Potential Solutions to the Fundamental Problem of Causal Inference: An Overview

By Chahir Zaki

Training on Applied Micro-Econometrics and Public Policy Evaluation

Economic Research Forum

The fundamental problem

- Impact evaluation in general faces the challenge of trying to rigorously separate out the amount of difference caused by the program being evaluated from differences caused by other factors.
- In program evaluation, want to know the impact of the program ("treatment") on participant outcomes.
- In the real world, participation in programs and the impact of public policies is difficult to identify.
  - Participation is likely to be related to characteristics that also affect outcomes
- Endogeneity: assignment to treatment is not random.
  - Not only depends on observables, but may also depend on unobservables.
  - Both observables and unobservables may affect the outcome.
- For instance, it is not possible to compare Takaful and Karama beneficiaries to non-beneficiaries to measure this impact, as the beneficiaries are generally poorer than non-beneficiaries.

Solutions

- Random assignment

- Quasi-experimental solutions
  - Type I: Conditional exogeneity of placement
    - Difference-in-difference
    - Panel data (fixed and random effects)
    - Propensity score matching

  - Type II: Rules or instruments of placement
    - Control function and instrumental variables techniques
    - Regression discontinuity design

Random Experiments

Random experiments

- Random experiments are often referred to as Randomized Controlled Trials (RCTs)
- Random allocation of intervention to program beneficiaries such that all units (within a defined set) have equal chance *ex ante* of receiving the treatment
- Assignment process creates treatment and control groups that are directly comparable
  - Should not have any observable or unobservable differences
- By eliminating selection bias, randomization allows direct comparison of participants and non-participants to detect impact of program
- Observed *ex post* differences in mean outcomes between treatment and control group can be attributed to program

Problems with Experimental Designs

- Ethical and political obstacles
- Difficult to randomize at level of individual beneficiaries
- Those assigned to treatment group may decline to participate or participate in a partial manner
  - This is referred to as selective compliance
  - Selection bias gets introduced through this self-selection process
- Those not selected and assigned to the control group may try to find alternative ways to get benefit of program
  - "Contamination" of control group

Case Study 1:
The Labor Market Impact of Youth Training in the Dominican Republic: Evidence from a Randomized Evaluation

Card et al. (2007)

Case study of a randomized evaluation

- From 2001 to 2005 the government of the Dominican Republic operated a subsidized training program, *Juventud y Empleo* (JE)
  - Targeted low-income youth (18-29) with less than a secondary education in urban areas
  - Several weeks of classroom training (basic skills & vocational skills) by private training institutions
  - Followed by an internship at a private sector firm

- Program was evaluated in:
  - Card, David, Pablo Ibarraran, Ferdinando Regalia, David Rosas, Yuri Soares (2007). "The Labor Market Impact of Youth Training in the Dominican Republic: Evidence from a Randomized Evaluation." National Bureau of Economic Research Working Paper 12883.

Structure of the evaluation

- JE program was unique in incorporating a randomized design
  - Each time 30 eligible applicants were recruited, 20 of the 30 were assigned to training (treatment), 10 to control.
  - Up to 5 individuals from control could be re-assigned to treatment if those assigned to treatment failed to show up for training (no-shows) or dropped out in the first two weeks (dropouts)

- Evaluation looks at the second cohort of the JE program
  - Trained in early 2004
  - Baseline data from registration form (prior to randomization)
  - Follow up survey in summer 2005 (~1 year after training)

Sample of the evaluation

- Second cohort consisted of 8,391 eligible applicants
  - 5,802 (69.1%) assigned to treatment
    - 1,011 dropouts or no shows
  - 2,589 (30.9%) controls
    - 966 reassigned

- Led to realized treatment group of 5,757 and realized control group of 1,623
  - Only these groups have follow-up data

- Problem of missing post-program data on no-shows and dropouts
  - Will bias results if this group is non-random.

