Difference-in-Difference

Contents

Basic Idea

Suppose we have two periods: before, t', and after, t. Individuals are divided into two groups , one with $D = 1$ (experiment group) and one with $D = 0$ (control group). At period t', no treatment happens.Everyone's outcome is denoted as $y_{0t'}$. At period t, experiment group gets treatment, so they have y_{1t}; individuals in control group do not get treatment, so they have y_{0t}

The logic flow of DID derivation is same as the treatment effect model. We want to estimate the following:

$$E(y_{1t}|x, D = 1) - E(y_{0t}|x, D = 1)$$

But $E(y_{0t}|x, D = 1)$ is not observable! To address this, we do the following algebra:

$$\begin{aligned} E(y_{1t}|x, D = 1) - E(y_{0t}|x, D = 1) &= E(y_{1t}|x, D = 1) - E(y_{0t'}|x, D = 1) + E(y_{0t'}|x, D = 1) - E(y_{0t}|x, D = 1) \\ &= (E(y_{1t}|x, D = 1) - E(y_{0t'}|x, D = 1)) - (E(y_{0t}|x, D = 1) - E(y_{0t'}|x, D = 1)) \end{aligned} \tag{1}$$

- $E(y_{1t}|x, D = 1) - E(y_{0t'}|x, D = 1)$ is the total observable change of the outcome variable (from t' to t) for individual x in the treatment group.
- $E(y_{0t}|x, D = 1) - E(y_{0t'}|x, D = 1)$ is the 'hypothetical' change of the outcome variable as if the individual x in the treatment group do not participate in the treatment. This measure the 'natural' change (i.e., the change if there is no policy treatment) of the outcome for individuals in the treatment group. unobservable!

The total observable change, minus the 'hypothetical' natural change, is the change brought by the treatment, i.e, the effect of the treatment. However, since the 'hypothetical change' is unosbservable, we need to again introduce the CIA assumption:

$$E(y_{0t} - y_{0t'}|x, D) = E(y_{0t} - y_{0t'}|x)$$

i.e., given the x, the trend of outcome (i.e., natural change of outcome without any treatment) is independent of 'whether getting treatment or not'. Having knowing this, we have

$$E(y_{0t} - y_{0t'}|x, D = 1) = E(y_{0t} - y_{0t'}|x, D = 0)$$

The right hand side is observable. Plugging this expression into (1), we have

$$E(y_{1t}|x, D = 1) - E(y_{0t}|x, D = 1) = E(y_{1t} - y_{0t'}|x, D = 1) - E(y_{0t} - y_{0t'}|x, D = 0) \tag{2}$$

Both terms in the right hand side are observable.

Paremetric Method : Linear regression.

We now present how to 'recover' the linear regression specification from (2). denote that

$$E(y_{0t'}|x, D = 0) = \alpha_1$$

i.e., α_1 is the level of y for people in the control group before the experiment happens. We next denote that

$$E(y_{0t}|x, D = 0) = \alpha_1 + \alpha_2$$

α_2 denotes the 'natural' change of y without any experiment effects (i.e., the change of y for people in the control group) Denote that

$$E(y_{0t'}|x, D = 1) = \alpha_1 + \alpha_3$$

α_3 denotes the difference of y between the control group and the experiment group before the experiment happens. Finally denotes that

$$E(y_{1t}|x, D = 1) = \alpha_1 + \alpha_3 + \alpha_2 + \alpha_4$$

we can see that the change of y for the people in the experiment group comes from two parts, α_2, the natural change of y, and α_4, the net effect of the treatment. Therefore we can see that - The α_2 appears only when the people is after the experiment. denote $G = 1(0)$ as after (before) the treatment. the α_2 is thus the coefficient of G. - The α_3 appears only when the people is at experiment group. Denote $D = 1(0)$ as in (not in) the treatment group. The α_3 is thus the coefficient of D. - The α_4 appears only the people is at experiment group, and the is after experiment. The α_4 is thus the coefficient of $G * D$.

Finally the model can be written as

$$y = \alpha_1 + \alpha_2 G + \alpha_3 D + \alpha_4 G * D + \gamma X$$

Non-Parametric Method: PSM

The algorithm for doing Non-parametric estimation is super simple:

- For each x value (in the $D = 1$ group):
  - estimate $E(y_{1t} - \widehat{y_{0t'}|x}, D = 1) \equiv \frac{\sum_{D_i=1, x_i=x}(y_{1ti} - y_{0t'i})}{\sum \left[1_{(D_i=1, x_i=x)}\right]}$
  - estimate $E(y_{0t} - \widehat{y_{0t'}|x}, D = 0) \equiv \sum_{D_j=0} w(x_j, x)\,(y_{0tj} - y_{0t'j})$, in which $w(x_j, x)$ is weight, which is larger when x_j is closer to x.

