

Program Evaluation and Difference in Difference

December 4, 2014

Program Evaluation

Imagine observing subjects over time and at some point in time some of the subjects receive an intervention. Suppose we want to examine the impact of that intervention on our dependent variable.

Example

In 1981 an incinerator was announced and built in the town of North Andover Massachusetts. We have data on homes sold in 1978 (pre incinerator) and 1981 (post incinerator). Can we estimate the impact of the incinerator on mean home prices near the incinerator in the town. Complications:

- Home prices in the area around the incinerator may have been different for reasons independent of the incinerator.
- Home prices may be changing due to market fluctuations from 1978 - 1981

The Data is Pure Cross Sectional

home_id	year	price	nearinc	y81
1	1978	\$20,000	1	0
2	1978	\$21,000	0	0
3	1981	\$18,000	1	1
4	1981	\$28,000	0	1

Difference in Difference Estimator for the Effect of Incinerator

The estimator is defined as

$$E[\Delta y_{treatment}] - E[\Delta y_{control}] \quad (1)$$

Note: we will be using standard regression approaches for estimating models. The above describes how we use regression estimates to then find the effect of the incinerator.

Terminology

Think about control and treatment groups defined by proximity to the incinerator I:

	1978	1981
Control	Homes Far from I	Homes Far from I
Treatment	Homes Near I	Homes Near I

Basic strategy:

- Difference between control and treatment in 1978 tells us about price differences independent of the incinerator
- Using differences between 1978 and 1981 on all properties, we can see how prices are changing due to fluctuations in the housing market
- Combine these to find the isolated effect from the incinerator

Estimation Strategy 1: Independent Equations for Years 1978 and 1981 and then combine after estimation

Simply pooling all these observations and estimating **price** = $\beta_0 + \beta_1 \text{NEARINC} + \epsilon$, probably won't get us what we want.

- Will probably tend to overstate the effect of the incinerator if the houses near the incinerator were low value to begin with.
- Will probably tend to understate the effect of the incinerator if house prices overall were trending higher

Estimate these two models¹:

$$\text{price}^{78} = \beta_0^{78} + \beta_1^{78} \text{NEARINC} + \epsilon^{78} \quad (2)$$

$$\text{price}^{81} = \beta_0^{81} + \beta_1^{81} \text{NEARINC} + \epsilon^{81} \quad (3)$$

¹This presentations focuses on the treatment variable only and surpresses other independent variables for convenience.

Put together for the Difference in Difference Estimator

	1978	1981	Difference
Control	β_0^{78}	β_0^{81}	$\beta_0^{81} - \beta_0^{78}$
Treatment	$\beta_0^{78} + \beta_1^{78}$	$\beta_0^{81} + \beta_1^{81}$	$\beta_0^{81} + \beta_1^{81} - (\beta_0^{78} + \beta_1^{78})$
Difference	β_1^{78}	β_1^{81}	$\beta_1^{81} - \beta_1^{78}$

So the estimated impact would be the so-called difference in a difference:

$$\beta_1^{81} - \beta_1^{78}$$

Estimation Strategy 2: Pool data and define covariates appropriately:

$$price_i = \beta_0 + \delta_0 y81_i + \beta_1 NEARINC_i + \quad (4)$$

$$\delta_1 y81_i \times NEARINC_i + \epsilon_i \quad (5)$$

It can be shown fairly easily by creating a table like the previous slide that the Difference in Difference estimator is δ_1 . This is the most common method for Diff-n-diff estimation.

Diff-in-Diff in a Panel Context

Suppose rather than a cross section, we have a panel of home sales prices collected in 1978 and then again for the same homes in 1981. Some of these homes are near the incinerator and others are not. The data for the first two observations might look like this:

home_id	year	price	nearinc	y81
1	1978	\$20,000	1	0
1	1981	\$21,000	1	1
2	1978	\$18,000	0	0
2	1981	\$28,000	0	1

Treatment and Control in this Context

Now we have an even better treatment and control strategy: we can observe the behavior of *each* cross section unit before (control) and after (treatment) the incinerator. Let the population regression function (with unobserved heterogeneity) be:

$$P_{it} = \beta_0 + \delta_0 y_{81it} + \beta_1 NEARINC_{it} + \delta_1 y_{81it} \times NEARINC_{it} + c_i + \epsilon_{it} \quad (6)$$

	1978	1981	Difference
Control	$\beta_0 + c_i$	$\beta_0 + c_i + \delta_0$	δ_0
Treatment	$\beta_0 + c_i + \beta_1$	$\beta_0 + c_i + \beta_1 + \delta_0 + \delta_1$	$\delta_0 + \delta_1$
Difference	β_1	$\beta_1 + \delta_1$	δ_1

Consider in a Fixed Effects Context

$$P_{it} = \beta_0 + \delta_0 y81_{it} + \beta_1 NEARINC_i + \delta_1 y81_{it} \times NEARINC_{it} + c_i + \epsilon_{it} \quad (7)$$

Now consider the first difference:

$$\Delta P_i = \delta_0 \Delta y81_i + \delta_1 \Delta y81_i \times NEARINC_{i,1981} + \Delta \epsilon_i \quad (8)$$

Note: the parameters in the preceding table that do not vary over t can't be identified with the fixed effects regression.