Blog 104: August 2019

Agricultural Extension in South Asia

WHAT, WHY, AND HOW TO DO IMPACT ASSESSMENT

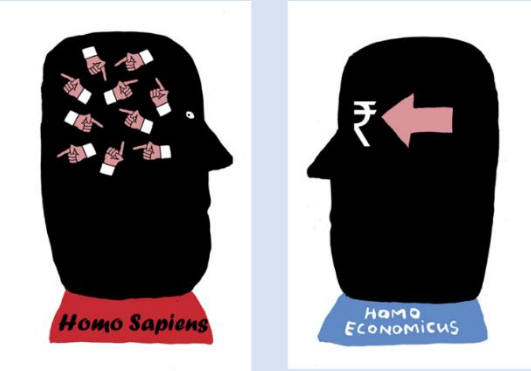
Investors and policy makers often need to know how far the programmes they are financing or implementing are having the intended impact. In this blog KS Aditya and SP Subash take us through the different methods of impact assessment and explain 'why' and 'when' to use these.

BACKGROUND

This blog is an honest attempt to explain the basic concepts of Impact Assessment in a language that Homo sapiens can understand and not just Homo economicus (Box 1). In a nutshell, we will be trying to convince you that scary methodologies used (sometimes many of us might have not even heard of these) and econometric juggleries that we employ in assessing impact is perfectly justified. However, a word of caution before you read any further, if you expect this blog to give you the 'best method to assess impact' we are very sorry to disappoint you – there is no 'gold standard' method, best fit for all cases. Our aim is just to introduce you to the different methods of impact assessment, and more importantly, to tell you why and when to use them, and also when not to use a particular method!

Box 1 (Trivia): *Homo sapiens vs Homo economicus*

Homo economicus or 'economic man' portrays humans (Homo sapiens) as rational ideal agents as defined in economic theory. Thaler and Sunstein (2009) in their book *Nudge* differentiates between the two and states such rational ideal agents don't exist and that humans have inherent biases.



Note: Visualization modified from https://universonline.nl/2017/11/20/little-nudge-right-direction

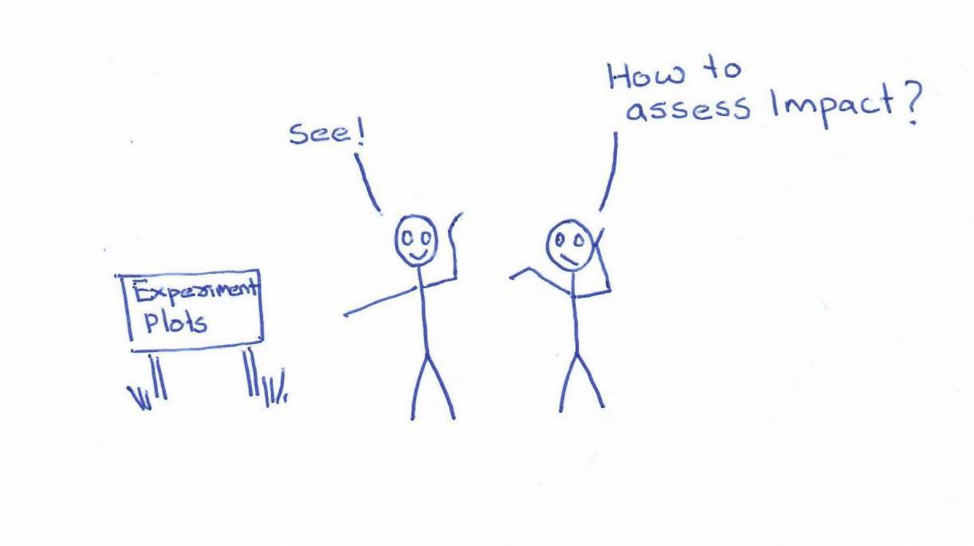
Source: Thaler & Sunstein (2009)

WHY IMPACT ASSESSEMENT?