Outcomes

- Labor market outcomes examined:
  - Employment
  - Hours of work
  - Hourly wages
  - Job with health insurance

Table 5: Summary of Labor Market Outcomes in the Follow-up Survey

Outcome:	Treatments	Controls	Raw Difference	Re-weighted Difference
Employment Rate	57.38%	55.95%	1.43%	0.02%
	1.77%	*2.09%*	*2.74%*	*2.74%*
Monthly Income (All Jobs)	$ 5,818	$ 5,289	$ 529	$ 438
	$ 195	*$ 202*	*$ 288*	*$ 284*
Hours worked per week (All Jobs)	43.43	44.27	-0.84	-1.11
	0.79	*0.98*	*1.25*	*1.27*
Hourly Wage (All Jobs)	$ 151.19	$ 133.92	$ 17.27	$ 14.50
	$ 9.91	*$ 7.02*	*$ 13.32*	*$ 11.84*
Health Insurance in Primary Job	38.0%	34.8%	3.1%	2.5%
	2.5%	*2.9%*	*3.9%*	*3.9%*

Notes: standard errors in italics. See note to table 2. The sample for employment includes everyone. The sample for income, hours per week, hourly wage, and health insurance includes those with positive earnings and between 10 and 85 hours per week. The value of earnings is censored at the 99th percentile.

- No impact on employment, hours of work, some differences in monthly earnings
- Some marginally significant impact on hourly wages of about 10%
- No significant differences in health insurance

Lessons from Card et al. 2007

- Randomization is the "gold standard" but reality of randomization usually less than perfect
  - Imperfect compliance
  - "Contamination" (reassignment) of controls
  - Potential selection bias due to no-shows and drop-outs
  - Potential dilution of impact for partial participation
- Still have to check assumptions and correct for selection in many randomized evaluations.

Case Study 2:
The Impact of Exporting: Evidence from a Randomized Experiment in Egypt

Atkin et al. (2017)

Design

- Small and medium enterprises (SMEs) are often important parts of developing country economies, but many programs designed to spur their growth have been unsuccessful.
- Researchers conducted a randomized evaluation in **Fowa**, a peri-urban Egyptian town with a population of 65,000.
- **Fowa** is known for its “carpet cluster” of hundreds of small textile firms. These firms usually consisted of a single owner who operates out of a rented space or his home and typically employed one to four individuals using hand looms.
- Researchers focused their evaluation on SMEs with fewer than five employees, the majority of which had not knowingly exported in the past.

Design

- To reach these SMEs, researchers partnered with Aid to Artisan (ATA), a U.S.-based non-profit organization that finds promising small-scale developing-country producers and fosters trade relationships between them and high-income OECD markets.
- At baseline, researchers surveyed 219 SMEs on their production techniques, product quality levels, owner characteristics, and household indicators.
- Of these 219 firms, researchers randomly selected 74 to serve as the treatment group. ATA offered these SMEs the opportunity to produce orders for delivery to U.S. and European retailers.
- The remaining 145 SMEs served as the comparison group and were not offered access to export opportunities.

Findings

- *Profits and productivity*:
  - Among SMEs offered the opportunity to export, operating profits increased 15 to 26 percent relative to comparison group firms.
  - The higher-quality rugs demanded by foreign retailers were more expensive and time-intensive to make, but the resulting increased purchase price was more than enough to offset costlier production.
  - This increase in profits, thus, came in tandem with increases in quality, as well as declines in output per hour.
- *Learning by exporting*:
  - Evidence from this evaluation suggests that the quality upgrades may have been a result of "learning by exporting." When the first orders arrived, productivity declined, but then rose steadily as production continued.
  - In the researchers' "quality lab," the rug quality produced by treatment group SMEs was higher across every dimension measured.
  - Communication logbooks confirmed that most improvements took place along dimensions discussed between buyers and sellers, suggesting that feedback was an important improvement channel.

