4.5 Advanced statistical techniques

Authors

Marcella Vigneri, Centre of Excellence in Development Impact and Learning, London School of Hygiene and Tropical Medicine, London, United Kingdom.

Howard White, Campbell Collaboration, New Delhi, India; and Centre of Excellence in Development Impact and Learning, London School of Hygiene and Tropical Medicine, London, United Kingdom.

4.5.1 Learning objectives

To understand the following more advanced factors to consider in developing an impact evaluation for health emergency and disaster risk management (Health EDRM):

1. Different approaches to estimating impact in the absence of random assignment.
2. Advantages and disadvantages of these different approaches.
3. Importance of baseline data for both intervention and comparison groups.

4.5.2 Introduction

Random assignment usually provides the most robust method for comparing the effectiveness of interventions (Chapter 4.1). However, it may not be possible in some settings related to Health EDRM. For example, the implementing agency might not be willing to accept randomization, or the impact evaluation may have to be designed after an intervention is already underway or even completed. When randomization is not possible, impact can still be estimated through a range of non-experimental techniques, which may be broadly divided into two categories: quasi-experimental methods (see also Chapters 4.14 and 4.15) and regression-based approaches.

Quasi-experimental (QE) methods identify a comparison group using statistical matching, such as propensity score matching and coarsened exact matching. Matching is also used to increase the power of designs such as difference in differences, which are explained below. Matching ensures that the comparison group is as similar to the intervention group as possible, such that the average characteristics (age, location and education, for example) of the intervention and control groups are similar at baseline (that is, pre-intervention). Impact is then calculated as either the difference in outcomes after the intervention (ex-post single difference) or the difference in the change in outcomes between baseline and endline (difference-in-differences).

Regression-based approaches include instrumental variables, Heckman sample selection models, endogenous switching regressions and fixed effects models. These approaches require the use of data in untreated or less treated units. Endogenous switching models and Heckman selection models are not covered in this chapter, and information on them is available elsewhere *(1)*. Regression based approaches are usually the only option if the intervention is measured as a continuous indicator (for example changes in the amount of exposure to the intervention), rather than as a binary indicator (that is, the intervention is either provided or not provided).

Non-experimental approaches are best based on specifying the underlying structural model, that is the set of behavioural relationships which lead to intervention impact (see Chapter 4.10). Applying non-experimental approaches requires data from both an intervention and a comparison population. Moreover, more reliable impact estimates are usually possible if baseline data are available that provide variables for matching that are unaffected by the intervention, since such data were collected before the intervention took place.

This chapter introduces three common matching techniques: propensity score matching, regression discontinuity and interrupted time series, as well as one regression-based approach: instrumental variable estimation. First, the following section explains how impact can be estimated using differencing.

4.5.3 Double difference estimates

When the intervention has taken place, impact can be estimated by single or double difference. Table 4.5.1 shows the different stages of an intervention (top row) and the data that are required to apply these approaches.

Table 4.5.1 Timing of intervention and surveys for large impact evaluations

Start of intervention	During intervention	At end of intervention	After intervention
B: Baseline	M: Mid-term	E: Endline	P: Post-endline

Description

Ex-post single difference impact estimators are calculated as the difference between the outcome indicator after the intervention (that is, at endline, time E) in the intervention group and the outcome indicator in the comparison group which did not receive the intervention. The double difference impact estimate is the difference in the change in the outcome indicator for the intervention and for the comparison groups between baseline and endline, rather than the difference in their endline values, as is the case for the single difference. Double differencing removes any difference in the indicator between intervention and comparison groups that was present at baseline. This is useful because these baseline differences cannot be a result of the intervention. If the values of the outcome indicators for the intervention and the comparison groups are the same at baseline, then the single and double difference estimates are equivalent.

Double differencing is a means of calculating the estimated impact. It is also used as an impact evaluation method. Double difference estimates require baseline data that should be collected immediately prior to the intervention. The validity of this approach relies on the 'parallel trends assumption', that is, the trend in the outcome in intervention and comparison populations should be the same without the intervention. The parallel trends assumption can be tested *(2)* if trend data from before the intervention are available, but unfortunately this is often not the case. Acquiring more data points (observations) before and after the intervention allows a visual inspection of whether the parallel trend assumption holds. If the assumption can be tested and does not hold, then using double differencing without matching cannot be expected to be free of bias. Matching can help to control for observable determinants of differences in changes over time and make the analysis less dependent on this assumption. Implementation of the method requires data on outcomes from the intervention and comparison groups at baseline and endline. If matching is to be used, then data for matching are also required.