"*In God we trust, rest bring data*" - Edward Deming

Policy makers need scientific and reliable estimates of how effective a programme or intervention (technology) is. Even the most promising projects might fail to generate the expected impacts. So, policy makers need to know how far the programmes are generating the intended impact. This equips them to take calls on reorienting the programmes as well as in allocating funds. In this line Impact Assessment is an effort to understand whether the impacts of a programme (Net welfare gain - only for readers who belong to species Homo economicus) are attributed to the programme and not to some other causes. Ultimately, the aim of Impact Assessment is to establish the causal link between the programme and the impact, and to arrive at reliable estimate of the 'size of impact'.

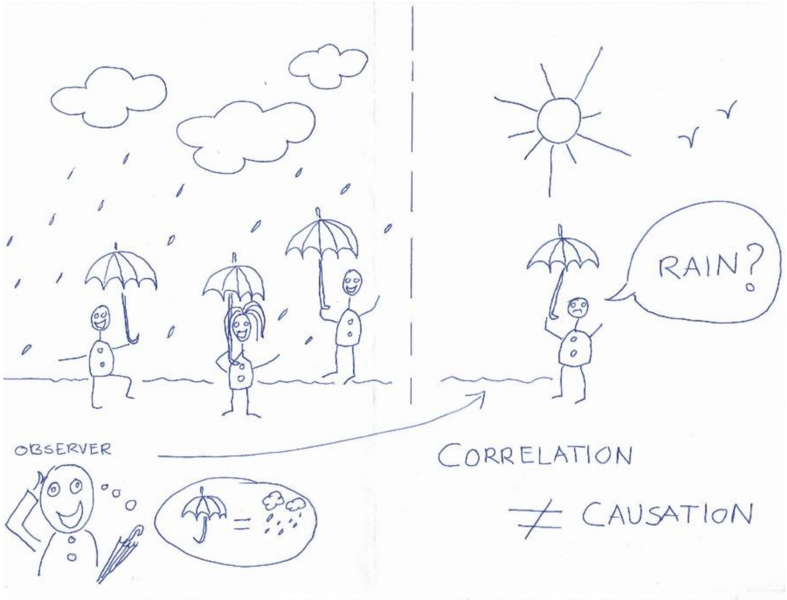
Let us take one case, where a new programme is launched to increase the income of beneficiaries. After a few years, the government wants to know the impact of the programme. One common and very popular approach is to collect data from a few beneficiaries (treatment) and non-beneficiaries (control) and estimate the difference in income between the two groups as impact. However, the difference in mean income between the two groups cannot be called impact, as we haven't yet established 'causation'- how can we say that the difference in income is only due to the programme and not due to other factors? How then to assess impact?



Source: Illustrated by authors

ATTRIBUTING IMPACT

In lab and field experiments carried out by biological scientists, three principles are used to establish causation: replication, randomization, and local controls. Randomization makes sure that the unobservable characters remain the same across treated and control groups. Local control ensures that all the variables, except for the treatment, are same across treatment and control. For example, if the purpose of the experiment is to know the effect of organic manure on crop yield, all other factors like variety, soil type, seed rate, chemical fertilizers, date of sowing, etc., must remain the same across treatment group and control group. If the only difference between treatment and control group is organic manure, we can safely say that increase (or decrease) in yield is due to manure use. To sum up, local control and randomization in case of experiments ensure that the treated and control groups are similar, which enable us to make 'causal claims' by simply taking the mean difference across groups (for which you need to test statistical significance).



Source: Illustrated by authors based on Doug Neill's work (read further https://commons.trincoll.edu/cssp/2013/12/09/10886/)

Let us shift our focus back to social science research. Mostly we do research based on 'observational data'; we collect data from observations (samples) where the researcher has no control over the variables unlike an experiment. So, in most cases, the treated and control groups are not similar with respect to many variables and the difference in outcome variable (Income for example) cannot be attributed to treatment (Programme or intervention) and would result in bias (Bias can be considered as a cousin of error!). More specifically, bias in estimate of impact arising due to the pre-treatment difference in covariates is called 'Selection Bias'. For example, let us say that we would like to know the impact of rice seed treatment on farmer's income. The usual research design would be to collect income and other data from both adopters and non-adopters. However, as per theoretical expectations, adopters of a technology are more motivated, have better education and extension contact compared to non-adopters. So, we cannot say that higher income of adopters is only due to seed treatment as it could also be due to pre-treatment differences in education, motivation, extension contact, etc. We can say that 'difference in income across treatment and control' as an estimate of impact suffers from 'sample selection bias'.

