Difference-in-Differences

Alberto Abadie
John F. Kennedy School of Government
Harvard University

IDB
Course on Synthetic Controls
December 3-4, 2015

Motivating Example: The Mariel Boatlift

- How do inflows of immigrants affect the wages and employment of natives in local labor markets?
- Card (1990) uses the Mariel Boatlift of 1980 as a natural experiment to measure the effect of a sudden influx of immigrants on unemployment among less-skilled natives
- The Mariel Boatlift increased the Miami labor force by 7%
- Individual-level data on unemployment from the Current Population Survey (CPS) for Miami and four comparison cities (Atlanta, Los Angeles, Houston and Tampa-St. Petersburg)

Motivating Example: The Mariel Boatlift

Difference-in-differences estimate on unemployment rates
(African-American workers)

	Year		
	1979	1981	1981−1979
Miami	8.3 (1.7)	9.6 (1.8)	1.3 (2.5)
Comparison cities	10.3 (0.8)	12.6 (0.9)	2.3 (1.2)
Miami-Comparison Difference	−2.0 (1.9)	−3.0 (2.0)	−1.00 (2.8)

Adapted from Card (1990) and Angrist and Krueger (1999).
Standard errors in parentheses.

Difference-in-Differences Setup

Two groups:

- $D = 1$: treated units
- $D = 0$: control units

Two periods:

- $T = 0$: pre-treatment period
- $T = 1$: post-treatment period

Potential outcomes:

- $Y_{1i}(t)$: outcome unit i attains in period t if treated before t
- $Y_{0i}(t)$: outcome unit i attains in period t if not treated before t

Difference-in-Differences Setup

Treatment effect for unit i at time t is

$$Y_{1i}(t) - Y_{0i}(t).$$

Observed outcomes $Y_i(t)$ are realized as

$$Y_i(t) = Y_{0i}(t)(1 - D_i(t)) + Y_{1i}(t)D_i(t).$$

Because the treatment occurs only after $t = 0$, we define

$$D_i = D_i(1).$$

It follows that,

$$\begin{aligned} Y_i(0) &= Y_{0i}(0), \\ Y_i(1) &= Y_{0i}(1)(1 - D_i) + Y_{1i}(1)D_i. \end{aligned}$$

Difference-in-Differences Result

Let

$$\alpha_{ATET} = E[Y_1(1) - Y_0(1)|D = 1].$$

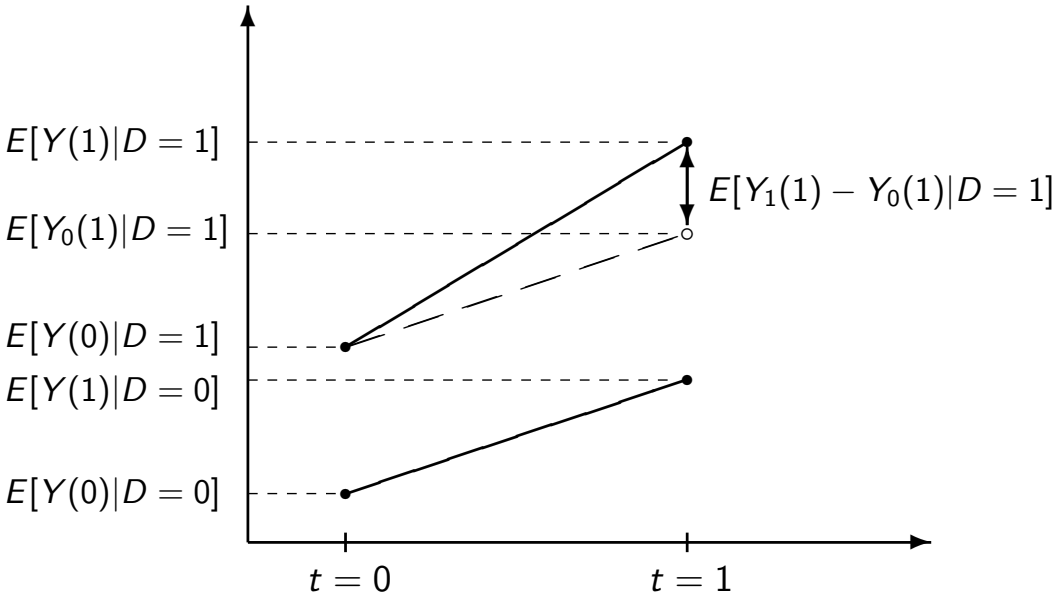
If the treated and non-treated would have exhibited the same trend in the absence of the treatment,

$$E[Y_0(1) - Y_0(0)|D = 1] = E[Y_0(1) - Y_0(0)|D = 0],$$

then:

$$\begin{aligned} \alpha_{ATET} = &\Big[E[Y(1)|D = 1] - E[Y(1)|D = 0]\Big] \\ &- \Big[E[Y(0)|D = 1] - E[Y(0)|D = 0]\Big]. \end{aligned}$$

