Non-experimental methodologies for quantitative analysis

Markus Frölich, Andreas Landmann, Markus Olapade, and Robert Poppe
6.1. Introduction

The ultimate objective of quantitative analysis is to establish causality. Researchers want to know the causal influence of a factor—the effect that can be attributed to this factor and to the factor only. Done correctly, quantitative analysis allows both quantifying the magnitude of this causal effect and computing statistical precision of estimation (confidence intervals). The goal of an impact evaluation is to measure the causal effect of a policy reform or intervention on a set of well-defined outcome variables.

Knowing causal relationships is useful for making predictions about the consequences of changing policies or circumstances; they answer the question of what would happen in alternative (counterfactual) worlds. As an example, one could try to identify the causal effect of introducing a community-based health insurance on health status or on out-of-pocket spending for health of the insured in a specific district of a developing country.

6.2. Selection bias and comparison issues

The fundamental problem of impact evaluation is the impossibility of observing an individual in two states at a moment in time; each individual is either in the programme under consideration or not, but not both. The impact of a development programme can only be identified by comparing realised outcomes of those who did receive and of those who did not receive an intervention. Thus, data on non-participating individuals needs to be collected as well. The issue of selection bias is of central concern in this context: selection bias may arise when treated and non-treated individuals are different with respect to observed and unobserved characteristics. One reason could be for example, a project manager who deliberately chooses some individuals to be eligible for the programme but not others. Another important source for selection bias is self-selection, i.e., when individuals themselves choose to be treated or not.
Thus, any impact evaluation should be based on a detailed understanding about why some individuals or communities participated, whilst others did not participate in the programme; otherwise, results are likely to be biased.

**Box 1: Potential outcomes and objects of interest**

To make ideas more precise, let $Y_i^1$ denote the outcome if individual or community $i$ is exposed to development intervention $D$. Before programme start, each individual (or community) has two hypothetical outcomes: a potential outcome $Y_i^1$ if individual $i$ participates in the programme, and a potential outcome $Y_i^0$ if individual $i$ does not participate in the programme. The causal effect of the intervention is defined as the difference between $Y_i^1$ and $Y_i^0$, i.e., the effect of participation in the programme relative to what would have happened had individual $i$ not participated in the programme. The individual effect of the intervention is usually averaged over the population of interest, defined as the average treatment effect $E[Y^1 - Y^0]$. It can be interpreted as the average treatment effect for a person randomly drawn from the population or, alternatively, as the expected change in the average outcome if the individual status indicator variable of development intervention $D$ were changed from 0 to 1 for every individual (provided that no general equilibrium effects occur), where individual $i$ either receives ($D=1$) or does not receive ($D=0$) the treatment. In a policy evaluation context of particular interest is the average treatment effect on the treated (ATT) defined as $E[Y^1 - Y^0|D=1]$. It may be more informative to know how the programme affected those who chose to participate in it than how it affected those who could have participated but decided not to.

One way to avoid selection bias is to randomly assign individuals to a treatment and a control group. We refer to randomised trials as methods that randomly assign individuals who are equally eligible and willing to participate into distinct groups; they are generally considered the most robust of all evaluation methodologies and sometimes referred to as the *gold standard* (Angrist 2004). Given appropriate sample sizes, the two groups will have approximately the same characteristics and differ only in terms of the treatment status. They will be approximately equal with respect to variables like race, sex, and age, and also for difficult to measure variables, such as lifestyle-related risks, quality of social networks, and health awareness.
In practice, there are several problems with randomisation. Firstly, it may be unethical to deny access to benefits for a subgroup of individuals. Secondly, it may be politically unfeasible to randomly deny access to a potentially beneficial intervention. Thirdly, there may not be any individuals who are unaffected if the scope of the programme is nationwide. Fourthly, problems arise when, after randomised assignment, individuals cross over to the treatment group. For example, people might travel to another municipality to buy insurance after having learned that a microfinance institution offers a new life insurance scheme there. Fifthly, individuals assigned to the treatment group may not take up treatment, or individuals assigned to the control group may seek similar treatment through alternative channels.

