The Resource Management Framework

Additional guidance on Evaluating Lapsing Programs - quantifying the effect of programs

Effective from 1 July 2019

Key points

- The **causal effect** of a program is the extent to which it has changed outcomes for participants, compared with what would have been the case without the program. The Neyman-Rubin Causal Model formalises this in a statistical framework.
- Accurately quantifying the effect of a program depends on deliberate decisions during program design. Good evaluation should be part of a program's design from the outset.
- A program evaluation report should explain and address potentially significant **confounding factors**—characteristics that affect both a person's outcome and their likelihood of being in the program. The effect of these factors on participants' outcomes needs to be distinguished from the effect of the program itself.
- As most Government programs are targeted in some way, it is generally not adequate to examine only data from participants, or to compare group outcomes between participants and non-participants. More sophisticated techniques need to be employed, informed by the types of confounding factors likely to be of concern, available data and program design.
- A program evaluation report should explain the limitations of the analysis used and what they mean for the conclusions of the report.
- Sometimes, isolating the causal effect of a program may not be feasible. In these cases, the Neyman-Rubin Causal Model is still useful in describing the effect that should be approximated to the degree feasible.

1. Causal effect of a program

The causal effect of a program on outcomes is a primary concern in program evaluation. For the purpose of lapsing program evaluations, DTF refers to the Neyman-Rubin Causal Model (NRCM) interpretation of causality. This model is recognised by bodies across the world, including the World Bank, as the standard by which government programs are evaluated against.

The NRCM frames causality in terms of the following question:

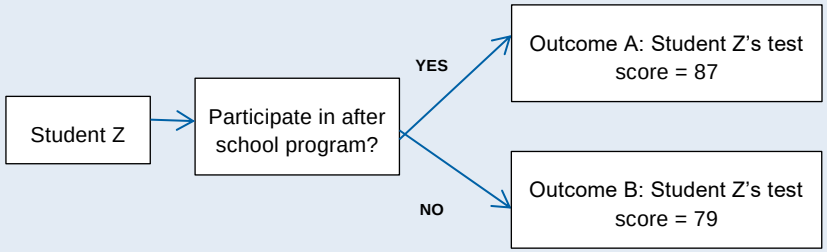
How has the program changed outcomes for participants, compared with what they would have been without the program?

The NRCM formalises the definition of a program's effect in a simple mathematical framework. For the purposes of program evaluation, it provides a mathematical statement of the above question. This allows practitioners to understand when certain statistical techniques will accurately isolate the effect of a program, and when they will be misleading.

NRCM compares actual outcomes for an individual that received a given treatment, with the potential had that individual not received the treatment. NRCM defines the difference between the actual and potential outcomes as the causal effect of the treatment. A detailed example of the NRCM, associated violations, and implications for program evaluation in practice is detailed in Box 1, below.

Box 1: Example of the NRCM and violations of NRCM

Assume that the Government introduces an after-school study program for students, and we are interested in how this program affects student test scores, as illustrated in **Chart 1**, below:



The NRCM definition of identifying causal effect

In this hypothetical example, the NRCM defines the causal effect of the after-school study program on Student Z as:

(1) Causal effect$_{(true)}$ = Student Z$_{(Outcome\ A)}$ – Student Z$_{(Outcome\ B)}$

→ 8 = 87 – 79

Approximation of NRCM

However, in reality, you cannot observe counterfactual outcome. That is, you cannot observe both Student Z$_{(Outcome\ A)}$ and Student Z$_{(Outcome\ B)}$. Student Z either participated or didn't participate in the after school program. Only one can be true.

Given this, analysts usually use the following approximation, which compares the observed outcome of a student that didn't participate in the after school program with a student that did:

(2) Causal effect$_{(observed)}$ = Student Z$_{(Outcome\ A)}$ – Student Y$_{(Outcome\ B)}$

If the approximation is valid, we would expect the following:

(3) Causal effect$_{(true)}$ = Causal effect$_{(observed)}$

Substituting **(1)** and **(2)** into **(3)**:

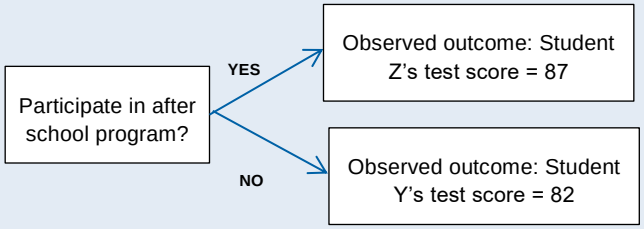
Student Z$_{(Outcome\ A)}$ – Student Z$_{(Outcome\ B)}$ = Student Z$_{(Outcome\ A)}$ – Student Y$_{(Outcome\ B)}$

→ **(4)** Student Z$_{(Outcome\ B)}$ = Student Y$_{(Outcome\ B)}$

The above equality represents the fundamental requirement of the NRCM. That any approximation of the causal effect of a program is only valid if Student Y$_{(Outcome\ B)}$ is representative of Student Z$_{(Outcome\ B)}$. Sections 2 and 3 of this document outlines statistical approaches that by its general design, seeks to satisfy this equality.