Difference-in-Differences: Graphical Interpretation



Difference-in-Differences Estimators

1. Panel Data/Sample Means

$$\left\{\frac{1}{N_1}\sum_{D_i=1} Y_i(1) - \frac{1}{N_0}\sum_{D_i=0} Y_i(1)\right\} - \left\{\frac{1}{N_1}\sum_{D_i=1} Y_i(0) - \frac{1}{N_0}\sum_{D_i=0} Y_i(0)\right\}$$

$$= \left\{\frac{1}{N_1}\sum_{D_i=1}\{Y_i(1)-Y_i(0)\} - \frac{1}{N_0}\sum_{D_i=0}\{Y_i(1)-Y_i(0)\}\right\},$$

where N_1 is the number of treated individuals and N_0 is the number of non-treated individuals.

Difference-in-Differences Estimators

2. Repeated Cross-Sections/Sample Means

   Let $\{Y_i, D_i, T_i\}_{i=1}^N$ be the pooled sample (the two different cross-sections merged) where T is a variable that indicates the period (0 or 1) in which the individual is observed.

   A DID estimator is given by:

$$\left\{\frac{\sum D_i \cdot T_i \cdot Y_i}{\sum D_i \cdot T_i} - \frac{\sum (1-D_i) \cdot T_i \cdot Y_i}{\sum (1-D_i) \cdot T_i}\right\} - \left\{\frac{\sum D_i \cdot (1-T_i) \cdot Y_i}{\sum D_i \cdot (1-T_i)} - \frac{\sum (1-D_i) \cdot (1-T_i) \cdot Y_i}{\sum (1-D_i) \cdot (1-T_i)}\right\}.$$

Difference-in-Differences Estimators

3. Repeated Cross-Sections/Regression

   The same estimator can be obtained using regression techniques. Consider the linear model:

$$Y_i = \mu + \gamma\, D_i + \delta\, T_i + \alpha\,(D_i \times T_i) + \varepsilon_i,$$

   where $E[\varepsilon|D, T] = 0$. Then, it is easy to show that

$$\begin{aligned} \alpha = & \{E[Y|D=1, T=1] - E[Y|D=0, T=1]\} \\ - & \{E[Y|D=1, T=0] - E[Y|D=0, T=0]\}. \end{aligned}$$

Difference-in-Differences Estimators

The equation

$$E[Y|D, T] = \mu + \gamma\, D + \delta\, T + \alpha\,(D \times T)$$

implies,

$$\begin{aligned}
&E[Y|D=1, T=1] = \mu + \gamma + \delta + \alpha \\
&\qquad\qquad\qquad\qquad\qquad\qquad \ominus \quad = \gamma + \alpha \\
&E[Y|D=0, T=1] = \mu + \delta \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \ominus \quad = \alpha \\
&E[Y|D=1, T=0] = \mu + \gamma \\
&\qquad\qquad\qquad\qquad\qquad\qquad \ominus \quad = \gamma \\
&E[Y|D=0, T=0] = \mu
\end{aligned}$$

Difference-in-Differences Estimators

Consider a regression version of the DID estimator including covariates:

$$Y_i = \mu + \gamma D_i + \delta T_i + \alpha D_i T_i + X_i'\beta + \varepsilon_i.$$

- introducing time-invariant X's in this way is not helpful (they get differenced-out)
- time-varying X's may be problematic because they are often affected by the treatment and may introduce endogeneity

More sensible: Interact time-invariant covariates with the time indicator:

$$Y_i = \mu + \gamma D_i + \delta T_i + \alpha D_i T_i + T_i X_i'\beta_1 + (1 - T_i) X_i'\beta_0 + \varepsilon_i$$

$\Rightarrow$ X is used to explain differences in trends.

