example, parental motivation) that influences the results, then we return to the selection bias problem. It is likely, for example, that a kid with supportive parents would have improved more than his classmate with the same grades, even without the tutoring program.

Apart from the fact that some characteristics cannot be observed, another challenge in matching is that it requires individuals with the same characteristics in both the treated group and the untreated group. This requirement is called the ***common support condition***. In the tutoring program example, if all students with very low grades received help from a tutor, it would not be possible to match based on grades.

Finally, the more characteristics we want to include in the matching, the harder it is to use this method. With many data points, like the census of all students in the country for example, it may be impossible to find a comparable student that did not have a tutor. On the other hand, with smaller samples, it is possible that certain individuals in the treatment group may not have an identical peer in the untreated group.

For these reasons ***"Propensity Score Matching" (PSM)*** was developed. PSM allows a matching with many characteristics. The number of characteristics is reduced to a single index that predicts the probability of participating in the program. In effect, the index is a weighted average of the subjacent characteristics. The matching is then done between individuals that have the same likelihood of participating in the program.

Summary: "Matching"

Description: Compares the outcomes of treated individuals with those of similar individuals that were not treated**.**

Counterfactual representation:

Exact Matching: For each participant, at least one identical individual (for selected characteristics) is selected that did not participate.

Propensity Score Matching (PSM): Participants of the program are compared to those that did not participate, and that according to their observable characteristics, had the same probability of participating in the program.

Key assumption: Those who do not participate are, on average, identical to the "paired participants," except for having participated in the program**.**

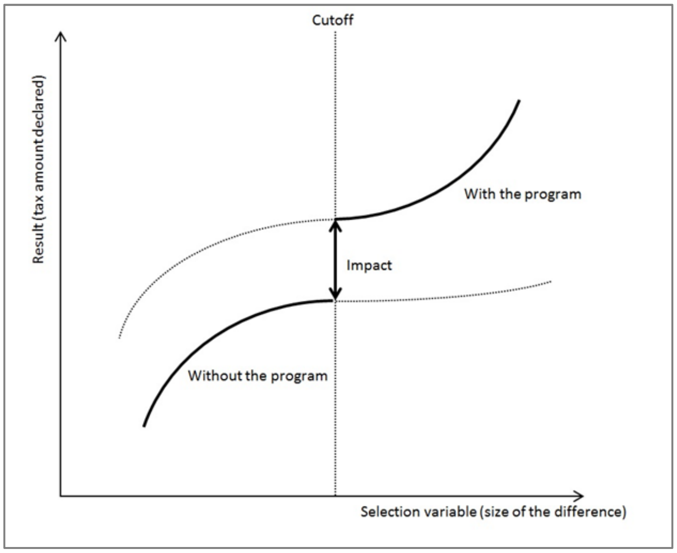
Advantages: Does not require randomization prior to the program. It can give us not only the average impact of the program, but also the distribution of the impact of the program.

Disadvantages: It is possible that there exist unobservable characteristics that affect the probability of participating in the program as well as the outcomes. This introduces a selection bias. Knowing the likelihood that unobservable characteristics will be important in this context requires fully understanding how the participants of the program were selected.

6. REGRESSION DISCONTINUITY

There is a methodology that allows causal conclusions that are as reliable as the randomized control trial that can be applied in certain special cases. Sometimes programs or policies have a specific threshold that determines who receives a treatment. A ***regression discontinuity design*** uses the fact that the individuals or firms near the threshold are basically identical. Under certain assumptions, it is possible to interpret the difference between the outcomes of the individuals just under the threshold that do not participate the program and the outcomes of those just above that do participate in the program as the impact of the intervention.

Let us assume, for example, that a program sends a letter to all firms with discrepancies between their reports and those supplied by a third party that are greater than 100 dollars. In this case, the size of the difference is the selection variable because the cutoff is defined by this variable.


Cutoff
Result (tax amount declared)
With the program
Impact
Without the program
Selection variable (size of the difference)

Error! Reference source not found. displays the concept of a regression discontinuity evaluation. The solid line represents the relationship between the size of the difference and the tax amount declared: the larger the difference, the more tax is declared. It is possible to see in the cutoff line region, the threshold over which the letter is sent, that there is a discontinuity or "jump" in tax payments. Under certain conditions it is possible to attribute this jump to the sending of the letter.