Sometimes, randomised trials are impractical. However, impact evaluations are most valuable when we use data to answer specific causal questions as if in a randomised controlled trial. In absence of an experiment, we may look for a natural or quasi-experiment that mimics a randomised trial in that there is a group affected by the programme and some control group that is not affected. If it is credible to argue that the groups do not differ systematically, such a natural or quasi-experiment can be used for evaluation instead of a randomised experiment.  

---

For a discussion see Burtless and Orr (1986).
Unfortunately, it is often hard to justify that programmes that were not ex-ante planned as randomised experiments do in fact fulfill this criteria. This is why it is preferable to think about evaluation and the appropriate design before the programme is implemented.

In non-experimental studies, researchers often try to approximate a randomised experiment by using statistical methods. We will discuss several non-experimental methods in the following paragraphs. The issue of selection bias is of central concern here. Rather complex statistical methods are required in order to deal with selection bias when using non-experimental data. The methods differ in the way they correct for differences in (observed and unobserved) characteristics between the treatment and the control group and by their underlying assumptions.2

Two non-experimental methods—differences in means and before-after estimation (also called reflexive comparisons)—usually do not give a satisfactory solution to the selection issue when using non-experimental data. In general, this is because changes in the outcomes cannot be attributed to the programme. The former method, differences in means, is based on cross-sectional data using the outcome of the non-participants to impute the counterfactual outcome for the participants. The underlying assumption is that individual characteristics, on average, do not play a role for the difference between the treated and the non-treated, which is a strong assumption. The latter method, the before-after estimator, is based on (at least) two cross-sections of data—one cross-section before programme start and one cross-section after the programme. It uses the participants’ pre-intervention outcome to impute the counterfactual outcome for the participants. The drawback of this method is that it is impossible to separate programme effects from general effects that occurred during the same period.

Throughout this chapter, a hypothetical example will display the different evaluation methods. By use of a microinsurance example for inpatient and outpatient hospital visits, we will explain the main concepts to determine the effect of this insurance on our outcome variable:

2 For a more general and accessible introduction to impact evaluation see further Leeuw and Vaessen (2009).
number of hospital visits. This example can then be generalised to any other insurance and outcome of interest.

Figure 1 displays the impact health insurance has on the number of hospital visits. At time $t=0$, the microinsurance for inpatient and outpatient services in a nearby hospital is introduced in our village. At that point of time the villagers visit the hospital 1.5 times a year on average. At time $t=1$, the number of hospital visits has increased to 3.5 visits per year (blue line). Without the introduction of microinsurance, the frequency would have increased to 2.5 times only (dotted red line). As we will explain, the outcome 2.5 constitutes the counterfactual outcome. Thus, the (true) impact of the microinsurance scheme is 1 hospital visit per year.

**Figure 1: Real impact of microinsurance on hospital visits**

The cross-sectional comparison by difference in means is our starting point for the impact evaluation. Hereby, we compare two villages at $t=1$, where it so happens that in the treatment village microinsurance is available. If the inhabitants from this village buy a policy, they can make claims to the insurer after a hospital visit. Inhabitants from the control village do not have this opportunity and need to pay the total hospital bill from their own money. In a cross-sectional comparison we directly compare the hospital visits of the treated and untreated groups.

Figure 2 displays the naïve estimator of the treatment effect using a cross-sectional comparison. The treatment village has 3.5 hospital visits a year per person on average, where village 2 has only two visits per year. The naïve estimator is simply the difference between those two outcomes, which is 1.5, a biased estimate of the treatment effect.

It is obvious that the effect might be partially attributable to the availability of a health insurance, but it cannot be ruled out that other reasons have also affected this result.