Difference-in-Differences Estimators

4. Panel Data/Regression

   With panel data we can use regression with the dependent variable in first differences:

$$\Delta Y_i = \delta + \alpha \cdot D_i + X_i'\beta + u_i,$$

   where $\Delta Y_i = Y_i(1) - Y_i(0)$, $\beta = \beta_1 - \beta_0$, and $u_i = \Delta\varepsilon_i$.

   $\Rightarrow$ Here X also explains differences in trends.

Difference-in-Differences Estimators

The fixed effects estimator generalizes DID in the context of panel data and multiple groups and time periods:

$$Y_{it} = \mu + \gamma_i + \delta_t + \alpha D_{it} + X_{it}'\beta + \varepsilon_{it}$$

- One intercept for each unit, γ_i, and time period, δ_t
- Cluster standard errors at the subject unit level
- Can add unit specific time trends:

$$Y_{it} = \mu + \gamma_{0i} + \gamma_{1i}t + \delta_t + \alpha D_{it} + X_{it}'\beta + \varepsilon_{it}$$

  with unit specific intercepts γ_{0i} and unit specific trend coefficients γ_{1i} that multiply time trend variable t.

Difference-in-Differences: Threats to Validity

1. **Compositional differences**: In repeated cross-sections we do not want that the composition of the sample changes between periods.

   The distribution of (D, X) should be similar for the pre-treatment and post-treatment periods.

2. **Non-parallel trends**: Different trends for treated and nontreated in the absence of the treatment.

   Falsification Test: Under the parallel trends assumption during the periods $t = -1, 0, 1$, we have:

$$E[Y(0) - Y(-1)|D = 1] - E[Y(0) - Y(-1)|D = 0] = 0$$

   Apply DID estimator to $t = -1, 0$ and test if $\alpha = 0$

DDD: Mandated Maternity Benefits (Gruber, 1994)

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. *Treatment Individuals: Married Women, 20–40 Years Old:*			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:	−0.062 (0.022)		
B. *Control Group: Over 40 and Single Males 20–40:*			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:	−0.008: (0.014)		
DDD:	**−0.054** **(0.026)**		

Terrorism in Cities (Abadie and Dermisi, 2008)

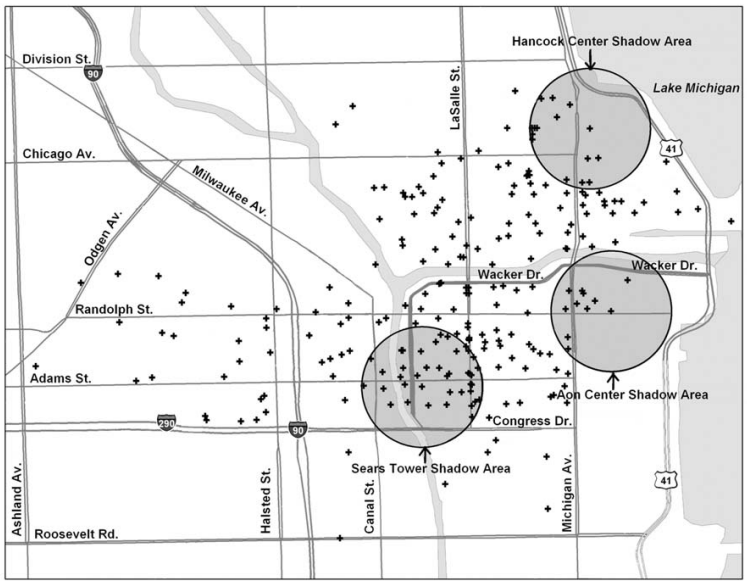


FIGURE I

Chicago's Central Business District Office Buildings and Shadow Areas

Crosses represent all Class A and Class B office buildings in Chicago's Central Business District. Shaded circles represent 0.3-mile radius "shadow areas" surrounding the three main Chicago landmark buildings: the Aon Center, the Hancock Center, and the Sears Tower.

