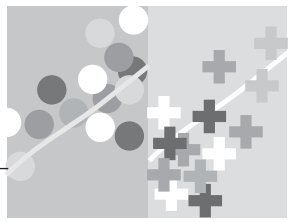


CHAPTER 6

Difference-in-Differences

Key Concept:
Difference-in-differences estimates the counterfactual for the change in outcome for the treatment group by calculating the change in outcome for the comparison group. This method allows us to take into account any differences between the treatment and comparison groups that are constant over time.

The three impact evaluation methods discussed up to this point—*randomized assignment, randomized promotion,* and *regression discontinuity design* (RDD)—all produce estimates of the counterfactual through explicit program assignment rules that the evaluator knows and understands. We have discussed why these methods offer credible estimates of the counterfactual with relatively few assumptions and conditions. The next two types of methods—*difference-in-differences* (DD) and *matching methods*—offer the evaluator an additional set of tools that can be applied in situations in which the program assignment rules are less clear or in which none of the three methods previously described is feasible. As we will see, both DD and matching methods can be powerful statistical tools; many times they will be used together or in conjunction with other impact evaluation methods.

Both difference-in-differences and matching are commonly used; however, both also typically require stronger assumptions than randomized selection methods. We also stress at the outset that both of these methods absolutely require the existence of baseline data.[1]

The difference-in-differences method does what its name suggests. It compares the *changes* in outcomes over time between a population that is enrolled in a program (the treatment group) and a population that is not (the comparison group). Take, for example, a road construction program that cannot be randomly assigned and is not assigned based on an index with a clearly defined cutoff that would permit an RDD. One of the program's objectives is to improve access to labor markets, with one of the outcome

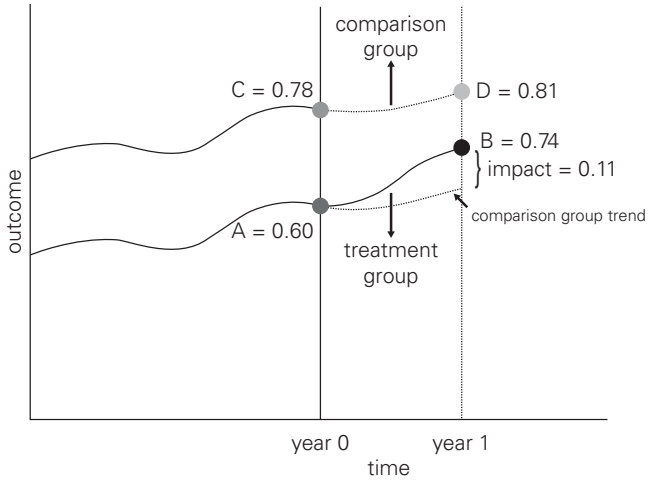
indicators being employment. As we saw in chapter 3, simply observing the before-and-after change in employment rates for areas affected by the program will not give us the program's causal impact because many other factors are also likely to influence employment over time. At the same time, comparing areas that received and did not receive the roads program will be problematic if unobserved reasons exist for why some areas received the program and others did not (the selection bias problem discussed in the enrolled–versus–not-enrolled scenario).

However, what if we combined the two methods and compared the before-and-after changes in outcomes for a group that enrolled in the program to the before-and-after changes for a group that did not enroll in the program? The difference in the before-and-after outcomes for the enrolled group—the first difference—*controls* for factors that are constant over time in that group, since we are comparing the same group to itself. But we are still left with the outside time-varying factors. One way to capture those time-varying factors is to measure the before-and-after change in outcomes for a group that did *not* enroll in the program but was exposed to the same set of environmental conditions—the second difference. If we "clean" the first difference of other time-varying factors that affect the outcome of interest by subtracting the second difference, then we have eliminated the main source of bias that worried us in the simple before-and-after comparisons. The difference-in-differences approach thus combines the two counterfeit counterfactuals (before-and-after comparisons and comparisons between those who choose to enroll and those who choose not to enroll) to produce a better estimate of the counterfactual. In our roads case, the DD method might compare the change in employment before and after the program is implemented for individuals living in areas affected by the road construction program to changes in employment in areas where the roads program was not implemented.