1. Were the two villages different from the outset? Different in observables like wealth, education, or different in unobservables such as trust in hospital staff?
2. Has one of the villages had other influences that might affect the number of hospital visits? The road leading to the hospital might have been blocked for village 2, hindering people from visiting the hospital and leading to an overestimation of our treatment effect.

Myriad scenarios can be constructed to answer these two questions. According to our example from figure 1, however, an actual comparable control village should have 2.5 hospital visits per year in $t=1$. This shows that using cross-sectional comparison, the impact evaluator cannot be sure whether the effect of microinsurance on hospital visits results only from the availability of microinsurance or also from other confounding factors.

Another naïve estimator would be the before-after comparison. Here, we need to conduct a baseline survey amongst the population of the treatment village (shortly) before the microinsurance is made available. We would ask explicitly how often a week the inhabitants go to the hospital. Then the insurance is made available and after a certain period of time the same survey questions are gathered again from the village. Of course, the time between the introduction of the microinsurance and the follow-up survey needs to be long enough for certain incidences of sickness and claims to occur. As depicted in figure 3, the estimated impact of our hypothetical insurance is the difference between the number of hospital visits

---

Figure 2: Cross-sectional comparison

![Cross-sectional comparison](image)

Neither before-after nor comparison of means analyses between two villages provide reliable impact results.

---

3 Of course, it might be more reliable to gather this information directly from the hospital: “How many inhabitants from village 1 visit the hospital a week?”
visits before and after introduction, which is two hospital visits per year. The interpretation would be that introducing microinsurance that covers the cost of inpatient and outpatient care increases the number of hospital visits by two visits per year.

However, this estimator relies on the important assumption that, in between our two surveys, no other factors have occurred that might cause a change in hospital visits. This means that, for the before-and-after estimator to produce reliable results, our researcher must be sure no other confounding effects have occurred between the two surveys. In fact, figure 1 shows that without microinsurance there still would have been an upwards trend in hospital visits and that, therefore, our before-after comparison delivers unreliable results. Reasons for an upwards trend in hospital visits independent of the treatment could be increases in prosperity, decreases in transportation costs, and many other scenarios.

In sections 6.3 through 6.6 of this chapter, the following four non-experimental approaches will be explained:

1. Instrumental variables
2. Regression discontinuity design (RDD)
3. Propensity score matching (PSM)
4. Difference-in-differences (DID)

Amongst them, the first two approaches, if applicable, usually give the most convincing results. We will also stress the importance of internal and external validity in each case. Internal validity is the extent to which the results are credible for the population under consideration. External validity is the extent to which this subpopulation is representative for the whole population (of interest). Some
methods give results with high internal validity, but low external validity, and vice versa. We conclude with a discussion of where non-experimental methods should be applied.

6.3. Instrumental variable

Selection bias occurs when an omitted variable has an effect on the outcome variable of interest and the treatment. It is also called selection on unobservables, whereby treatment selection is affected by a variable that the researcher cannot observe in the data. For example, individuals with insurance could have had a higher (unobserved) awareness for health issues from the outset. Consequently, they would show different health behaviour, than those without insurance. Figure 4 illustrates this simple case with arrows indicating directions of influence and dashed lines indicating unobserved variables (health awareness). The instrument affects insurance take-up without being itself affected by different levels of awareness about insurance. In the absence of a good instrument, one could not tell apart the insurance’s effect and the awareness’ effect on hospital visits. Instrumental variable methods solve this problem of omitted control variables. An instrumental variable is a variable which has an effect on whether an individual takes up or does not take up treatment and at the same time is permitted to affect the outcome variable of interest via the treatment variable only. This is called the exclusion restriction. In other words, individuals with different values of the instrument differ in their treatment status. But otherwise, these individuals are comparable. Often, the exclusion restriction will be only valid conditionally, that means when controlling for individual characteristics.