Quasi-experiments

Causal Effects in Non-Experimental Evaluations

- We want to identify the causal impact of a program or policy
  - Typically we do not have experimental data (undertaking a non-experimental evaluation)
  - Referred to as **quasi-experiments**

- To estimate a causal effect in non-experimental evaluations we need "identifying assumptions"

- Non-experimental methods can be classified into two types depending on the identification assumptions they make.
  - Type I: "conditional exogeneity of placement" or "conditional exogeneity of placement to changes in outcomes"
  - Type II: instrumental variables or discontinuities that can explain placement can be found

Non-Experimental Methods

- Type I Non-Experimental Methods

1- Regression Methods

2- Propensity Score Methods

3- Difference in Difference Methods

4- Panel data (fixed or random effects) models

- Type II Non-Experimental Methods

4- Instrumental Variable Methods

5- Regression Discontinuity Design Methods (RD or RDD)

Causal Inference in Type I Non-Experimental Methods

- Type I non-experimental methods make the following identification assumptions:
  - Conditional exogeneity of placement
    - Placement only depends on exogenous observable characteristics X and not on unobservables.
    - Often referred to as "selection on observables"

OR

  - Exogeneity of placement with respect to changes in outcomes:
    - Unobservable factors affecting changes in outcomes do not affect the probability of placement.
    - Unobservables that determine placement can affect initial conditions but are assumed not to affect changes in outcomes over time

Type I Methods: First and second differences

- Under "conditional exogeneity of placement", all we need to do is compare outcomes for a treatment and control group at one point in time controlling for observables X
  - This is called a first difference approach (see D(X) estimator)

- Under the weaker "exogeneity of placement to changes in outcomes" we need to compare the difference from before and after the program for a treatment group to the same difference for a control group.
  - This is called difference-in-difference or a second difference approach

Type I Methods: Propensity Score Matching & Weighting

- Assumes conditional exogeneity of placement (selection on observables)

- Models the selection process with a probit or logit model to predict the probability of participation, Pr(T=1) based on observable characteristics (X)

- Creates “matched” treatment and control groups
  - After matching or weighting, no observable differences between groups

- Can then estimate program impacts by looking at mean differences (ATT) between matched/weighted T and C groups

Type I Methods: Random and Fixed Effects

- Often concerned about unobservables that are going to be related to an observable unit (school, family, city)

- Panel data models assume that after controlling for the effect of that unit, the remainder of selection is fully observable

- Random effects (RE) models assume the unobservable effects have some underlying (normal) distribution
  - REs assumed to be unrelated to observable X

- Fixed effects (FE) models do not require parametric assumptions
  - FEs can be related to observable X

Causal Inference in Type II methods

- Identifying assumptions:
  - There exists at least one (instrumental) variable (IV) that affects participation (placement) but that does not affect the outcome conditional on participation and other covariates (X))
  - i.e. that the IV can be excluded from the outcome regression without causing omitted variable bias. This is called an “identifying restriction”
  - To be valid this IV must be exogenous
  - This called the instrumental variables approach

- Regression discontinuity design (RD or RDD) is based on a similar assumption.
  - The instrument is some cutoff for eligibility/participation in the program
  - RD focuses on the differences in outcomes around that cutoff to model program impacts

Case Study 3: Takaful and Karama

IFPRI (2018)

How to study the impact of Takaful and Karama?

- Ideally, we would randomly assign some households during a pilot period to receive the program and a similar set of households not to receive the program (Randomized Control Trial). The households which do not receive the program represent our control group to which the treated households can be compared, and since the two groups were initially similar, any difference will be caused only by the program.

- Another common approach is to survey households before starting the program and then compare beneficiary outcomes before the intervention to beneficiary outcomes after the intervention (Differences-in-Differences).

Because Takaful and Karama had already started by the time we designed the impact evaluation strategy, neither of these first two approaches was feasible.

How to study the impact of Takaful and Karama?

- A third approach is to try to find a group of non-beneficiary households that are similar to beneficiary households as far as a set of observable characteristics (Matching).
- Matching techniques work best when we are confident that households in the beneficiary and non-beneficiary groups are also similar with regards to non-observable characteristics.
  - In Takaful and Karama, we expect that households that registered for the program are likely to be quite different from households that did not register in both observable and non-observable ways.
  - For example, between two women who look similar on paper, the fact that one woman went to the effort to get the documents together and apply for the program means that she may be more organized and have more initiative than the woman who did not apply, so we do not know if her family is better-off because of the program or because she does a better job in general at managing her household.
- This concern about unobservable differences led the researchers to concentrate on comparing non-beneficiary households to beneficiary households **only among households that registered** for Takaful and Karama.