It is important to note that the counterfactual being estimated here is the *change* in outcomes for the comparison group. The treatment and comparison groups do not necessarily need to have the same preintervention conditions. But for DD to be valid, the comparison group must accurately represent the change in outcomes that would have been experienced by the treatment group in the absence of treatment. To apply difference-in-differences, all that is necessary is to measure outcomes in the group that receives the program (the treatment group) and the group that does not (the comparison group) both before and after the program. The method does not require us to specify the rules by which the treatment is assigned.

Figure 6.1 illustrates the difference-in-differences method. A treatment group is enrolled in a program, and a comparison group is not enrolled. The

Figure 6.1 Difference-in-Differences



Source: Authors.

before-and-after outcome variables for the treatment group are *A* and *B,* respectively, while the outcome for the comparison group goes from *C,* before the program, to *D* after the program has been implemented.

You will remember our two counterfeit counterfactuals—the difference in outcomes before and after the intervention for the treatment group *(B − A)* and the difference in outcomes[2] after the intervention between the treatment and comparison groups *(B − D)*. In difference-in-differences, the estimate of the counterfactual is obtained by computing the change in outcomes for the comparison group *(D − C)*. This counterfactual change is then subtracted from the change in outcomes for the treatment group *(B - A)*.

In summary, the impact of the program is simply computed as the difference between two differences:

DD impact = *(B − A) − (D − C) = (B − E)* = (0.74 − 0.60) − (0.81 − 0.78) = 0.11.

The relationships presented in figure 6.1 can also be presented in a simple table. Table 6.1 disentangles the components of the difference-in-differences estimates. The first row contains outcomes for the treatment group before *(A)* and after *(B)* the intervention. The before-and-after comparison for the treatment group is the first difference *(B − A)*. The second row contains outcomes for the comparison group before the intervention *(C)* and after the intervention *(D)*, so the second (counterfactual) difference is *(D − C)*.

Table 6.1 The Difference-in-Differences Method

	After	Before	Difference
Treatment/enrolled	*B*	*A*	*B – A*
Comparison/ nonenrolled	*D*	*C*	*D – C*
Difference	*B – D*	*A – C*	*DD = (B – A) – (D – C)*

	After	Before	Difference
Treatment enrolled	0.74	0.60	*0.14*
Comparison/ nonenrolled	0.81	0.78	*0.03*
Difference	*–0.07*	*–0.18*	*DD = 0.14 – 0.03 = 0.11*

Source: Authors.

The difference-in-differences method computes the impact estimate as follows:

1. We calculate the difference in the outcome *(Y)* between the before and after situations for the treatment group *(B − A)*.
2. We calculate the difference in the outcome *(Y)* between the before and after situations for the comparison group *(D − C)*.
3. Then we calculate the difference between the difference in outcomes for the treatment group *(B − A)* and the difference for the comparison group *(D − C)*, or DD = *(B − A)* − *(D − C)*. This "difference-in-differences" is our impact estimate.

How Is the Difference-in-Differences Method Helpful?

To understand how difference-in-differences is helpful, let us start with our second counterfeit counterfactual, which compared units that were enrolled in a program with those that were not enrolled in the program. Remember that the primary concern with this was that the two sets of units may have had different characteristics and that it may be those characteristics rather than the program that explain the difference in outcomes between the two groups. The *unobserved* differences in characteristics were particularly worrying: by definition, it is impossible for us to include unobserved differences in characteristics in the analysis.

The difference-in-differences method helps resolve this problem to the extent that many characteristics of units or individuals can reasonably be assumed to be constant over time (or *time-invariant*). Think, for example, of *observed* characteristics, such as a person's year of birth, a region's location close to the ocean, a town's level of economic development, or a father's level of education. Most of these types of variables, although plausibly related to outcomes, will probably not change over the course of an evaluation. Using the same reasoning, we might conclude that many *unobserved* characteristics of individuals are also more or less constant over time. Consider, for example, a person's intelligence or such personality traits as motivation, optimism, self-discipline, or family health history. It is plausible that many of these intrinsic characteristics of a person would not change over time.