Advantages and disadvantages of double differencing

Double differencing is easy to implement and easy to understand. However, pre-intervention trend data may not be available to test its validity. Hence, it is more rigorous when used with a matching technique.

4.5.4 Propensity score matching

Propensity score matching (PSM) creates a comparison group from observations on a population that did not receive the intervention by matching intervention observations to one or more observations from the sample without the intervention, based on observable characteristics. Matching is based on the propensity score, which is the estimated probability of being in the intervention group given the observable characteristics. The propensity score is estimated using a regression model of participation (taking part in the intervention). Propensity score matching cannot incorporate selection on unobservables, so may give biased estimates if these are important. Additional information is available elsewhere *(3–5)*.

Description

Perfect matching would require matching each individual or unit in the intervention group with a person or unit in the comparison group that is identical on all relevant observable characteristics (for example, age, education, religion, occupation, wealth, attitudes to risk and so on). Clearly, this is not possible nor is it necessary. 'Balance' between intervention and comparison group units (which is necessary for unbiased estimates) requires that the average characteristics of the intervention and comparison groups are the same before the intervention. A good example on the methods used for variable selection in PSM is provided by Brookhart and colleagues *(6)*.

In PSM, matching is not achieved on every single characteristic but on a single number: the propensity score. This is the likelihood of a person taking part in the intervention given their observable characteristics. This probability is obtained from the 'participation equation': a probit or logit regression in which the dependent variable is dichotomous, taking the

value of 1 for those who took part in the intervention and 0 for those who did not. The right-hand side of the equation includes all observed variables (individual, household or firm and community or market) that may affect participation, but that are not affected by the intervention. Baseline values of all variables, including outcomes, cannot be affected by the intervention, so having baseline data helps to obtain a stronger match.

Observations outside the 'region of common support' are discarded before matching. The region of common support is the area of overlapping propensity scores. Therefore, those observations with very low scores (which typically come from the comparison group) or very high scores (typically from the intervention group) are discarded. The observations retained from those who did not receive the intervention are used as the comparison group, which ensures that the comparison is 'like with like'.

Each member of the intervention group is matched to one or more members of the comparison group. This is done through a variety of matching algorithms such as the nearest neighbour matching, caliper matching and kernel matching. An example is the study by Boscarino and colleagues *(7)* which uses PSM to estimate the impact of mental health interventions received by employees at the worksite after the World Trade Center attacks among workers in New York City. The authors used data from telephone interviewees with adults in a household survey conducted one and two years after 9/11 to match intervention cases to non-intervention control cases based on a bias-corrected nearest-neighbour algorithm. Their findings from matching with PSM suggest that about 7% of approximately 425 000 adults reported positive outcomes (such as reduced alcohol dependence, binge drinking, depression, severity of post-traumatic stress disorder and anxiety symptoms) resulting from receiving employer-sponsored, worksite crisis interventions related to the attacks.

In PSM, those members of the comparison group that do not match those in the intervention group are discarded. Once matching is completed, a balancing test is performed to ensure there is no statistically significant difference between the mean characteristics of the matched intervention and comparison groups. Finally, the impact is estimated by calculating the difference between the outcome indicator of interest for the intervention units and the average value for the matched comparison individuals, and then averaging over all these differences. Another interesting application of PSM is the study by Gomez and colleagues *(8)* which exploits data collected as part of a large-scale evaluation of an early childhood education intervention related to earthquakes in Santiago, Chile. The data included 4-year old children who had experienced, and who had did not experienced, the severe earthquake episodes of 2010. These children were then matched through PSM to find that the earthquake affected lower scores on some early language and pre-literacy assessments of children that had experienced the earthquake. A further example is provided as Case Study 4.5.1, which assessed the impact of humanitarian aid on food security in the Republic of Mali.

There are several statistical packages (such as Stata and R) that allow to implement PSM analysis through pre-built commands.

4.5

Advantages and disadvantages of propensity score matching

The two main advantages of PSM are that it easily lends itself to establish the propensity score of being treated through a binary model, and that it can be done ex post, including in the absence of baseline data. If baseline data are not available, matching uses time invariant characteristics (such as sex and religion) and recall information on pre-intervention characteristics that can be reliably recollected. These features suggest the greater flexibility of the PSM model to accommodate many covariates.

Case study 4.5.1
Using PSM to measure the impact of humanitarian aid on the food security of rural populations in Mali *(9)*

PSM was used to measure the impact of humanitarian aid on the food security of rural populations in the Mopti region of Northern Mali.