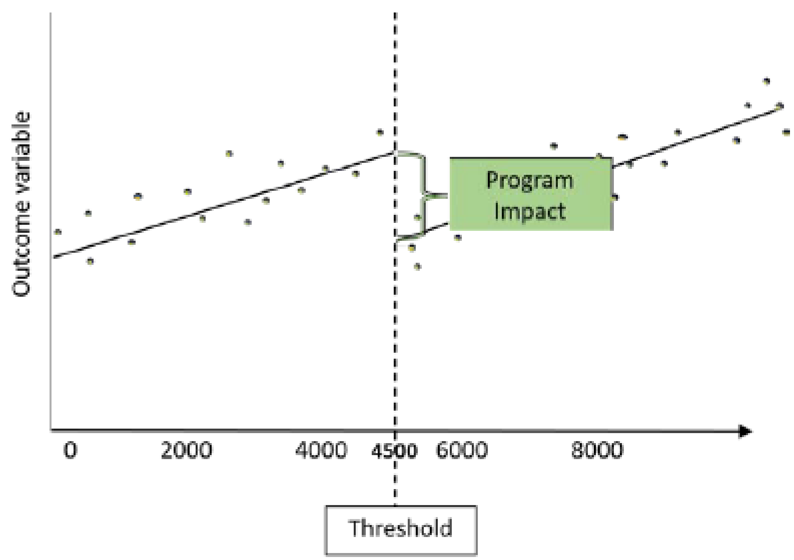
How to study the impact of Takaful and Karama?

- Having decided to concentrate on registrants, it was noticed that a very large number of observable variables are already used in the proxy mean test (PMT) score and that the PMT score is the primary factor determining if a household is in the program.
- This implies that it is not feasible to find a set of non-beneficiary households that registered for the program and are similar to beneficiary households in terms of having identical observable characteristics because if they were similar enough to use for matching, they would have PMT scores similar to beneficiary households and they would be in the program.

How to study the impact of Takaful and Karama?

- Thus, RDD must be used: while it is hard to find non-beneficiary registrant households that are similar to all beneficiary households, we can compare households just below and just above the cutoff score.
- While this strategy is very effective at determining the true impact of the program as distinct from the influence of any other factors, and in the case of this impact evaluation was the only option, the disadvantage is that it estimates that impact only among households near the cutoff.

RDD design



Outcome variable
Program
Impact
0
2000
4000
4500
6000
8000
Threshold

Example: Takaful and Karam

- The impact evaluation was conducted using a regression discontinuity (RD) design.
- The program was targeted by selecting households who fell below a threshold level on a proxy means test (PMT) score using data collected during three waves of registration.
- The RD approach compares outcomes for beneficiaries below each threshold for eligibility to outcomes for non-beneficiaries above the threshold.
- The available impact evaluation data were well suited to conducting the analysis using the RD approach: there is a large number of households with a PMT score near the eligibility thresholds (except for the first threshold) and the PMT score is continuous at the eligibility threshold.

Example: Takaful and Karam

- To assess the impact of the Takaful and Karama program, a household survey was conducted from July 15 – August 30, 2017.
- The survey collected information on outcomes related to household expenditure and poverty, well-being and income, schooling, child dietary diversity and anthropometry, child morbidity, household dietary diversity, health care utilization, infant and young child nutrition knowledge and practices, women's decision making, shocks, and illness and disability.
- The sample for the evaluation includes 6,541 households in the impact evaluation sample plus an additional 1,692 households in a nationally representative sample for targeting analysis.

References

- Atkin, D., Khandelwal, A. K., & Osman, A. (2014). *Exporting and firm performance: Evidence from a randomized trial* (No. w20690). National Bureau of Economic Research.
- Breisinger, C., ElDidi, H., El-Enbaby, H., Gilligan, D., Karachiwalla, N., Kassim, Y., ... & Thai, G. (2018). *Egypt's Takaful and Karama cash transfer program: Evaluation of program impacts and recommendations*. Intl Food Policy Res Inst.
- Card, D., Ibarrarán, P., Regalia, F., Rosas-Shady, D., & Soares, Y. (2011). The labor market impacts of youth training in the Dominican Republic. *Journal of Labor Economics*, *29*(2), 267-300.
- Slides of Ragui Assaad and Caroline Kraft.

Thank you for your attention