When the same individual is observed before and after a program and we compute a simple difference in outcome for that individual, we cancel out the effect of all of the characteristics that are unique to that individual and that do not change over time. Interestingly, we are canceling out (or controlling for) not only the effect of *observed* time-invariant characteristics but also the effect of *unobserved* time-invariant characteristics such as those mentioned above.

The "Equal Trends" Assumption in Difference-in-Differences

Although difference-in-differences allows us to take care of differences between the treatment and the comparison group that are constant over time, it will not help us eliminate the differences between the treatment and comparison groups that change over time. In the roads example above, if treatment areas also benefit from the construction of a new seaport at the same time as the road construction, we will not be able to account for the seaport construction by using a difference-in-differences approach. For the method to provide a valid estimate of the counterfactual, we must assume that no such time-varying differences exist between the treatment and comparison groups.

Another way to think about this is that in the absence of the program, the differences in outcomes between the treatment and comparison groups would need to move in tandem. That is, without treatment, outcomes would need to increase or decrease at the same rate in both groups; we require that outcomes display *equal trends in the absence of treatment*.

Unfortunately, there is no way for us to prove that the differences between the treatment and comparison groups would have moved in tandem in the absence of the program. The reason is that we cannot observe what would

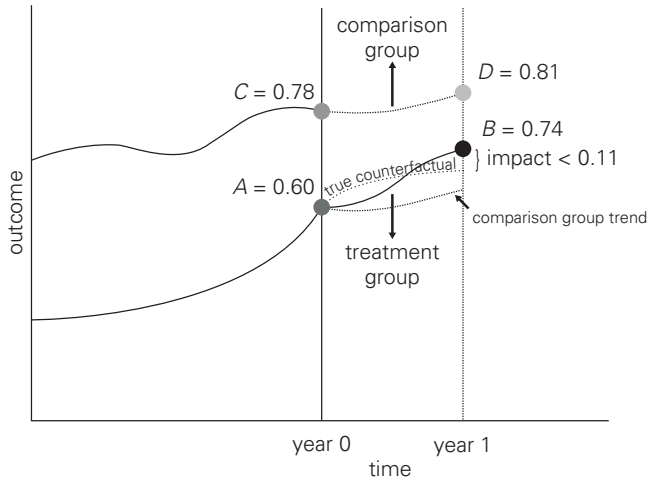
have happened to the treatment group in the absence of the treatment—in other words, we cannot observe the counterfactual!

Thus, when we use the difference-in-differences method, we must *assume* that, in the absence of the program, the outcome in the treatment group would have moved in tandem with the outcome in the comparison group. Figure 6.2 illustrates a violation of this fundamental assumption, which is needed for the difference-in-differences method to produce credible impact estimates. If outcome trends are different for the treatment and comparison groups, then the estimated treatment effect obtained by difference-in-difference methods would be invalid, or biased. The reason is that the trend for the comparison group is not a valid estimate of the counterfactual trend that would have prevailed for the treatment group in the absence of the program. As we see in figure 6.2, outcomes for the comparison group grow faster than outcomes for the treatment group in the absence of the program, so using the trend for the comparison group as a counterfactual for the trend for the treatment group leads to an underestimation of the program's impact.

Testing the Validity of the "Equal Trends" Assumption in Difference-in-Differences

The validity of the underlying assumption of equal trends can be assessed even though it cannot be proved. A good validity check is to compare

Figure 6.2 Difference-in-Differences when Outcome Trends Differ



Source: Authors.

changes in outcomes for the treatment and comparison groups *before* the program is implemented. If the outcomes moved in tandem before the program started, we gain confidence that outcomes would have continued to move in tandem in the postintervention period. To check for equality of preintervention trends, we need at least two serial observations on the treatment and comparison groups before the start of the program. This means that the evaluation would require three serial observations—two preintervention observations to assess the preprogram trends and at least one postintervention observation to assess impact with the difference-in-differences formula.