Box 2: Example

Allow us to give you another hypothetical example. Suppose you want to measure the impact of a particular 'badminton coaching center'. You enroll me into the coaching classes for a training course of one-month's duration. So, I belong to the treated group. To measure the impact, you have to pick someone else, who has not attended the one-month coaching class in that center as counter-factual or control. If you pick Saina Nehwal to play against me as control, does it make sense? Based on this, can you say that the coaching class is a total waste because the training did not help me to score a single point? Ideally speaking, if you like to measure the impact of the training, the opponent has to be a clone copy of me, minus the training!

You might wonder if there are cases where there can be no sample selection bias. Yes, if the beneficiaries for a programme is selected 'at random' then, by definition, the beneficiary group and non-beneficiary group would be similar on average (please note that we are using the term 'similar', not 'same'). Also, the two groups will be similar on average, you can't expect each person in the beneficiary group to be similar to each person in the non-beneficiary group (on different variables). In this case, simple difference of means across treatment and control can be treated as impact and hence we say 'randomness is economists' best friend'. Unfortunately, in most programmes, the selection of beneficiaries is based on some observed characters and not a random assignment (except for one or two programmes in the world, like conditional cash transfer scheme for improving school enrolment in Mexico known as 'Progressa'). It is because few programmes are targeted for a specific target group (like people below poverty line or small farmers) where random assignment is impossible by design. In the case of other programmes, random allotment is simply not practical on a large scale due to socio-political factors.

Isn't sample selection bias due to sampling error? Definitely not. Let me try to convince you that 'random sampling' cannot cure selection bias. Selection bias arises due to pre-treatment differences in beneficiary and non-beneficiary groups with respect to some variables, say education and land holding size, for our convenience. Let us assume that the beneficiaries of the programme are mostly large farmers and well-educated farmers. So, when you take random sample, it is quite obvious that most of the beneficiaries are large farmers and well-educated and vice versa with non-beneficiaries, so random sampling cannot eliminate selection bias. The point we want to make is that random sampling is not the solution for selection bias. However, we acknowledge the importance of random sampling in social science research.

If selection bias is the problem, then why not take the value of outcome variable before implementing the programme as a baseline, and take a second measurement after implementation? Will the difference between the value of outcome variable after the programme and before programme become a measure of impact? Sadly 'no'. The outcome variable is measured at two different periods of time and in between many things might have changed. We cannot attribute the effect only to the programme and causation cannot be established.

The next common misconception is regression of outcome variable against a dummy variable indicating that treatment and all other control variables will be sufficient to account for selection bias and partial regression coefficient of the dummy variable as an unbiased estimate of impact. However, in this scenario, the dummy variable for treatment is not exogenous (as the selection into either treated or control group depends partly on the observed control variables included in the model), which is a violation of ordinary least square (OLS) assumption. Also, if the selection of treatment and control depends on unobservable (like motivation), then the error term will be correlated with dummy variable which is again a violation of OLS assumption. In this case, the estimate of impact will be biased.

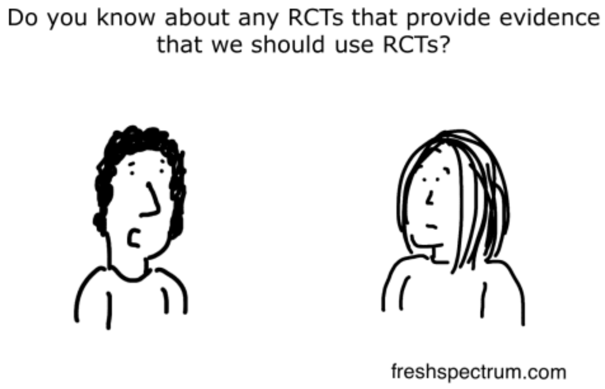
By this time, we have made our point clear that selection bias is inherent in observational studies and estimation without accounting for selection bias, which tends to be biased (over/underestimation). Next important question is what should/can we do to account for sample selection to ***minimize*** the bias? We would like to make one thing clear: if anyone tells you that some method will eliminate bias don't trust them. Because, no method can completely eliminate bias, each of the methodologies that we discuss here have their own advantages and disadvantages. The purpose is to minimize the bias in estimates and make it as accurate as possible. Moreover, there is no statistical test to tell us the best method for a particular data set, unlike the Huassman test for selection between fixed effect model and random effect model in panel data regression (econometric juggleries☺). So, the selection of method is left to the discretion of the researcher, who has to take a call based on the research question, size of sample, type of data and other factors.