Figure 4: Setup with instrumental variable
Box 3: How the instrumental variables method solves the problem of unobservables

More formally, instrumental variable $Z$ affects treatment status $D$, but there is no direct relationship between $Z$ and the outcome variable $Y$. Hence $Z$ is allowed to affect $Y$ only indirectly via $D$. Suppose we have

$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$ where $Y_i$ is the outcome variable, $D_i$ is the treatment indicator variable, $X_i$ is an unobserved variable that is correlated with $D_i$, and $\epsilon_i$ is a random error term.

If we now estimate (since $X_i$ is unobserved),

$Y_i = \alpha + \beta D_i + \epsilon_i$

$D_i$ will be correlated with the residual $\epsilon_i$ because, in effect, we have

$Y_i = \alpha + \beta D_i + (\gamma X_i + \epsilon_i)$.

$X_i$ and $D_i$ are correlated, so $D_i$ and $(\gamma X_i + \epsilon_i)$ are correlated as well and, thus, $D_i$ is endogenous.

The instrument can now be used to get an unbiased estimate of the effect of the endogenous variable. Researchers use a method that is called two-stage least squares: in the first stage, the instrument(s) $Z$ is used to give estimated values of the endogenous treatment variable $D$ for every individual (or community):\(^4\)

$\hat{D}_i = \alpha + \beta Z_i + \epsilon_i$

Then, in the second stage, this new variable is plugged in the equation of interest:

$Y_i = \alpha + \beta \hat{D}_i + \epsilon_i$.

The coefficient $\beta$ gives now an unbiased estimate of the treatment effect.

Using an instrument for the evaluation of microinsurance has not been done often and, in general instruments are hard to find. We therefore use a hypothetical example to illustrate our point in the context of this guide. Suppose the government sets up a health insurance programme for the poor. Everybody who is interested has to register

\(^4\) Note that using predicted values as additional regressors in the way presented here only works in linear models.
and purchase the product at the local insurance administration centre of the neighbourhood or municipality. Now imagine two households that are very close, but on different sides of the border between two neighbourhoods. The distance to the administrative centre might differ considerably, but otherwise the two neighbours should be very similar. For such pairs, distance to the administrative centre could be used as a predictor of insurance take-up that is otherwise unrelated to individual characteristics—in other words, a good instrument. Here, the instrument is correlated with insurance take-up but not with awareness. Thus, instead of comparing the treated to the untreated, we compare those with high values and low values of the instrument. This example is analogous to the famous distance-to-school instrument used by Card (1995) for schooling. Typically, an instrument requires including additional $X$ variables, e.g., quality of the neighbourhood, degree of urbanisation, family background, etc.

We may also generate instrumental variables ourselves by randomly assigning incentives or encouragements to individuals (random encouragement design). This approach looks very much like a proper randomised experiment, except that we have imperfect control over the beneficiaries. An encouragement or incentive is given to the individuals in the treatment group, whilst the individuals in the control group do not receive such an encouragement or incentive (or receive a different one). It is up to the individuals whether they sign up for the actual treatment. For example, imagine that the price of insurance is varied randomly across communities, creating a random incentive to buy insurance for the population facing a lower price. The instrumental variable that is generated here helps resolve the problem of selection bias and allows consistent estimation of the effect that insurance take-up has on health and other outcome measures. Similarly, we may vary the effort related to take-up by, for example, varying service hours, density of offices in a community, etc., from the insurer’s side. If areas or individuals cannot be exclusively chosen for a programme at random, we may at least give them varying incentives to do so.

If there is an instrument that fulfills the exclusion restriction as explained above, internal validity is high. However, external validity depends on another quality of the instrument. If the instrument predicts treatment status accurately, external validity is also likely to be high. Otherwise the instrumental variable results cannot be generalised to the whole population. The reason is that only those who are induced to take up treatment by the instrument can be used for the estimation of the treatment effect.
6.4. Regression discontinuity design

Although not as rigorous as random assignment, the regression discontinuity design (RDD) approach may give more convincing results than propensity score matching (PSM) and difference-in-differences (DID) methods (see below). The idea of RDD is to exploit some cut-off point that is important for selection into treatment and compare people near this cutoff. Thus, this approach implicitly compares treated subjects to a control group that is very similar.