Violation of NRCM and a fundamental problem of observational statistical analysis

In practice, there are a range of distorting factors that would distort **(4)** such that Student Z$_{(Outcome\ B)}$ ≠ Student Y$_{(Outcome\ B)}$. For example, take the following observation and assume that in reality, we only observe Student Z participating in the program and Student Y not participating, as outlined in **Chart 2**:



As shown, Student Y's observed test score is 82 (Student Y did not participate in the after-school program). This may be due to the fact that Student Y has a higher academic ability. Given student Y's higher ability, he has not chosen to participate in the after-school program. Therefore:

Student Y$_{(Outcome\ B)}$ > Student Z$_{(Outcome\ B)}$ = 82 > 79

and

Causal effect$_{(observed)}$ = 87 – 82 = 5

A clear distortion has been introduced into the observed effect, since:

Causal effect$_{(true)}$ > Causal effect$_{(observed)}$

[(87-79) = 8] > [(87-82) = 5]

This illustrates a fundamental problem of observational statistical analysis. A selection bias problem has been created, where those who choose not to participate, tend to be students who have a higher academic ability and would score highly on tests, thereby distorting the measured causal effect of the after-school program.

In this case, we refer to a student's academic ability as a 'confounding' factor, which is hidden and not measured in the comparison, but which is correlated to both a student's probability of participating in the program and his test scores, thereby distorting the measured effect of the program.

The magnitude of the distortion/selection bias can be quantified by **(4)**:

Student Y$_{(Outcome\ B)}$ - Student Z$_{(Outcome\ B)}$ = 82 – 79 = 3

Traditional statistical techniques are highly vulnerable to confounding variables, as these are not designed to holistically account for confounding variables. Program evaluations relying on these techniques will, therefore, likely misestimate causal effects. DTF has set out a range recommended statistical approaches in Section 3, aimed at overcoming these issues.

2. Designing a program evaluation

Most programs' effects cannot be measured directly. A program evaluator does not observe what the outcome would have been for program recipients, had they not been in the program. Statistical techniques attempt to address this by constructing a counterfactual benchmark, representing a comparable group who did not receive the program.

The quality of a statistical analysis depends on the quality of data available. A good program evaluation depends on deliberate decisions that may not otherwise arise during program design. As many potential weaknesses in a statistical analysis cannot be addressed retroactively, **good evaluation must be designed up-front**.

The construction of a counterfactual benchmark requires that **data must be collected from non-participants.** Only using data from participants will undermine an evaluation's ability to isolate the effect of the program on participants' outcomes. For example, an improvement in employment may arise due to participation in a program, or it may be due to an overall improvement in the economy.

A program evaluator must pay attention to **confounding factors:** characteristics of participants that affect both their outcomes and their likelihood of being in a program. These will be present for most Government programs, which target or are used by people with some (dis-)advantage. For example, an employment program may target people most at risk of poor employment outcomes, or it may target vulnerable people closest to being job-ready.

Potential confounding factors need to be identified before a program is rolled out, so that data can be collected on them. It is not possible to retroactively address a certain (dis-)advantage if appropriate data has not been collected during the program.

3. Statistical techniques for program evaluation

The presence of confounding factors means it is often misleading to compare group outcomes of program participants with outcomes for some wider population. Doing so will conflate the effect of the (dis-)advantage that brought people into the program with the effect of the program itself.

There are several statistical techniques which, by its general design, try to isolate the effect of a program by constructing a comparable benchmark.