There are different approaches available in literature which could help us in doing impact assessment. We will now briefly discuss these approaches.

IMPACT ASSESSMENT METHODS

Randomized Control Trials: The gold standard of impact assessment

We hope you are clear by now that observation studies suffer from selection bias because of 'non-random assignment'. What if can assign the units into either of the groups randomly? Or in other words, if the researcher has control over the treatment assignment, he could conduct an experiment where the treatment allocation is done randomly such that participation in the programme is independent of either observed or unobserved covariates. By definition, random allotment would mean that the treated unit and control unit are similar to each other on average and are comparable (we need to perform balancing test after randomization to make sure of this). Simple difference in mean outcomes across the group will be an estimate of impact. This looks simple on paper, however, it is difficult to implement in the field. This approach can be used only when the treatment allocation is under the control of the researcher. RCT needs to be planned before a programme/ intervention is implemented. Furthermore, in cases where there is possibility of spillovers, villages or clusters may need to be randomized. Even after taking care of all these things, the RCT method is criticized for lacking External Validity.



Source: https://designmonitoringevaluation.blogspot.com/2010/05/quotes-related-to-evaluation.html

This is an ideal approach for impact assessment, regretfully, most researchers won't get the luxury of doing it. Most of the impact assessments are ex-post observational studies and for such cases quasi-experimental approaches are available. A few commonly used approaches are discussed below.

Heckman two step model for impact assessment

In this approach, in the first step, a selection equation is estimated to capture the probability that an individual belonging to a treatment group, is dependent on a set of observed explanatory variables. This is usually estimated using Probit regression. From this regression, we estimate expected value of a truncated normal random variable, commonly known in literature as Inverse Mills ratio (IMS) or Hazard function [technically speaking Inverse Mills ratio tell us the probability that an individual will be in a treated group (or beneficiary group) over cumulative probability of the decision. Which explain that part of the error term which captures the difference in outcome variables due to the selection and not the programme itself. Is it too much jugglery? Just ignore ☺]. In the second stage, outcome variable is regressed upon dummy variable for treatment, along with a set of control variables, including IMS as an explanatory variable to minimize the effect of endogeneity (In simple terms, endogeneity in this context implies that the participation in a programme is determined by a set of observed and unobserved variables and is not exogenous). However, the Heckman model is developed for improving the explanatory power of the model in special case where sample self-selection leads to truncated dependent variable and OLS estimates are biased. So, the Heckman model is not specifically developed to establish the causal relationship. Hence, whenever possible, it is better to use models which are developed specifically for establishing causation. If the choice of method is limited by the smallness of a sample, it is better to use Heckman model in addition to other simple methods, such as Regression Adjustment as robustness check.

Regression Adjustment

Another very simple method (at the cost of efficiency though) for measuring impact is Regression Adjustment. We will try to explain the method in the simplest terms (though at the cost of technical fineness). The Regression Adjustment model fits two separate regressions – for the treated and control units – and estimate the partial regression coefficients for all the control variables included in the model (dependent variable - outcome variable like income). In the next step, the model estimates ‘Potential Mean Outcomes’ (PMO). PMO is the average value of the outcome if all the units in the sample are either in treated or control. (For example, what would be the mean income in case all the units in our sample were to be a beneficiary of the programme?) The Regression Adjustment model first calculates the **expected value of dependent variable for the entire sample** based on **coefficients of regression estimated on treated units.** Mean of the expected value is termed as PMO of treated group. Similarly, **expected value of dependent variable for the entire sample** based on **coefficients of regression estimated on control units** is used to estimate PMO for control units. The difference between PMO of treated and control groups is considered as estimate of impact. Again, a word of caution, the Regression Adjustment method is very sensitive to functional form of the outcome equation and model specification. In many cases, the estimate of impact changes drastically with addition/deletion of a control variable indicating model dependency leading to bias. In spite of these limitations, RA can be used as a method to assess impact, particularly when the size of sample is not large enough for semi-parametric matching methods, such as Propensity Score Matching.