A standard application is when enrollment into treatment is limited and selection of participants is conducted according to an objective rule. Frequently, such form of targeting is done on the basis of a poverty index: individuals above the threshold receive the treatment, whilst it is withheld from those individuals below the threshold. When comparing individuals very close to this threshold, their characteristics barely differ, except with respect to their treatment status. It is basically random as to whether an individual is below or above the cutoff given that the individual is close to the cut-off value. However, for this approach to be valid, individuals must be unable to manipulate their value of the index such that they would become eligible for the treatment. To test for the plausibility of this assumption, we can use statistical methods. Although discontinuities in evaluation studies are often unplanned, they may also be integrated ex-ante.

To give an illustrative example, imagine the government wants to introduce microinsurance especially for the poor. The village administration is responsible for the distribution of the insurance and relies on a poverty index to determine the eligible households. Only households that are considered poor under this index are eligible and can buy the insurance. Such discontinuities do not necessarily have to be

\[5\] Even if they have some influence, the approach is feasible as long as they are unable to manipulate their assignment precisely. The solution in this case is the so-called fuzzy regression discontinuity design.
planned as part of the intervention (even though it is certainly beneficial to have it planned beforehand). Instead, the evaluator could detect and exploit any rule used in practice to determine participation in the programme.

Figure 5 shows the number of hospital visits by ranking in the poverty index sometime after the programme started in our hypothetical village.

We see that the number of hospital visits increases and that there is a jump exactly at the poverty line. This jump is a result of the microinsurance programme and the restriction that only poor people have access to this programme. Any household that is above the poverty line has no access to the insurance product. RDD assumes that households just above the poverty line are, in fact, similar to those slightly below the index in all relevant aspects.

Therefore, we can use the households that are eligible for the insurance and very close to the threshold as our treatment group, whilst those slightly above the threshold serve as control group.

The benefit of RDD is that it does not need actual randomisation. However, the interpretation of the estimated impact is limited to the population that is close to the threshold. As a result, external validity of this approach is rather limited. Further, it usually requires a large sample for estimation.

### 6.5. Propensity score matching (PSM)

The basic idea of propensity score matching is to match at least one non-participant to every participant with identical or highly similar values of observed characteristics $X$. The difference in outcome, $Y$, between these
two individuals is then averaged over the whole population. A practical simplification is to match non-participants to participants on the basis of the propensity score, which is defined as the probability of treatment.

Box 4: How matching eliminates selection bias

Using matching, the average treatment effect on the treated (ATT) can be defined as

$$\text{ATT} = E[Y^1|D = 1, X] - E[Y^0|D = 1, X] = E[Y^1|D = 1, X] - E[Y^0|D = 0, X].$$

As in the randomised trial, the expected counterfactual outcome, $E[Y^0|D = 1, X]$ can be replaced by the expected observed outcome of the non-participants, $E[Y^0|D = 0, X]$, but only conditional on a set of observable covariates, $X$. If treated and non-treated differ in terms of observable characteristics $X$ only and not in terms of unobservables (the so-called selection on observables assumption), it holds that $E[Y^0|D = 1, X] = E[Y^0|D = 0, X]$. Otherwise, selection bias will remain an issue.