A second way to test the assumption of equal trends would be to perform what is known as a "placebo" test. For this test, you perform an additional difference-in-differences estimation using a "fake" treatment group, that is, a group that you know was not affected by the program. Say, for example, that you estimate how additional tutoring for grade 7 students affects their probability of attending school, and you choose grade 8 students as the comparison group. To test whether seventh and eighth graders have the same trends in terms of school attendance, you could test whether eighth graders and sixth graders have the same trends. You know that sixth graders are not affected by the program, so if you perform a difference-in-differences estimation using grade 8 students as the comparison group and grade 6 students as the fake treatment group, you *have to* find a zero impact. If you do not, then the impact that you find must come from some underlying difference in trends between sixth graders and eighth graders. This, in turn, casts doubt on whether seventh graders and eighth graders can be assumed to have parallel trends in the absence of the program.

A placebo test can be performed not only with a fake treatment group but also with a fake outcome. In the tutoring example, you may want to test the validity of using the grade 8 students as a comparison group by estimating the impact of the tutoring on an outcome that you know is not affected by it, such as the number of siblings that the students have. If your difference-in-differences estimation finds an "impact" of the tutoring on the number of siblings that the students have, then you know that your comparison group must be flawed.

A fourth way to test the assumption of parallel trends would be to perform the difference-in-differences estimation using different comparison groups. In the tutoring example, you would first do the estimation using grade 8 students as the comparison group, and then do a second estimation using grade 6 students as the comparison group. If both groups are valid comparison groups, you would find that the estimated impact is approximately the same in both calculations.

Using Difference-in-Differences to Evaluate the Health Insurance Subsidy Program

Difference-in-differences can be used to evaluate our health insurance subsidy program (HISP). In this scenario, you have two rounds of data on two groups of households, one group that enrolled in the program and another that did not. Remembering the case of the selected enrolled and nonenrolled groups, you realize that you cannot simply compare the average health expenditures of the two groups because of selection bias. Because you have data for two periods for each household in the sample, you can use those data to solve some of these challenges by comparing the change in expenditures for the two groups, assuming that the change in the health expenditures of the nonenrolled group reflects what would have happened to the expenditures of the enrolled group in the absence of the program (see table 6.2). Note that it does not matter which way you calculate the double difference.

Next, you estimate the effect using regression analysis (table 6.3). Using a simple linear regression, you find that the program reduced household health expenditures by $7.8. You then refine your analysis by using multivariate linear regression to take into account a host of other factors, and you find the same reduction in household health expenditures.

QUESTION 6

A. What are the basic assumptions required to accept this result from case 6?
B. Based on the result from case 6, should the HISP be scaled up nationally?

Table 6.2 Case 6—HISP Impact Using Difference-in-Differences (Comparison of Means)

	After (follow-up)	Before (baseline)	Difference
Enrolled	7.8	14.4	−6.6
Nonenrolled	21.8	20.6	1.2
Difference			DD = −6.6 − 1.2 = −7.8

Source: Authors.

Table 6.3 Case 6—HISP Impact Using Difference-in-Differences (Regression Analysis)

	Linear regression	Multivariate linear regression
Estimated impact on household health expenditures	−7.8**	−7.8**
	(0.33)	(0.33)

Source: Authors.

Note: Standard errors are in parentheses.

** Significant at the 1 percent level.

The Difference-in-Differences Method at Work

Despite its limitations, the difference-in-differences method remains one of the most frequently used impact evaluation methodologies, and many examples appear in the literature. For example, Duflo (2001) analyzed the schooling and labor market impacts of school construction in Indonesia. DiTella and Schargrodsky (2005) examined whether an increase in police forces reduces crime. Another key example from the literature is described in box 6.1.

Box 6.1: Water Privatization and Infant Mortality in Argentina

Galiani, Gertler, and Schargrodsky (2005) used the difference-in-differences method to address an important policy question: whether privatizing the provision of water services can improve health outcomes and help alleviate poverty. During the 1990s, Argentina initiated one of the largest privatization campaigns ever, transferring local water companies to regulated private companies covering about 30 percent of the country's municipalities and 60 percent of the population. The privatization process took place over a decade, with the largest number of privatizations occurring after 1995.

