Causality and plausibility in experimental and quasi-experimental evaluations

Conny Wunsch

WWZ

University of Basel

Vienna | 21 June 2019

Outline

Introduction

Evaluations aim at answering the following types of questions:

- What are the effects of a policy/measure/instrument/program?
- Is the policy/measure/instrument/program effective in achieving its objectives?
- Is the policy/measure/instrument/program cost-effective?
- Is the policy/measure/instrument/program optimal?

Example:

- Does labour market training reduce the unemployment duration of jobseekers?

Answering these questions necessarily requires isolating the ***causal effect*** of the policy/measure/instrument/program.

Introduction

Ex-ante evaluation predicts the consequences of a new policy:

- Since the policy is new, no relevant data are available.
- Specify microeconomic or macroeconomic model.
- Estimate or calibrate structural parameters of the model.
- Simulate the consequences of the policy.
- Can be used to find the optimal policy.

Ex-post evaluation studies the consequences of an implemented policy:

- Use actual data.
- Reduced-form methods (experimental or quasi-experimental) allow assessing the (cost) effectiveness of the policy.
- Theory-based methods possibly allow for normative statements on top.

Introduction

Causality versus correlation:

Consider a training program for unemployed workers.

- The average unemployment duration of non-participants is 7.2 months.
  The average unemployment duration of participants is 9 months.
- Participation in training correlates positively with unemployment duration.
- Participation in training is associated with longer unemployment duration.

Does that mean that the training fails to reduce unemployment duration and is even harmful for participants?

Introduction

Causality versus correlation:

Consider a training program for unemployed workers that assigns participants based on employability:

Training	Employability	Share	Unemployment duration	
No	High	80%	Low	6 months
No	Low	20%	High	12 months
Yes	High	20%	Low	5 months
Yes	Low	80%	High	10 months

The causal effect of participation in training on unemployment duration is negative and equal to -1 month for workers with high employability and -2 months for workers with low employability: participation in training reduces unemployment duration.

Employability is a confounder.

Failure to isolate the causal effect leads to wrong policy conclusions.

Introduction

Causality versus plausibility:

An evaluation is plausible if it is credible that it isolates and quantifies the causal effect of the policy/measure/instrument/program.

- plausibility = causality
- plausibility = credibility
- plausibility = reliability
- plausibility = internal validity

The first and most important objective of any evaluation study must be to ensure high internal validity.

Introduction

Internal validity:

- Plausibility or credibility of an evaluation study (causality).

External validity:

- Validity of applying the conclusions of an evaluation study outside the context of that study: extent to which the results of a study can be generalized to and across other situations, people, and times.

Introduction

Experimental methods create exogenous variation by randomization:

- high internal validity, possibly low external validity
- lab experiment
- field experiment
- lab-in-the-field experiment

Quasi-experimental methods exploit existing variation:

- high external validity
- internal validity depends on the validity of the identifying assumptions
- as-good-as-random assignment conditional on observed confounders
- exploit policy reform
- exploit policy discontinuity
- exploit instrumental variable

Outline

Methods to isolate causal effects

Randomization:

- random assignment of unemployed workers to training or no training
- confounders do not exist by construction
- requires compliance with treatment assignment

Construct (or find) instrumental variable (IV):

- randomize the offer of training in the group of eligible workers
- take-up is endogenous
- the offer of training serves as instrument for participation in training
- requires that only workers who received the offer participate and that receiving the offer does not discourage anyone from participating (monotonicity)

Methods to isolate causal effects

Randomization conditional on observed determinants of selection:

- all variables that jointly determine selection into training and unemployment duration are known and observed in the data:

  selection on observables, unconfoundedness, ignorability,
  conditional independence assumption (CIA)
- due to budget or capacity constraints, there are observationally identical unemployed workers who do and do not receive training (common support)
- need profound knowledge of the relevant institutions and the selection process

Methods to isolate causal effects

Time-constant unobserved confounders:

- exploit introduction or abolishment of training
- need data for more than one period: one without and one with training
- outcomes in period without training must be unaffected by availability of training in subsequent/previous period
- find or collect data on all time-varying confounders
- measure impact of time-constant (observed and unobserved) confounders
- first difference estimation
- individual fixed effects (FE)
- difference-in-differences (DiD) with an unaffected group
- need profound knowledge of the relevant institutions and the selection process

Methods to isolate causal effects

Exploit a policy discontinuity:

- assignment to training is based on a continuous measure of employability (e.g. predicted probablity to become long-term unemployed)
- jobseekers below a certain cutoff value of the employability measure do not receive training
- all or a certain share of jobseekers above the cutuff receive training
- implies that there is a discrete increase in the probability of receiving training at the employability cutoff: regression discontinuity design (RDD)
- requires no sorting around the cutoff: local IV

Pros and cons of the different methods

Randomization:

- gold standard
- non-compliance and external validity are an issue

Construct (or find) instrumental variable (IV):

- very hard to find valid instruments

Randomization conditional on observed determinants of selection (CIA):

- very hard to justify that there are no unobserved confounders
- much more credible when many pre-treatment outcomes are observed
- common support can be an issue

Pros and cons of the different methods

Time-constant unobserved confounders (FE/DiD):

- often considered more plausible because it allows for unobserved confounders
- often hard to justify that there are no time-varying unobserved confounders
- other changes over time are a problem
- functional form dependence is a problem