Propensity Score Matching

Another very popular and widely used (or should we say abused?) method for assessing impact is Propensity Score Matching (PSM). Earlier, we had explained that the problem in observational studies is that we don’t have a proper counterfactual for assessing impact because of non-random assignment of treatment. However, what if we can select units from the control group, which are similar to the treated units and construct a quasi-counterfactual group? PSM, and also many related matching methods, use the same logic for impact assessment. The objective is to find the counterfactual for each treated unit from the control group we have. Suppose, in a treated group we have a farmer with 10 acres of land, 15 years of experience, who belongs to OBC group, and similar data on many other variables. The matching methods try to identify one (or more depending on type of matching we use) farmer from the control group who is very similar to the treated unit with respect to all these characteristic features. If we want to match with respect to one character, it is fairly straightforward, however, as the number of control variables increase, matching becomes increasingly difficult. We call it 'Curse of Dimensionality'. So, Rosenbaum and Rubin (1983) came up with a solution that if we can calculate 'propensity score' for each unit in the sample for each individual, which is a function of all the explanatory variables, then this propensity score can be used as a base for matching. This is as good as reduction of dimension, information on a set of control variables is captured in a single propensity score.

The propensity score is usually calculated based on logit or probit regression of treatment participation on a set of control variables. All those control variables which can impact either programme participation or the outcome should be included in the model. Once the propensity scores are calculated, the treated units are matched with the control units having similar propensity scores. The mean of difference in outcomes between treated and control units within each matched pair is considered as estimate of impact. (The basic logic is that the treated unit and the control unit in a matched pair are very similar to each other with respect to all covariates except for treatment. So observed difference in outcome is directly attributed to the treatment.) But before that we need to make sure that the propensity scores are good enough to achieve matching on the control units we have used for estimation. For this, the entire data is divided into different strata based on the value of propensity scores. Remember the basic assumption – the matching method will work if, and only if, observations having similar propensity scores also have similar values of control variables (on an average). This needs to be tested using a balancing test. Further, there could be a chance that many of the treated units have propensity scores for which no control variables are available for matching, which is termed as 'lack of common support'.

Of late PSM has received a fair share of criticism due to some serious drawbacks. We won't discuss all of these in detail, however, a few of them are worth mentioning here. PSM being a semi-parametric method, needs a bigger sample size to achieve proper matching and subsequent reduction in bias (so avoid using PSM for small samples). Secondly, PSM is centered on the assumption that selection bias arises due to observed variables (like age, education, caste, etc.) and not on unobserved characters (like motivation). So, if there is selection bias due to unobservable factors, PSM suffers from 'hidden bias'.

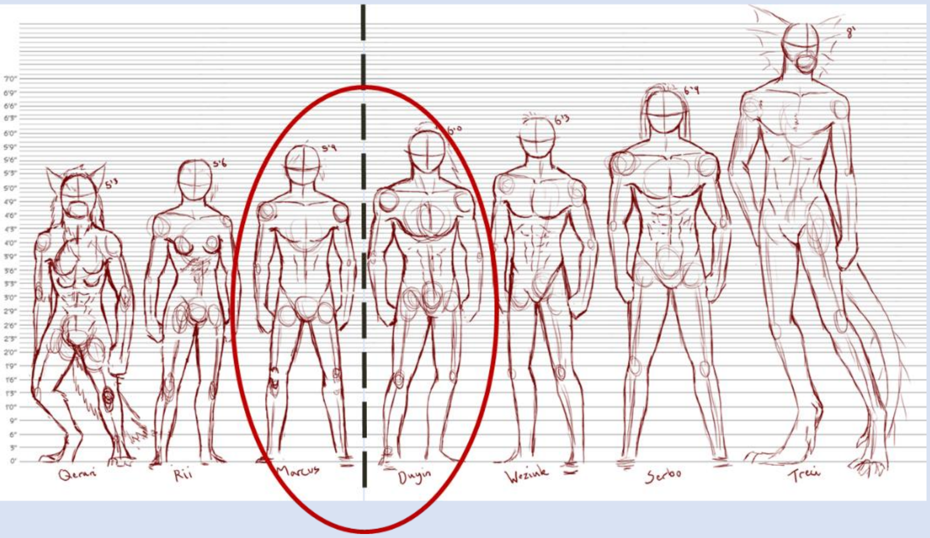
Box 3: Best practices in using PSM for measuring impact