Galiani, Gertler, and Schargrodsky (2005) took advantage of that variation in ownership status over time to determine the impact of privatization on under-age-5 mortality. Before 1995, the rates of child mortality were declining at about the same pace throughout Argentina; after 1995, mortality rates declined faster in municipalities that had privatized their water services. The researchers argue that, in this context, the identification assumptions behind difference-in-differences are likely to hold true. First, they show that the decision to privatize was uncorrelated with economic shocks or historical levels of child mortality. Second, they show that no differences in child mortality trends are observed between the comparison and treatment municipalities before the privatization movement began.

They checked the strength of their findings by decomposing the effect of privatization on child mortality by cause of death and found that the privatization of water services is correlated with reductions in deaths from infectious and parasitic diseases but not from causes unrelated to water conditions, such as accidents or congenital diseases. In the end, the evaluation determined that child mortality fell about 8 percent in areas that privatized and that the effect was largest, about 26 percent, in the poorest areas, where the expansion of the water network was the greatest. This study shed light on a number of important policy debates surrounding the privatization of public services. The researchers concluded that in Argentina, the regulated private sector proved more successful than the public sector in improving indicators of access, service, and most significantly, child mortality.

Source: Galiani, Gertler, and Schargrodsky 2005.

Limitations of the Difference-in-Differences Method

Difference-in-differences is generally less robust than the randomized selection methods (randomized assignment, randomized offering, and randomized promotion). Even when trends are parallel before the start of the intervention, bias in the estimation may still appear. The reason is that DD attributes to the intervention *any differences in trends* between the treatment and comparison groups that occur *from the time intervention begins*. If any other factors are present that affect the difference in trends between the two groups, the estimation will be invalid or biased.

Let us say that you are trying to estimate the impact on rice production of subsidizing fertilizer and are doing this by measuring the rice production of subsidized (treatment) farmers and unsubsidized (comparison) farmers before and after the distribution of the subsidies. If in year 1 the subsidized farmers are affected by drought, whereas the unsubsidized farmers are not, then the difference-in-differences estimate will produce an invalid estimate of the impact of subsidizing fertilizer. In general, any factor that affects only the treatment group, and does so at the same time that the group receives the treatment, has the potential to invalidate or bias the estimate of the impact of the program. Difference-in-differences *assumes* that no such factor is present.

Notes

1. Although randomized assignment, randomized promotion, and regression discontinuity design theoretically do not require baseline data, in practice having a baseline is very useful for confirming that the characteristics of the treatment and comparison groups are balanced. For this reason, we recommend including a baseline as part of the evaluation. In addition to verifying balance, a number of other good reasons argue for collecting baseline data, even when the method does not absolutely require them. First, having preintervention (exogenous) population characteristics can enable the evaluator to determine whether the program has a different impact on different groups of the eligible population (so-called heterogeneity analysis). Second, the baseline data can also be used to perform analysis that can guide policy even before the intervention starts, and collecting the baseline data can serve as a massive pilot for the postintervention data collection. Third, baseline data can serve as an "insurance policy" in case randomized assignment is not implemented; as a second option, the evaluator could use a combination of matching and differences-in-differences. Finally, baseline data can add statistical power to the analysis when the number of units in the treatment and comparison groups is limited.
2. All differences between points should be read as vertical differences in outcomes on the vertical axis.

References

DiTella, Rafael, and Ernesto Schargrodsky. 2005. "Do Police Reduce Crime? Estimates Using the Allocation of Police Forces after a Terrorist Attack." *American Economic Review* 94 (1): 115–33.

Duflo, Esther. 2001. "Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment." *American Economic Review* 91 (4): 795–813.

Galiani, Sebastian, Paul Gertler, and Ernesto Schargrodsky. 2005. "Water for Life: The Impact of the Privatization of Water Services on Child Mortality." *Journal of Political Economy* 113 (1): 83–120.