Listed below are some common ones (supporting academic papers are in the reference list):

- **Difference-in-difference** studies look at the changes in outcomes for people who have and those who have not received the program. If confounding factors largely do not change before and after a program, focusing on the change in outcomes removes their effect.
- **Propensity score matching** groups people according to how likely they are to have been in the program, based on observed characteristics. Within each group, those who received a program can be compared with those who did not.
- **Regression discontinuity** makes use of the fact that confounding factors generally do not change abruptly around the cut-off points for program eligibility. They attribute any sharp change in outcomes for people near the cut-off point to the abrupt change in program eligibility.
- **Randomised control trials** attempt to create the benchmark by randomly assigning people into a program. Random assignment means the effect of any confounding factor does not consistently influence the benchmark in any one direction.
- **Synthetic control:** similar to a difference-in-difference approach but with a more rigorous method for selecting appropriate 'control' groups based on observing outcomes

These techniques are conceptually simple, well explored in contemporary academic literature, can be implemented through most statistical software packages, and have been used in past lapsing program evaluations. These provide greater transparency compared with propriety modelling approaches, such as in CGE and input-output modelling, which are constructed from up to thousands of equations/inputs, and are difficult to objectively verify.

Because the causal effect of a program must be indirectly estimated, no technique will produce a perfect estimate. Each has its own trade-offs and will be appropriate in different situations. Residual confounding factors may be unavoidable and will be of less concern if they are unlikely to substantially distort results. A program evaluation report should **explain the limitations** of the analysis and what they mean for the conclusions of the report.

Sometimes, accurately identifying the causal effect of a program may not be feasible. In these cases, the causal effect of the program should be approximated as closely as possible. The NRCM provides a framework for understanding the consequences of the approximations used. An evaluation report should discuss this and how a program may be modified to allow a more robust estimate of its causal effect.

4. Modelling approaches that do not satisfy the NRCM

There are some commonly used modelling approaches in program evaluation, which by themselves, do not adequately approximate the causal impact of programs. These include:

- **Computable general equilibrium (CGE) modelling,** which is constructed using a series of equations that represent assumed consumer and producer behaviour. Linkages within industries and between industries and consumers are parametrised through National Accounts input output tables, though these linkages at the regional/state level are largely assumed and not based on robust data. CGE is effectively a theoretical approach to modelling, often used to estimate how the benefit from a single program propagates to the rest of the economy. By itself, it cannot represent the causal effect of a program given there is little/minimal real data input.
- **Input-output (IO) modelling** essentially constructs multipliers from the National Accounts input-output tables to extrapolate program level benefits to economy-wide benefits. By itself, IO modelling cannot establish the causal effect of a program, as it is only used to generate multipliers with which to extrapolate a program's direct benefit. DTF does not endorse the use of IO modelling for any purpose, as the multipliers estimate make unrealistic assumptions such as unlimited supply of production inputs at the current price, and constant industry input and household consumption structures.
- **Most time series approaches, including vector auto regressions (VARs)** which naively compare correlations between variables. This approach may be sufficient for forecasting purposes, where causal links between variables is a secondary consideration to the primary objective of predicting the absolute value of a given variable. However, it does not eliminate confounding factors by design, and therefore, does not approximate causal effects.
- **Naïve regression approaches** which seek to control for specific confounding variables by including them as explanatory variables in a regression. These do not treat confounding variables in a general manner, through the design of the identification strategy. Coupling this with the difficulty in obtaining appropriate data for all relevant confounding variables, this approach increases the risk of selection bias.
- **Other assumptions-based approaches** which directly assume a causal relationship between the program and some outcome, without demonstrating this empirically using evidence. CGE modelling and simpler, spreadsheet-based approaches fall under this category.

References

World Bank Guide to Impact Evaluation:

Gertler P, Martinez S, Premand P, Rawlings L, Vermeersch, Christel M (2016) Impact Evaluation in Practice, Second Edition. Washington, DC: Inter-American Development Bank and World Bank.

http://www.worldbank.org/en/programs/sief-trust-fund/publication/impact-evaluation-in-practice

Difference-in-difference:

Yang C, Hansen J, Dabo A (2017) The Value of Economic Access. Victoria's Economic Bulletin 1:23-33

https://www.dtf.vic.gov.au/sites/default/files/2018-01/Victorias-Economic-Bulletin-Volume-1.pdf

Propensity Score Matching:

Caliendo M, and Kopeinig S (2008). Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys, 22: 31-72

https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-6419.2007.00527.x

Regression Discontinuity Design:

Cameron L, Ergal N, Gangadharan L, Meng X (2013) Little emperors: Behavioral impacts of China's one-child policy. Science **339**(6122):953–957.

https://science.sciencemag.org/content/339/6122/953

Synthetic Control:

Abadie A, and Gardeazabal J (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. American Economic Review 93 (1): 113-132.

https://www.aeaweb.org/articles?id=10.1257/000282803321455188