Terrorism in Cities (Abadie and Dermisi, 2008)

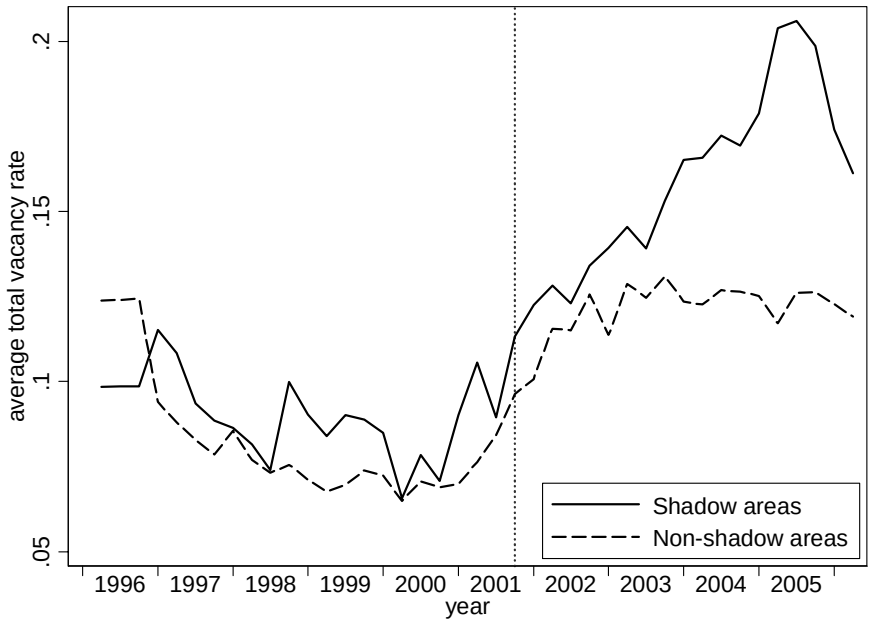


FIGURE II

Average Vacancy Rates in Shadow and Non-shadow Areas

Terrorism in Cities (Abadie and Dermisi, 2008)

TABLE I

DESCRIPTIVE STATISTICS – MEANS AND STANDARD DEVIATIONS
(Class A and B office buildings in downtown Chicago)

	(1) Entire sample	(2) Inside shadow areas	(3) Outside shadow areas	(4) Diff. (2)-(3) (s.e.)
Characteristics of the buildings:				
shadow (= 1 if in shadow area, = 0 otherwise)	.27 [.45]			
Class A (=1 if Class A building, =0 if Class B building)	.21 [.41]	.44 [.50]	.13 [.34]	.31** (.07)
distance to anchor (miles)	.46 [.26]	.19 [.08]	.56 [.24]	-.38** (.02)
height (hundred feet)	2.76 [2.46]	4.43 [2.90]	2.14 [1.94]	2.29** (.39)
number of stories	19.77 [18.80]	32.59 [21.67]	14.96 [15.08]	17.63** (2.90)
rentable building area (sq. feet)	353,683 [499,847]	665,705 [604,842]	236,675 [397,123]	429,031** (80,243)
Vacancy rates (fraction):				
First quarter of 2001	.0803 [.0949]	.0901 [.0903]	.0699 [.0989]	.0202 (.0174)
First quarter of 2006	.1491 [.1306]	.1740 [.1302]	.1228 [.1266]	.0512** (.0248)
Rent per square foot (current USD):				
First quarter of 2001	30.40 [5.43]	32.22 [5.59]	28.08 [4.25]	4.14** (1.23)
First quarter of 2006	28.08 [5.97]	29.09 [5.30]	26.78 [6.54]	2.31* (1.28)
Number of buildings in the sample	242	66	176	

Terrorism in Cities (Abadie and Dermisi, 2008)

TABLE II

9/11 AND VACANCY RATES IN DOWNTOWN CHICAGO OFFICE BUILDINGS
(Fixed-effects estimates with clustered standard errors, 1996-2006)

Dependent variable: Building vacancy rate

	(1)	(2)	(3)	(4)
shadow area×post-9/11	.0303* (.0166)			
distance to anchor×post-9/11		-.0617* (.0362)		
distance to non-shadow area×post-9/11			.2302** (.0633)	
height×post-9/11				.0052** (.0022)
R-squared	.39	.39	.39	.39
Number of observations	9,922	9,922	9,922	9,922

Note: The sample is a quarterly panel of Class A and Class B office buildings in the extended Chicago Central Business District between the second quarter of 1996 and the second quarter of 2006. See text of the article for the exact limits of the area of the City of Chicago included in our sample. Observations are weighted by the rentable area of the buildings. All specifications include building fixed effects and a full set of year×quarter dummies. Standard errors (in parentheses) are clustered at the building level.
* indicates statistical significance at the 10% level.
** indicates statistical significance at the 5% level.