Instead of matching on $X$, it suffices to match on the propensity score $p(X)$, i.e., the probability of treatment defined as $p(X) = Pr(D = 1|X)$ (Rosenbaum and Rubin 1983). We can therefore also write

$$\text{ATT} = E[Y^1|D = 1, p(X)] - E[Y^0|D = 1, p(X)] = E[Y^1|D = 1, p(X)] - E[Y^0|D = 0, p(X)].$$

It is important to include all variables in $X$ that affect the outcome and selection into the programme at the same time. Another requirement is that the $X$ variables need to be unaffected by the treatment, i.e., they should be measured before treatment starts. PSM requires both a thorough understanding of the selection process and a large data basis. Qualitative interviews with local project managers and participants may be helpful to determine which variables to collect in order to ensure that all important variables are included in $X$.

Panel data, if available, would allow testing for the plausibility of the underlying assumptions by conducting a pseudo treatment test. The idea is to pretend that the participants received the treatment before the start of the intervention and then to estimate the impact. Because the intervention had not been in place yet, the estimated effect should be zero. If the estimation leads to a different finding, then this should be taken as evidence that participants and non-participants differ with respect to unobserved characteristics. If we are willing to assume
that these differences are time-invariant, then we can use a DID matching approach. If, however, we suspect that these differences change over time, then we need more or better \( X \) variables or a better understanding of the selection process.

Propensity score matching gives rather low internal validity due to its reliance on the selection on observables assumption. In other words, the results might be biased if there are variables that are correlated with insurance take-up and the outcome of interest (such as hospital visits), but cannot be observed in the data. External validity can be high, except in the case that we cannot find sufficiently comparable untreated individuals to be matched with every treated individual (the so-called common support requirement). These treated individuals would then need to be excluded from the analysis, which would reduce external validity.

6.6. Difference-in-differences (DID)

Relying on the assumption that selection is on observables only can be difficult to justify. Often we need a method that can also take care of confounding variables that are unobserved. However, as already mentioned, good instruments are hard to find. Therefore, we would like have other tools to deal with unobservables. The DID estimator uses data with a time or cohort dimension to control for unobserved but time-invariant variables. It relies on comparing participants and non-participants before and after the treatment. The minimum requirement is to have data on an outcome variable, \( Y \), and treatment status, \( D \), before and after the intervention. (It can be carried out with or without panel data and with or without controlling for characteristics, \( X \).) In its simplest form, we take the difference in \( Y \) for the participants before and after the treatment and subtract the difference in \( Y \) for the non-participants before and after the treatment. As a result, time-invariant differences in characteristics between participants and non-participants are eliminated, allowing us to identify the treatment effect. Consequently, this approach accounts for unobserved heterogeneity as long as it is time-invariant.
Panel data are helpful but not strictly required. Having cross-sectional data before and after the treatment may suffice. For instance, if villages participate in the intervention entirely, whilst other villages do not participate, it will suffice to conduct representative surveys in the villages before and after the intervention, i.e., interviewing the same individuals in the villages is not required. Thus, this method allows us to avoid problems with attrition commonly found in panel surveys.

The simple DID approach eliminates time-invariant heterogeneity. However, it fails to account for systematic differences in time trends between participants and non-participants. Therefore, we should include additional control variables, $X$, if we can argue that time trends are the same at least for treated and non-treated with the same $X$. This can for example be done with PSM. Another extension is to use additional differences of unaffected comparison

---

**Box 5: How the DID estimator accounts for time-invariant unobservables**

To make ideas more precise, suppose we have

$$Y_i = \alpha + \beta D_i + c + \epsilon_i,$$

where $c$ is a time-invariant variable.

With observations at time $t=0$ and $t=1$ we can take the first difference

$$Y_{i1} - Y_{i0} = \alpha + \beta (D_{i1} - D_{i0}) + (c - c) + \epsilon_i.$$

Importantly, the time-invariant characteristic $c$ drops out. As discussed in box 3, using just one cross-section of data will lead to a biased treatment effect $\beta$ if $c$ is correlated with $\epsilon_i$. By using the DID approach we get rid of the problematic unobservable.
Non-experimental methodologies for quantitative analysis

groups. For instance, imagine an insurance product applies only to individuals below the age of 40. We can then compare the time trend of individuals above the age of 40 in the treatment villages with the time trend of those above 40 in the control villages. This difference in time trends can be used to eliminate differences in time trends of those under 40. A further possibility is to use data for more than one point in time before the treatment is introduced. This would also allow eliminating differences in time trends. Having more than one survey after the treatment implementation additionally allows the estimation of time-varying and long-run treatment effects.