- Use PSM only if you have a large sample size;
- Common support assumption is satisfied;
- All the relevant covariates/controls are used (only those variables which influence programme participation/value of outcome variable must be used in the analysis);
- Ensure that balancing property is satisfied;
- Try different methods of matching (Nearest neighbor, Caliper, etc.);
- Do sensitivity analyses for the estimates.

Regression Discontinuity Design

Regression Discontinuity Design (RDD) is a quasi-experimental approach of impact assessment just like PSM. RDD suits a situation where the probability of assignment in a treatment changes discontinuously with some forcing variable (Continuous variable). For instance, a particular programme is designed exclusively for small farmers. Here land cultivated by a farmer is the forcing variable and a farmer who has less than 2 hectares of land is automatically enrolled in the programme as beneficiary. So, ‘land cultivated’ becomes the forcing variable and 2 ha becomes the cutoff point. The key assumption is that the discontinuity design creates a randomized experiment around the cut-off value of the forcing variable. (In simpler terms, a farmer who has 2.1 ha of land, who is not a beneficiary of the programme, is a good counterfactual for a farmer who has 1.9 ha of land who is a beneficiary of the programme). In other words, the units which are near to either side of the cutoff (to both left and right of cutoff point) are good counterfactuals for measuring impact and only those units will be used to measure impact (see Box 4 for illustration). Well, how far must the units go to be included in the model? To decide this, optimum band width is selected (distance of units from the cutoff to be included in the model) based on selected optimization criteria. Then amongst the selected units, find the good counterfactuals, average treatment effect is usually estimated by fitting two separate regression functions (one to left of cutoff and one to the right). The biggest advantage of RDD is that the jump in regression line at the point of cutoff can be easily visualized. However, this method is criticized for using only the observations that are close to cutoff, leading to loss of information. It is also important to note that the method needs a large sample size.

Box 4: RDD illustration



Let’s consider an example were a programme was designed to give benefits to people who are more than 6 feet in height. Comparing the person who is on the extreme right (who is a beneficiary) with a person on the extreme left (who is in the control group) is not the best thing to do as they are not good counterfactuals. As discussed earlier, the RDD approach could be used to compare the impact of the programme by comparing the beneficiaries who are just above and below the programme cutoff (6 feet height). The optimum bandwidth would result in comparing people who are within a specific height parameter (circled).

Source: Modified by the authors based on work by Waspino (https://www.deviantart.com/waspino/art/Height-Anatomy-chart-WIP-285426107).

Difference in Difference method

Another popular method of impact assessment is Difference in Difference estimator. If you remember, we had said earlier that before and after comparisons suffer due to the ‘trend effect’ – many things (variables from methodology perspective) change over a period of time and observed change in impact cannot be attributed to programme participation. However, DID assumes that average change in income (or any outcome variable) due to 'other factors' or 'trend' would be the same across treated and control group. In other words, average difference in income before and after the implementation of the programme in the control group is due to 'trend effect' as they have not received the treatment. So, by subtracting the average difference in income before and after the implementation of the programme in the treated group from that of the control group (which capture trend effect), can we get the estimate of impact? As you might have noted already, DID rests on one very crucial assumption: DID assumes that average change in income (or any outcome variable) due to 'other factors' or 'trend' would be the same across treated and control group. (On a technical note, this means 'the variables that affect the value of outcome other than treatment are either time invariant or the time varying variables are group invariant. Confusing isn't it?) This holds true, if and only if, the income in treated and control group moves parallelly in the pre-treatment period. This we call as 'parallel trend assumption', which needs to be tested before using DID. If the parallel trend assumption is violated, we may have to use matching methods before using DID. Also, if there is spillover effect of a treatment, then the DID estimates may be biased.