Terrorism in Cities (Abadie and Dermisi, 2008)

TABLE III

TIME SINCE 9/11 AND VACANCY RATES IN DOWNTOWN CHICAGO OFFICE BUILDINGS
(Fixed-effects estimates with clustered standard errors, 1996-2006)

Dependent variable: Building vacancy rate

	(1)	(2)	(3)	(4)
shadow area×post-9/11	-.0046			
	(.0173)			
shadow area×quarters since 9/11	.0037**			
	(.0017)			
distance to anchor×post-9/11		.0156		
		(.0379)		
distance to anchor×quarters since 9/11		-.0081**		
		(.0039)		
distance to non-shadow area×post-9/11			.0614	
			(.0639)	
distance to non-shadow area×quarters since 9/11			.0178**	
			(.0060)	
height×post-9/11				-.0003
				(.0022)
height×quarters since 9/11				.0006**
				(.0002)
R-squared	.39	.39	.39	.39
Number of observations	9,922	9,922	9,922	9,922

Note: The sample is a quarterly panel of Class A and Class B office buildings in the extended Chicago Central Business District between the second quarter of 1996 and the second quarter of 2006. See text of the article for the exact limits of the area of the City of Chicago included in our sample. Observations are weighted by the rentable area of the buildings. All specifications include building fixed effects and a full set of year×quarter dummies. Standard errors (in parentheses) are clustered at the building level.
* indicates statistical significance at the 10% level.
** indicates statistical significance at the 5% level.

Terrorism in Cities (Abadie and Dermisi, 2008)

TABLE V

REGRESSIONS USING PRE-9/11 DATA ONLY
(Fixed-effects estimates with clustered standard errors, 1996-2001)

Dependent variable: Building vacancy rate

	(1)	(2)	(3)	(4)
shadow area×after 1998	. 0120			
	(.0165)			
distance to anchor×after 1998		-.0313		
		(.0370)		
distance to non-shadow area×after 1998			.1017	
			(.0839)	
height×after 1998				.0026
				(.0025)
R-squared	.48	.48	.48	.48
Number of observations	5,324	5,324	5,324	5,324

Note: The sample is a quarterly panel of Class A and Class B office buildings in the extended Chicago Central Business District between the second quarter of 1996 and the third quarter of 2001. See text of the article for the exact limits of the area of the City of Chicago included in our sample. Observations are weighted by the rentable area of the buildings. All specifications include building fixed effects and a full set of year×quarter dummies. Standard errors (in parentheses) are clustered at the building level.
* indicates statistical significance at the 10% level.
** indicates statistical significance at the 5% level.

Terrorism in Cities (Abadie and Dermisi, 2008)

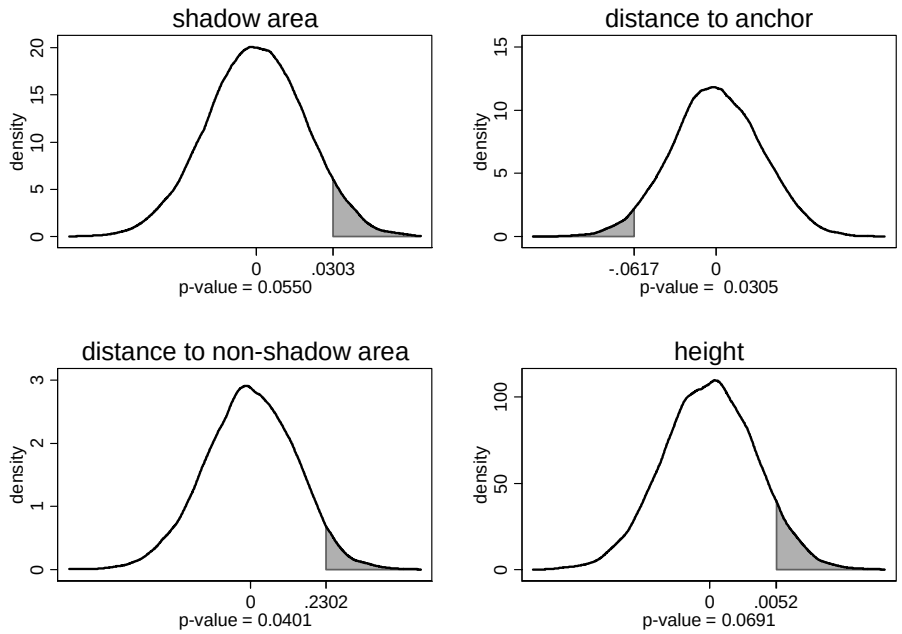


FIGURE III

Permutation Distributions and p-Values