The evaluation exploited data from a unique pre-crisis baseline in the region to use matched difference-in-difference methods to estimate whether access to different forms of food assistance improved household food expenditures, food and nutrient consumption, and the long-term nutritional status of children. The existence of baseline data enabled the matching of 'intervention' households with comparable 'comparison' households.

The measures used for matching were all pre-intervention (and so unaffected by it) and relate to both the selection into intervention and the outcome of interest (household expenditures, food consumption and a proxy for child nutritional status). The matching variables were both village-level measures (the presence of a secondary school within 5 km and the presence of a market within 5 km) and household-level measures (including whether children were involved in past projects, feelings of safety and age of the household head).

The impact evaluation found that food assistance increased household non-food and food expenditures and micronutrient availability.

A disadvantage of PSM is that it relies upon matching on observables. If selection (participation) into the intervention is affected by unobservables, PSM will yield biased impact estimates for ex-post single difference estimates. When panel data are available, PSM is biased if the unobservables are time varying or affect differences over time. However, time invariant observable factors can be removed by double differencing, so that PSM would again be unbiased.

4.5.5 Regression Discontinuity Design and Interrupted Time Series

Regression discontinuity designs (RDD) are used when there is a threshold rule for allocation to the intervention (such as administration of a drug if patient has a heartrate or temperature above a specific value, or the poverty line, or villages on either side of an administrative boundary). The assumption, which is tested as part of the procedure, is that units in proximity to either side of the boundary are sufficiently similar for those excluded from the intervention for these to be a valid comparison group. The difference in outcomes between those near either side of the boundary, as measured by the discontinuity in the regression line at that point, is attributable to the intervention, and so is the measure of the intervention's impact.

Interrupted time series (ITS) is a specific application of RDD in which the threshold is the point in time at which the intervention came into effect. This can be a particularly relevant method where intervention effectiveness is sudden, rather than gradual, such as the completion of a bridge or major power transmission connection, or the sudden availability of relief services.

Description

RDD can be used when there is a threshold rule that determines eligibility for the intervention, where the threshold is based on a continuous variable assessed for all potentially eligible units of assignment (such as individuals, households or communities). For example, households above or below the poverty line, children born before or after the cutoff date for school enrolment in a specific academic year, or students above a certain test score are awarded a scholarship. If the threshold is imperfectly applied, a variation on the approach, called 'fuzzy RDD', can be used.

The threshold variable must not be one which can be manipulated to become eligible for the intervention, as that might lead to selection bias. As an example, an impact evaluation of the Tropical Cyclone Winston social protection top up transfers was conducted by the World Bank in 2016 *(10)*. The goal of the intervention was to provide additional assistance in the form of top-up transfers to the most vulnerable, as a key component of its disaster response, and the intervention and control groups were constructed based on the Poverty Benefit Scheme (PBS) eligibility (poverty score) threshold. The treatment group was formed from PBS recipient households (20% below threshold) in affected areas in the Republic of Fiji that would also receive the intervention (top-up PBS benefit) after the cyclone. The control group was formed from the PBS-evaluated (before the cyclone) households in affected areas that were not eligible for PBS, as they were above (but within 20%) the threshold. The disaster responsive social protection intervention, in the form of top-up transfers to beneficiaries, was found to be an effective response following the cyclone.

In ITS, the threshold is the point in time at which the intervention or policy was introduced. In the case of a policy, this point in time is common to all households but other interventions (such as electrification or connection to a sewage disposal system) may affect different communities at different points in time. The threshold should be unique to the intervention. Clearly, those on either side of the threshold have some differences. In addition, the threshold criteria may be correlated with the outcome, so that there is

selection bias if simple comparisons are made. For example, scholarships are awarded to improve learning outcomes, but those with better learning outcomes are given the scholarships. Older women are more likely to get breast cancer, and it is older women who are selected for screening for this cancer. However, those near either side of the threshold are also much more similar. Regression discontinuity is based on a comparison of the difference in average outcomes for these two groups.

Another interesting application of this method comes from the study of Mezuk and colleagues *(11)* who used the September 11 2001 attack as the discontinuity (cut-off) point to investigate its impact on the average monthly suicide rate in New York City. Using average monthly suicide rates data between 1990 and 2006, the study found no net change in suicides rates just before and immediately after the attacks, suggesting that factors other than exposure to that particular traumatic event may have been driving the risk of suicide in the population studied.