Exploit a policy discontinuity (RDD):

- if there is no sorting, credible source of locally exogenous variation
- large toolkit to assess plausibility of no sorting
- if there is sorting, bunching approaches may offer an alternative
- only local effect is identified: external validity is an issue

Identification versus estimation

Identification strategy:

- set of assumptions required to ensure causality

Estimation strategy:

- choice of estimator for a given set of identifying assumptions

Identification	Examples of estimators
IV	2SLS, LATE, MTE
CIA	OLS, IPW, Matching
DiD	OLS, IPW, Matching, FE
RDD	OLS/2SLS, Wald estimator, Local linear regression

Identification versus estimation

Non-parametric identification:

- minimum set of assumptions required for identification
- no functional form or distributional assumptions
- no parameter restrictions

Parametric identification:

- functional form or distributional assumptions or parameter restrictions necessary for identification in addition to minimum set of assumptions

Identification versus estimation

Non-parametric estimation:

- no assumptions added to the set of minimal assumptions

Semi-parametric estimation:

- no restrictions on the relationship between the outcome, the treatment, the observed and the unobserved variables
- typically restrictions on the relationship between the treatment, the observed and the unobserved variables

Parametric estimation:

- restrictions on the relationship between the outcome, the treatment, the observed and the unobserved variables

Identification versus estimation

Non-parametric estimation: exact matching

- minimal assumptions are conditional independence (CIA) and common support

Semi-parametric estimation: propensity score matching

- no restrictions on the relationship between the outcome, the treatment, the observed and the unobserved variables
- probit or logit model for the relationship between the treatment, the observed and the unobserved variables

Parametric estimation: OLS

- linear relationship between the outcome, the treatment, the observed and the unobserved variables
- effect homogeneity

Outline

Scientists as evaluators

Incentives for scientists to become evaluators:

- genuine interest in the evaluation
- new policy/measure/instrument/program
- access to unique data
- possibility to apply new methods
- fund raising

Constraints of scientists:

- need to or want to publish: innovation and high internal validity is a must
- personnel: labour intensive projects are a problem
- time: tight deadlines are a problem

Scientists as evaluators

Scientists are more likely to become evaluators if the evaluation itself or something that is possible with the same data has (high) publication potential, i.e. is innovative and promises high internal validity.

Evaluation of active labour market policies in Germany:

Nr.	Time	Contract research	Derivatives
1	2000-2005	JEEA, GER, ZAF	JOLE, ReStat
2	2003-2007	Kyklos, Economics of Transition	ReStat
3	2006-2008	GER	Health Economics
4	2007-2011	Labour Economics, GER	J'Emetrics, JHE, ILRR

Scientists have a strong incentive to deliver evaluations with high internal validity.

But are they better evaluators?

Scientists as evaluators

Are scientists better evaluators?

Not necessarily because

- practitioners often have better knowledge of the institutions and the aspects that are relevant in practice
- scientists have fewer incentives to study all possible effects and heterogeneities (papers need to be short)
- scientists are often slow and produce results when they are already outdated for policy makers or practitioners
- scientists often use methods that are very complex and difficult to understand for policy makers and practitioners

The practitioners' perspective

Practitioners:

- scientists who conduct contract research
- research institutes
- think tanks
- for-profit evaluators
- groups of evaluators within the institutions that implement the policies
- ...

All evaluations that are being conducted without the primary objective of publication in a peer-reviewed scientific journal.

The practitioners' perspective

Constraints of practitioners:

- on-time delivery within relatively short deadlines
- limited resources
- circumstances may limit internal validity

Which compromises are ok?

- For a given identification strategy (e.g. CIA), choose an estimator that is easy to implement and minimizes computing time (e.g. OLS or IPW instead of matching).
- Often not necessary to implement latest state-of-the-art technical refinements.

The practitioners' perspective

What if circumstances limit internal validity?

- Evaluators should always make the most of the knowledge and data they have.
- They should always maximize internal validity within the limitations they face.
- They should consider involving scientists as consultants to maximize the credibility of their studies.
- Evaluations should always be fully transparent about their limitations. Otherwise, consumers may draw false conclusions.
- Results from studies that to not manage to establish causality should not be used for policy advice.
- But these studies are still important to obtain first ideas and to understand what is required to be able to establish causality. This information should serve as input for the design of a follow-up evaluation study that is able to establish causality.

The practitioners' perspective

Ideally, given today's knowledge about credible evaluations, institutions that implement new measures and wish to have them evaluated at some point, should make sure from the onset that credible evaluations are possible:

- create some randomness
- document selection process
- conduct surveys on determinants of selection
- collect data on individual characteristics and outcomes
- ...

Outline

Take aways

- Failure to isolate causal effects leads to wrong policy conclusions.
- Plausibility = causality = credibility = reliability = internal validity.
- The first and most important objective of any evaluation study must be to ensure high internal validity.
- Scientists have different incentives than other evaluators.
- Scientists are not necessarily better evaluators.
- Some pragmatism is ok but not at the expense of internal validity.
- Non-causal evaluations can provide some insights but their results should not be used for policy advice.

Thank you very much for your attention!

University of Basel
Faculty of Business and Economics

WWZ

conny.wunsch@unibas.ch
https://wwz.unibas.ch/en/laboureconomics