One of the most important assumptions for the use of a regression discontinuity design is that there was no strategic change in the behavior of the firms around the threshold. If, for instance, the firms just under the 100 dollar difference had good accountants that knew what to do in order to come out just under the limit, then there would be a difference between the firms just under the threshold and those just above. Such a difference around the threshold introduces a selection bias. The manipulation around the threshold is referred to as a **behavioral response to the threshold**.

The advantage of the regression discontinuity method is that the assumption that there is no behavioral response to the threshold can be tested. If a manipulation occurred, there would be a higher concentration of firms (bunching) just above or below the threshold, which can be verified. In the same manner, it is possible to verify that there are no differences in the key characteristics between the firms just above or below the threshold.

Finally, a regression discontinuity design also requires that no other programs or policies are applied to the same threshold. For example, if the firms with differences greater than 100 dollars are also visited by an auditor, it would not be possible to distinguish the impact of the visit from that of the letter.

Both problems, the behavioral response to the threshold and other policies applied to the same threshold, are more frequent when the cutoff known by everyone. Therefore, optimal thresholds for the use of this methodology are secret, or defined ex-post, and are applied in the implementation of a single program.[3]

In a regression discontinuity analysis, the results of firms or individuals above or below a threshold are not simply compared. A regression that controls for a change in the selection variable both in a linear and nonlinear manner (using polynomials) is run. For further detail, refer to the bibliography.

[3] An example of a threshold chosen ex-post is a declared sales amount, defined after declarations are made, under which a certain treatment is applied. In this case, the firms can't adjust their declared sales based on the threshold because the cutoff was not known prior to the declaration. In the case that the threshold is public and known prior to the declaration date, it is very important to determine if there was a behavioral response to the threshold before running the regression discontinuity. If there was manipulation around the cutoff, the regression discontinuity estimator will not be valid.

Summary: Regression Discontinuity

Description: Compares the results of individuals that are just below a threshold that qualifies them for the treatment with the results of the individuals that are just above this threshold (cutoff).

Counterfactual representation: Individual outcomes close to the cutoff, but that fall on the just below the cutoff and do not participate in the program, represent the counterfactual of the individuals that fall just above the threshold and consequently receive the treatment.

Key assumptions: The individuals just above the cutoff are identical to those that fall just below the threshold. There is no manipulation around the threshold and no other policies that are applied based on the same cutoff.

Advantages: Produces very reliable impact estimations. In tax administration, there are many policies that are applied according to some cutoff, and frequently the administrative data required for the analysis already exists. Most of the assumptions can be tested.

Disadvantages: The conclusions can only be applied to individuals or firms around the cutoff. It is not possible to know what the impact would be on those far from the threshold.

Bibliography

General texts

Angrist, Joshua D., and Jorn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press, 2009.

Imbens, Guido, and Jeffrey Wooldridge. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*. 47.1 (2009): 5-86.

Experimental evaluation

Glennerster, Rachel, and Kudzai Takavarasha. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, 2013.

Banerjee, Abhijit, and Esther Duflo. "The Experimental Approach to Development Economics." *Annual Reviews of Economics*. 1. (2009): 151-178.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. "Using Randomization in Development Economics Research: A Toolkit." *Handbook of Development Economics*. 4. (2007): 3895-3962.

Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan. "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives*. Forthcoming (2011).

<u>Differences in differences</u>

Duflo, Esther. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review*. 91. (2001): 795-813.

Abadie, Alberto. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies*. 72. (2005): 1-19.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How Much Should We Trust Differences-In-Differences Estimates?" *Quarterly Journal of Economics*. 119.1 (2004): 249-275.

<u>Matching</u>

Dehejia, Rajeev, and Sadek Wahba. "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*. 94. (1999): 1053-1062.

<u>Regression discontinuity design</u>

Imbens, Guido and Thomas Lemieux. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics*. 142. (2008): 615-635.

Lee, David, and Thomas Lemieux. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*. 48.2 (2010): 281–355.