An iterative approach is used to determine the margin around the eligibility threshold. Initially, one sets a small margin and checks for balance of the resulting intervention and comparison group units. If the match is good, the margin may be widened a little and balance checked again. This can be repeated until the samples start to become dissimilar (that is, there is no longer balance between the two groups). When the sample is established, a regression line is fitted to the sample around the threshold. The sample for the regression is restricted to observations just on either side of the threshold. Specifically, the outcome indicator is regressed on the selection variable (such as test scores and an intercept dummy). The intercept dummy is a dichotomous variable, taking the value 0 for observations below the threshold and 1 at the threshold and above it.

Advantages and disadvantages of RDD

RDD controls unobservables better than other quasi-experimental matching methods. It can also often use administrative data, thus reducing the need for data collection (see Chapters 2.4 and 4.4). The main limitation of RDD is that it is usually valid only for observations relatively close to the discontinuity point. Hence, a challenge for RDD is often to find a sufficiently large sample of observations on either side of the threshold. Further, the impact is being estimated only for the population close to the threshold. The estimate is what is called a local area treatment effect (LATE), rather than an average effect for the whole population in the intervention group. In principle, this limitation restricts the external validity of the approach.

Case Study 4.5.2 provides an example of how RDD was used to measure the impact of a winter cash assistance programme for Syrian refugees in Lebanon.

Case Study 4.5.2
Using RDD to measure the impact of a winter cash assistance programme to Syrian refugees in Lebanon *(12)*

The evaluation assessed the impact of cash on household well-being among Syrian refugees in Lebanon and whether cash might attract refugees to regions with assistance. The RDD design exploited the targeting approach of the cash assistance programme itself. Cash was given at high altitudes to target assistance for those living in the coldest areas during the winter months (households did not know beforehand that there would be an altitude eligibility cutoff). When the eligibility cutoff was set at 500 meters, households residing at 501 meters and above (intervention group) were included, while households residing at 499 meters or below (comparison group) were excluded. Intervention and comparison groups had very similar characteristics before the start of the programme, so differences measured after the programme's implementation represent the causal impact of cash assistance.

The impact evaluation found that the current value of cash assistance was inadequate because beneficiaries' income was so low that they were forced to use the cash assistance to satisfy other basic needs, in particular food. It also found that cash assistance increased access to school, reduced child labour and that the cash assistance programme had no pull factor on refugees settling in communities where cash was distributed.

4.5.6 Instrumental variables approach

The instrumental variable (IV) method is a regression-based estimation of the outcome variable of interest on either a project dummy or a measure of participation in the intervention group *(13)*.

In the conventional ordinary least squares (OLS) approach, the outcome is regressed on a dichotomous intervention dummy variable. The problem with this approach is that selection bias can affect the estimate of the impact coefficient. If selection is entirely based on observables, and the regression has included variables on all those observables, then OLS will indeed yield a valid impact estimate. However, if – as is more frequently the case – there are time varying unobservables, then cross sectional OLS models on differences will yield biased impact estimates. IV estimation is the technique used to remove the bias. It is an OLS regression in which the variable which is the source of the endogeneity problem is replaced by an instrument satisfying the following two conditions:

i. To be correlated with the probability of intervention (programme participation)
ii. To be uncorrelated with the outcome, except through its effect on the intervention.

When more than one instrumental variable is identified, the procedure is implemented as two-stage least squares: first one regresses the endogenous variable (the one measuring intervention participation) on the instruments and calculates its fitted value, then the outcome equation is estimated replacing the endogenous variable with the fitted values from the first stage. The estimated impact is the coefficient on the instrument. It is

important to have determined the instruments before data collection starts, so that the relevant questions are included in the survey instruments.

Advantages and disadvantages of IV

The advantage of IV is that if a valid instrument is found, both observable and unobservable sources of selection bias are controlled for. The main disadvantage of the method is that it may be difficult to find a valid and defendable instrument, because many factors that affect decisions to use an intervention typically also affect outcomes.

Case Study 4.5.3 provides an example of the use of IV to measure the political effects of environmental change.

Case Study 4.5.3
Using instrumental variables to measure the political effects of environmental change to understand the disaster–violence nexus *(14)*

In 2004, Sri Lanka was hit by a massive tsunami that killed more than 35 000 people and destroyed over 78 000 homes in that country alone. By May 2006, the Government of Sri Lanka had spent more than US$200 million on recovery, reconstructing at least 40 000 houses *(14)*. This study examined whether post-disaster reconstruction triggered further intrastate violence to explain civil unrest after the disaster.

The author addressed the endogeneity problem between reconstruction processes and violence (that is, that reconstruction is endogenous to violent events, but noted that there may be also a reverse causation if future violence limits current reconstruction efforts in disaster zones) by using the wave heights in the tsunami as an IV for post-war housing reconstruction.