The estimated ATT is $E(y_{1t} - \widehat{y_{0t'}|}D = 1) = \sum_{D_i=1} \left[E(y_{1t} - \widehat{y_{0t'}|x}, D = 1)p(x_i)\right]$

Difference-in-Difference-in-Difference Method

DID requires that the the control group and experiment group have the same time trend (common natural change of target variable.).Sometimes, however, this may not be true, especially when the control group and the experiment group are conceptionally different. To illustrate this, suppose that at year t, a province A introduced a health policy targeting on ALL the elder people (over 65). Suppose we have the the health data for all-age people at t_1 and t_2, with $t_1 < t < t_2$. We want to evaluate the effect of such policy on elder people's heath.Following the basic idea, we want to see

$$E(y^1_{A,old,t_2}|x, old) - E(y^0_{A,old,t_2}|x, old)$$

The second is unobservable. Following (2),A (bad) approach is to check

$$E(y^1_{A,old,t_2} - y^0_{A,old,t_1}|x, old) - E(y^0_{A,young,t_2} - y^0_{A,young,t_1}|x, young)$$

This approach relies on the CIA assumption:

$$E(y^0_{A,young,t_2} - y^0_{A,young,t_1}|x, young) = E(y^0_{A,old,t_2} - y^0_{A,old,t_1}|x, old)$$

This is unrealistic: elder people and young people apparently have different trends of health status!

A another (not so bad) approach is to check

$$E(y^1_{A,old,t_2} - y^0_{A,old,t_1}|x, old) - E(y^0_{B,old,t_2} - y^0_{B,old,t_1}|x, old)$$

where $E(y^0_{B,old,t_2} - y^0_{B,old,t_1}|x, old)$ is the trend of change for the elder people at province B, where there is no such policy.This relies on the CIA assumption:

$$E(y^0_{B,old,t_2} - y^0_{B,old,t_1}|x, old) = E(y^0_{A,old,t_2} - y^0_{A,old,t_1}|x, old)$$

This is less unrealistic, but still skeptical: elder people in A may have different natural trend in health status from elder people in B.

To address this, we can impose the following assumption:

$$E(y^0_{A,young,t_2} - y^0_{A,young,t_1}|x, young) - E(y^0_{A,old,t_2} - y^0_{A,old,t_1}|x, old)$$

$$= E(y^0_{B,young,t_2} - y^0_{B,young,t_1}|x, young) - E(y^0_{B,old,t_2} - y^0_{B,old,t_1}|x, old)$$

That is, the difference in the natural trend of health status between elder and young people in province A and B are the same.Given this assumption, we have

$$E(y^1_{A,old,t_2}|x, old) - E(y^0_{A,old,t_2}|x, old)$$

$$= E(y^1_{A,old,t_2} - y^0_{A,old,t_1}|x, old) - E(y^0_{A,old,t_2} - y^0_{A,old,t_1}|x, old)$$
$$= E(y^1_{A,old,t_2} - y^0_{A,old,t_1}|x, old) + E(y^0_{B,young,t_2} - y^0_{B,young,t_1}|x, young)$$
$$-E(y^0_{B,old,t_2} - y^0_{B,old,t_1}|x, old) - E(y^0_{A,young,t_2} - y^0_{A,young,t_1}|x, young)$$

Now we construct a linear model based on this framework.

- Denote the $E(y^0_{B,young,t_1}|x, young)$ as α_1
- Denote the $E(y^0_{B,young,t_2}|x, young)$ as $\alpha_1 + \alpha_2$. α_2 is natural trend for young people at province B.
- Denote the $E(y^0_{B,old,t_1}|x, old)$ as $\alpha_1 + \alpha_3$, α_3 is the difference of the ex-ante heath status between elder and young at B.
- Denote the $E(y^0_{B,old,t_2}|x, old)$ as $\alpha_1 + \alpha_3 + \alpha_2 + \alpha_4$. α_4 is the difference between natural trend for young people and the natural trend for elder people at B.
- Denote the $E(y^0_{A,young,t_1}|x, young)$ as $\alpha_1 + \alpha_5$. α_5 is the across-province difference.
- Denote the $E(y^0_{A,young,t_2}|x, young)$ as $\alpha_1 + \alpha_5 + \alpha_2 + \alpha_6$. α_6 measures the difference between the natural trend of young people in A and that of young people in B.
- Denote the $E(y^0_{A,old,t_1}|x, old)$ as $\alpha_1 + \alpha_3 + \alpha_5 + \alpha_7$. α_7 indicates the difference between the ex-ante health status between elder and young at A.
- Denote the $E(y^0_{A,old,t_2}|x, old)$ as $\alpha_1 + \alpha_3 + \alpha_5 + \alpha_2 + \alpha_6 + \alpha_7 + \alpha_4$.
- Denote the $E(y^1_{A,old,t_2}|x, old)$ as $\alpha_1 + \alpha_3 + \alpha_5 + \alpha_2 + \alpha_6 + \alpha_7 + \alpha_4 + \alpha_8$. α_8 is the true effect of the policy

Now it is easy to get: - α_2 appears whenever at t_2.Denote $G = 1(0)$ as being in $t_2(t_1)$, then α_2 is the coefficient of G. - α_3 appears whenever old. Denote $O = 1(0)$ as being old (young), then α_3 is the coefficient of O. - α_4 appears whenever old at t_2. then α_4 is the coefficient of $G * O$ - α_5 appears whenever at province A. Denote $A = 1(0)$ as being in state A (B). Then α_5 is the coefficient of A. - α_6 appears whenever at province A and t_2. then α_6 is the coefficient of $A * G$. - α_7 appears whenever at province A and old. Then α_7 is the coefficient of $A * O$. - α_8 appears whenever at province A and old and t_2.Then α_8 is the the coefficient of $A * O * G$.

There we can write the linear regression model based on the above analysis.