We are providing one selected paper for each of the approaches discussed below for you to read and understand (Box 5).

Box 5: Articles using different impact assessment approaches

RCT

Emerick K, de Janvry A, Sadoulet E, and Dar MH. 2016. Technological innovations, downside risk, and the modernization of agriculture. American Economic Review 106(6):1537-1561.

Heckman two step model for impact assessment

Aditya KS, Subash SP, Praveen KV, Nithyashree ML, Bhuvana N and Sharma A. 2017. Awareness about Minimum Support Price and its impact on diversification decision of farmers in India. Asia and The Pacific Policy Studies 4(3):514–526.

Regression Adjustment

Agula C, Akudugu MA, Mabe FN and Dittoh S. 2018. Promoting ecosystem-friendly irrigation farm management practices for sustainable livelihoods in Africa: the Ghanaian experience. Agricultural and Food Economics 6:13.

Propensity Score matching

Aditya KS, Khan M.T and Kishore A. 2018. Adoption of crop insurance and impact: Insights from India. Agricultural Economics Research Review 31(2):163-174.

Regression Discontinuity Design

Sekhri S. 2014. Wells, water, and welfare: The impact of access to groundwater on rural poverty and conflict. American Economic Journal: Applied Economics 6(3):76–102.

Difference in Difference method

Khan MT, Kishore A, Pandey D and Joshi PK. 2016. Using zero tillage to ameliorate yield losses from weather shocks. IFPRI Discussion Paper 01562. Washington D.C.: International Food Policy Research Institute.

Apart from these methods, there are many other methods/approaches for impact assessment. Some of the others worth noting are: Inverse probability weighting, Inverse probability weighting regression adjustment, Endogeneity Switching Regression, PSM-DID, Coarsened Exact Matching (CEM), Instrumental Variable technique and synthetic control.

IMPACT PATHWAYS AND THEORY OF CHANGE

Though the objective of the blog is to highlight the need for, and methods of, impact assessment, it would be incomplete if we forget to mention the concept of impact pathway. The first step of any impact assessment exercise has to be development of impact pathways. Impact pathways are developed based on theoretical expectations regarding the expected outcome of a project and various pathways through which the impact is manifested. Impact pathways are developed based on 'theory of change'- the process through which the changes occur, leading to long term desired changes. Let us take a simple example of women's participation in SHGs. Participation in SHG activities, such as training and sharing information among participants, helps in increasing knowledge and skill sets. This might lead to a few women going on to explore entrepreneurial options like vegetable cultivation or kitchen gardens, leading to higher incomes. In turn, higher income can empower these women. Empowerment of women can then be linked to better nutritional and health outcomes. Such an impact pathway acts as a guide for conducting impact assessment. Assessing impact without impact pathways is akin to what George Fuechsel says, "Garbage in, garbage out!"

TO CONCLUDE

As discussed earlier, through this blog we intended to orient readers on the need for impact assessment and to introduce a few methods of impact assessment. The discussion on each method is driven by the principle of parsimony: the simpler is better than the better. However, we acknowledge that, in our endeavor to simplify, we might have missed out on a few technical things. The blog is only for understanding the basic principles behind each of the methods rather than to acquaint readers with the details on how to use it. As you are aware, by this time, there are many methods to measure impact. Each has its own premise, set of assumptions, advantages and disadvantages. Researchers must cautiously choose the right method after in-depth understanding of the research question, method and data availability. A detailed review of each method should be done so as to understand the 'good practices' and 'robustness checks' that each method demands. Moreover, reading recent literature is always advisable as Impact Assessment is an ever-evolving field where new methodologies/modifications and post-estimation tests for existing methodologies are developing at a rapid pace.

References

Rosenbaum PR and Rubin DB. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70(1):41-55.

Thaler Richard H and Sunstein CR. 2009. Nudge: Improving decisions about health, wealth, and happiness.

Aditya KS, Scientist, Division of Agricultural Economics, ICAR- Indian Agricultural Research Institute, New Delhi. Email id: adityaag68@gmail.com

Subash SP, Scientist, ICAR-National Institute of Agricultural Economics and Policy Research, New Delhi. Email id: subashspar@gmail.com