The results suggest that an increase in housing construction is associated with the number of violent events, while the number of destroyed houses has no discernible impact on violence. Therefore, the paper plausibly concludes that reconstruction is a manipulable strategy that policy makers can use to respond to disasters through different post-disaster measures.

4.5.7 Conclusions

The chapter introduces some of the non-experimental quantitative methods that are available for impact evaluation studies in Health EDRM. These approaches are likely to be appropriate in establishing impact of interventions when random assignment is not be possible. Strengths and limitations of these approaches are illustrated with references to specific studies from disasters and other health emergencies. In general, best practice in planning a research study is to consider which approach is most appropriate and feasible at the design stage in order to prepare data collection tools and think of the best sampling strategy to get a good match. For example, PSM requires that data collection includes suitable matching variables and IV requires that data is available for one or more valid instruments. Oversampling will be necessary if observations will be discarded in establishing the regional of common support.

Moreover, where possible, it is best to use a combination of methods to ensure the most reliable and credible results on the impact of the intervention being assessed. For example, it is much better when possible to exploit baseline data for matching and using the difference-in-difference strategy. Similarly, if an assignment rule exists for the project, it would be ideal to match on this rule and subsequently do a regression discontinuity design.

4.5.8 Key messages

- **Impact estimates are possible in the absence of randomization, but still need data from a comparison group that did not receive the intervention.**
- **The available methods may be subject to selection bias.**
- **It is important to test for baseline balance to check if bias based on observables has been removed.**
- **The reliability of matching and the ability to calculate a double difference estimate are enhanced by the availability of baseline date for the intervention and comparison groups.**

4.5.9 Further reading

Allaire MC. Disaster loss and social media: Can online information increase flood resilience? Water Resources Research; 2016: 52(9): 7408-23.

White H, Sabarwal S. Quasi-experimental Design and Methods, Methodological Briefs: Impact Evaluation 8. Florence, Italy: UNICEF Office of Research. 2014.

4.5.10 References

1. White H, Raitzer D, editors. Impact Evaluation of Development Interventions: A Practical Guide. Asian Development Bank. 2017.
2. Angrist JD, Pischke JS. Mostly Harmless Econometrics. Princeton University Press. 2019.
3. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research. 2011: 46(3): 399-424.
4. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology: 1974: 66(5) 688-701.
5. Rubin DB. Estimating causal effects from large data sets using propensity scores. Matched Sampling for Causal Effects. 2006.
6. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. American Journal of Epidemiology. 2006: 163(12): 1149-56.
7. Boscarino JA, Adams RE, Foa EB, Landrigan P. A propensity score analysis of brief worksite crisis interventions after the World Trade Center disaster: Implications for intervention and research. Medical Care. 2006: 44(5): 454-62.
8. Gomez CJ, Yoshikawa H. Earthquake effects: Estimating the relationship between exposure to the 2010 Chilean earthquake and preschool children's early cognitive and executive function skills. Early Childhood Research Quarterly. 2017: 38: 127-36.
9. Tranchant JP, Gelli A, Bliznashka L, Diallo AS, Sacko M, Assima A, et al. The impact of food assistance on food insecure populations during conflict: Evidence from a quasi-experiment in Mali. World Development. 2019: 119: 185-202.
10. Mansur A, Doyle J, Ivaschenko O. Cash Transfers for Disaster Response: Lessons from Tropical Cyclone Winston. SSRN Electronic Journal. 2018. https://ssrn.com/abstract=3143459 (accessed 29 February 2020).
11. Mezuk B, Larkin GL, Prescott MR, Tracy M, Vlahov D, Tardiff K, Galea S. The influence of a major disaster on suicide risk in the population. Journal of Traumatic Stress. 2009: 22(6): 481-8.
12. Masterson D, Lehmann C. Emergency Economies: the Impact of Cash Assistance Program in Lebanon. International Rescue Committee. 2014. https://www.rescue.org/sites/default/files/document/631/emergencyeconomiesevaluationreport-lebanon2014.pdf (accessed 29 February 2020).
13. Angrist JD, Imbens GW, Rubin DB. Identification of Causal Effects Using Instrumental Variables. Journal of the American Statistical Association. 1996: 91(434): 444-55.
14. Kikuta K. Postdisaster Reconstruction as a Cause of Intrastate Violence: An Instrumental Variable Analysis with Application to the 2004 Tsunami in Sri Lanka. Journal of Conflict Resolution. 2019: 63(3): 760-85.