The role of evaluation in social research: current perspectives and new developments

Lorraine Dearden, Institute of Education & Institute for Fiscal Studies

Email: ldearden@ifs.org.ukuk

Presentation for the Social Research Association (SRA) Annual Conference, British Library, 6th December 2012

Introduction

- Understanding what works and what doesn't work is crucial in social research
- To do this effectively, you need good qualitative work coupled with robust impact evaluation and an assessment of both costs and impact.
- Why?
  - Inappropriate quantitative methods can often find a correlation between a policy and an outcome that is not *causal* which is of no use for social research or policy making (Patrick's talk)
  - Finding a *causal* ex post quantitative impact on an outcome of interest is generally only part of the story and will often not tell you *why* a policy is having an impact – need qualitative work to help this this (William's talk) or more sophisticated evaluation models (dynamic structural models)
  - Just because something has a quantitative impact doesn't mean it is a good policy if the costs are large - need some assessment of both benefits *and* costs

The Evaluation Problem

- Most empirical questions in social research can be set up in an evaluation framework
- What most empirical social researchers want to ask is:
  - what is the *causal* impact/effect of some program/variable of interest on an outcome of interest?
- Good quantitative evaluation methods try to utilise methods that can estimate this *causal* impact in a robust way
- However, there is no off the shelf evaluation technique that can be used in all circumstances which is a mistake often made in social research
- Best method depends on nature of the intervention; the way it was introduced; the richness of the data; whether outcomes are also measured before the intervention…

The missing counterfactual and selection

- Question we want to answer:
  - What is the effect of some program/treatment on some outcome of interest *compared to the outcome if the program/treatment had not taken place*
- Problem is that we never observe this missing *counterfactual*
- Fine if program/treatment is randomly assigned, but in most social research settings this is not the case
- Generally have to construct a counterfactual group from those who don't get treatment
- But these two groups are generally systematically different from each other in both observed and unobserved characteristics which means they are often not a good counterfactual group – *selection problem*

So how do we get around this selection problem?

- Non-experimental evaluation techniques use a variety of statistical methods to identify the causal impact of a treatment on an outcome of interest
- Generally rely on having good quality data; and/or a natural or social experiment (policy accident/pilot study)
- Methods differ in the assumptions they make in order to recover the missing counterfactual but try to replicate a randomised control trial
- Take you through a brief tour of some of these:
  - Matching methods
  - Regression Discontinuity Design (Patrick already discussed)
  - Instrumental and Control Function methods (won't discuss)
  - Difference- in-Difference (DID) methods
  - Dynamic structural models

Matching Methods

- Need to have a well defined treatment and control group
- Relies on having a rich set of pre-program/treatment variables for those who get treatment and those who don't – *matching variables*
- The matching variables need to be good predictors of whether you get treatment or not and/or the outcome of interest
- They need to be measured before the treatment – you cannot match on any variable which has the potential to get affected by the program/treatment
- Crucial Assumption: assume ALL relevant differences between the groups pre-treatment can be captured by the matching variables
  - Conditional Independence Assumption (CIA)

How do you match? Regression Models

- Standard regression models are matching models but have quite strong assumptions
- Simply regress outcome of interest on matching variables and treatment variable (dummy variable of whether or not you receive program/treatment)
  - Coefficient on treatment dummy variable gives you effect
- Some Key Assumptions:
  - that there is only selection on the basis of the matching variables
  - That a linear model can accurately specify the relationship between the matching variables and the outcome of interest
  - Effect of the matching variables on outcome of interest doesn't change as a result of the intervention
  - Can relax (test) this last assumption using a regression framework by interacting matching variables with treatment variable

Propensity Score Matching (PSM)

- More flexible matching method but more computationally difficult
- Involves selecting from the non-treated pool a control group in which the distribution of observed/matching variables is as similar as possible to the distribution in the treated group
  - This is done by deriving weights which make the control group look like treatment group in terms of matching variables
- There are a number of ways of doing this but they almost always involve calculating the propensity score
- The propensity score is the predicted probability of being in the treatment group, given your matching characteristics
  - Can do this using traditional regression techniques and significant variables in this estimation procedure will pick up matching variables that systematically differ between two groups
- Rather than matching on the basis of all matching variables can match on basis of this propensity score (Rosenbaum and Rubin (1983))

How do we match using propensity score?

- Nearest neighbour matching
  - each person in the treatment group choose the individual in the control group with the closest propensity score to them
  - can do this with (most common) or without replacement
  - not very efficient as discarding a lot of information about the control group (throw away all people not matched and may use some individuals a lot of times)
- Kernel based matching
  - each person in the treatment group is matched to a weighted sum of individuals (adding to one) who have similar propensity scores with greatest weight being given to people with closer scores
  - Some methods use ALL people in non-treated group (e.g. Gaussian kernel) whereas others only use people within a certain range (e.g. Epanechnikov)

Estimated impact with PSM

- Compare mean outcome in treated group to the appropriately *weighted* mean outcome in the control group (using propensity score weights)
- So just comparing two (weighted) means
- No guarantee that you can come up with matching weights that make the two groups look the same in terms of matching variables
  - Quite common if two groups are fundamentally different
  - Can drop those for whom you can't find matches (imposing common support) but then effect is measured only on a sub-sample
- Don't have to specify how the matching variables affect outcome so much more flexible and robust than regression methods but much less efficient

Difference-in-difference methods

- DID approach uses a natural experiment to mimic the randomisation of a social experiment
- Natural experiment – some naturally occurring event which creates a policy shift for one group and not another
  - E.g. It may be a change in policy in one jurisdiction but not another
- The difference in outcomes between the two groups *before* and *after* the policy change gives the estimate of the policy impact
- Requires either longitudinal data on same person/firm/area or repeated cross section data on similar persons/firms/areas (where samples are drawn from the same population) before and after the intervention
- Assumes that change that occurs to control group would have happened to treatment group in absence of policy change so any *additional* change is the impact of the policy (ATT)
- Can do *matched* DID (DID using propensity score weights)

Dynamic Structural Models

- Evaluation techniques described so far are for analysing impact of changes we have observed (ex post)
  - Results are specific to the policy, time and environment
- How can we model the impact of future (ex ante) policy reforms?
- Build a dynamic structural model of impact based on theory which can disentangle impact of programme on incentives from how incentives affect individual decisions (cf traditional evaluation methods)
- Use existing quasi-experimental results/data to estimate and validate (calibrate) models
- When/If succeed in doing this use model to simulate policy impact of new policies.
- Very difficult and computationally complex so models to date tend to be very simple but increasing computer power means that this is a new and exciting area in evaluation and of extreme policy interest.

Conclusions

- Number of options available when evaluating whether something has impact or is likely to have impact in social research
- Depends on nature of intervention, available data, question you want to answer……
- Each methods has advantages and disadvantages and involves assumptions that may or may not be credible and all these factors have to be carefully assessed
- New PEPA Node based at IFS will be looking at improving/developing programme evaluation methods for policy analysis as well as running a comprehensive training and capacity building program.