In order to apply the DID estimator to our hypothetical example, we need data at two points in time for two different villages: one village where insurance is available and another one where it is not. The treatment effect is then calculated as the difference in hospital visits between the two villages after the introduction of our insurance in village 1 ($t=1$) and the difference between the two villages before insurance was introduced ($t=0$). Two scenarios show how this method relies on the assumption that time-trends of the two villages are equal. In figure 6, the two villages have parallel time-trends. This means that the counterfactual of the village with insurance (red dashed line) changes over time in the same way as the control village (the green line). In this scenario the treatment effect is $(3.5-2)-(1.5-1)=1$. We receive a reliable result for our treatment effect.

However, the second scenario shows that if the time-trends are different the DID estimator does not produce reliable results: $(3.5-1.5)-(1.5-1)=1.5$. Here, our estimated treatment effect overestimates the true treatment effect by 0.5 hospital visits because the treatment and control villages do not have parallel time-trends. Reasons for the difference in time-trends can be, for example, macroeconomic effects that affect treatment and control villages differently or any other confounding factors that influence the number of hospital visits in one village but not in the other.

Figure 6: Difference-in-differences with parallel time-trends

- **1) Blue = with microinsurance scheme**
- **2) Red = without microinsurance scheme (counterfactual)**
- **3) Green = control group**

<table>
<thead>
<tr>
<th># of hospital visits</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>t=0</th>
<th>t=1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>3.5</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

(Selected points for demonstration purposes)
Internal validity of the DID approach hinges on the assumption that participants’ and non-participants’ outcome variables under consideration have the same time trend. As explained, there are tests to check for the plausibility of this assumption. In addition to the common trend assumption, we require that there are no spill-over effects from the participants to the non-participants. If these assumptions are fulfilled, internal validity is high. External validity is high, as long as the sample our data is based on is representative for the population of interest.

6.7. Fields of application of non-experimental methods

Non-experimental quantitative impact evaluation can be applied to many areas. However, the methods described belong to the field of microeconometrics, which are suited to evaluate interventions on the micro- or meso-level. A central element is that there exist different units—individuals, firms, hospitals, water works, villages, local administrations, neighbourhoods, districts, etc.—some of which were exposed by the treatment, whilst others were not. In order to evaluate projects on the macro level, such as budget support for balancing the national budget, other econometric methods are more suitable.

Convincing evaluations based on non-experimental methods require a detailed understanding of the selection mechanism and comprehensive and representative data on the treatment and comparison groups. Moreover, we usually need more than 1000 observations in order to obtain sufficiently precise estimates of the impact. Non-experimental methods are not suitable as a monitoring instrument for projects in the phase of introduction and should only be applied after resolving initial problems. Apart from initial obstacles, there are often larger modifications of the originally planned intervention, making a precise definition of the treatment more difficult.

Quantitative methodologies can be used to evaluate the impact of an intervention compared to a situation without. It is more informative, however, to
evaluate the impact of an intervention relative to other interventions or, alternatively, to evaluate different variants of an intervention, keeping context and data collection procedure constant. This, for example, would allow to look at the impact of different incentives or cost sharing arrangements for subsidized insurance.

Non-experimental methods generally give less convincing results than experimental methods. Moreover, if the confidence intervals turn out to be very wide, we should not interpret these non-significant results as evidence for the absence of an impact. This interpretation is only valid if the confidence intervals are very narrow. The correct interpretation would be that the sample size was too small to draw reliable conclusions.

References


