# Lecture 5: Difference-in-Differences

P.J. Messe[1]

[1]Le Mans Université GAINS-TEPP, CEET, LEMNA

Master in Applied Econometrics

# Outline

# Difference-in-Differences (DiD)

- A method that exploits the time dimension to control for a time-invariant unobserved factor that may affect selection into treatment
- Useful when coupled with events or treatments that arise from a "natural experiment"

# A classic DiD example: the Mariel boatlift

What is the impact of immigration on a local labor market unemployment rate?

- Hard to estimate a causal effect due to self-selection of immigrants

# A classic DiD example: the Mariel boatlift

What is the impact of immigration on a local labor market unemployment rate?

- Hard to estimate a causal effect due to self-selection of immigrants
- Card (1990) exploits a natural experiment provided by the Mariel boatlift
  - A mass emigration of Cubans who traveled from Cuba's Mariel harbor to the US between 15 April and 30 october 1980
  - After Castro's decree that stated that Mariel would be opened to anyone wishing to leave Cuba

# A classic DiD example: the Mariel boatlift

- About 125 000 low-skilled Cubans arrived to Miami in 1980, increasing labor force by 7 percent.
- We would like to know the causal effect on unemployment rate in Miami (treated local labor market) in 1981:

$$ATT = E[Y_i(1)|Miami, 1981] - E[Y_i(0)|Miami, 1981]$$

- where $Y_i(D_i)$ is employment rate with/without the treatment

# A classic DiD example: the Mariel boatlift

- About 125 000 low-skilled Cubans arrived to Miami in 1980, increasing labor force by 7 percent.
- We would like to know the causal effect on unemployment rate in Miami (treated local labor market) in 1981:

$$ATT = E[Y_i(1)|Miami, 1981] - E[Y_i(0)|Miami, 1981]$$

- where $Y_i(D_i)$ is employment rate with/without the treatment
- Naive estimator: a before-after comparison

$$E[Y_i(1)|Miami, 1981] - E[Y_i(0)|Miami, \mathbf{1979}]$$

# The variation in unemployment rates in Miami over the period 1979-1985 (Card, 1990)

The before-after (1979-1981) comparison would suggest that this mass immigration has increased unemployment rate among Blacks by 1.3 percentage points.

*Table 4.* Unemployment Rates of Individuals Age 16–61 in Miami and Four Comparison Cities, 1979–85. (Standard Errors in Parentheses)

| *Group* | *1979* | *1980* | *1981* | *1982* | *1983* | *1984* | *1985* |
|---|---|---|---|---|---|---|---|
| *Miami:* | | | | | | | |
| Whites | 5.1 | 2.5 | 3.9 | 5.2 | 6.7 | 3.6 | 4.9 |
| | (1.1) | (0.8) | (0.9) | (1.1) | (1.1) | (0.9) | (1.4) |
| Blacks | 8.3 | 5.6 | 9.6 | 16.0 | 18.4 | 14.2 | 7.8 |
| | (1.7) | (1.3) | (1.8) | (2.3) | (2.5) | (2.3) | (2.3) |
| Cubans | 5.3 | 7.2 | 10.1 | 10.8 | 13.1 | 7.7 | 5.5 |
| | (1.2) | (1.3) | (1.5) | (1.5) | (1.6) | (1.4) | (1.7) |
| Hispanics | 6.5 | 7.7 | 11.8 | 9.1 | 7.5 | 12.1 | 3.7 |
| | (2.3) | (2.2) | (3.0) | (2.5) | (2.1) | (2.4) | (1.9) |

# A classic DiD example: the Mariel boatlift

- Why the before-after comparison is not a relevant estimator of the ATT?
  - Because unemployment rates may change over time due to other shocks not related to the Mariel boatlift
  - For instance, the recession occured in US in the early 1980's

# A classic DiD example: the Mariel boatlift

- Why the before-after comparison is not a relevant estimator of the ATT?
  - Because unemployment rates may change over time due to other shocks not related to the Mariel boatlift
  - For instance, the recession occured in US in the early 1980's
- Idea of the DiD: Find **comparison cities** that experienced similar shocks as Miami BUT did not receive the treatment (here, the mass emigration of Cubans from Mariel harbor)
- Card selects Atlanta, Houston, Los Angeles and Tampa

# A classic DiD example: the Mariel boatlift

- Let $D_{it} = 1$ if the city receives the treatment (Miami), and $D_{it} = 0$ for comparison cities
- Let $P_{it} = 1$ if the unemployment rate is observed in the second time period ($t = 1981$), $P_{it} = 0$ if it is observed in the first-time period ($t = 1979$)

# A classic DiD example: the Mariel boatlift

- Let $D_{it} = 1$ if the city receives the treatment (Miami), and $D_{it} = 0$ for comparison cities
- Let $P_{it} = 1$ if the unemployment rate is observed in the second time period ($t = 1981$), $P_{it} = 0$ if it is observed in the first-time period ($t = 1979$)

- The treatment is defined by the **interaction** of $D_{it}$ and $P_{it}$ since the treatment is received in the second time period in Miami
- ATT is identified comparing the difference in unemployment rates between Miami and comparison cities, before and after 1980:

$$ATT = E[Y_{it}|D_{it} = 1, P_{it} = 1] - E[Y_{it}|D_{it} = 0, P_{it} = 1] -$$

$$E[Y_{it}|D_{it} = 1, P_{it} = 0] - E[Y_{it}|D_{it} = 1, P_{it} = 0]$$

# The variation in unemployment rates in Miami and comparison cities over the period 1979-1985 (Card, 1990)

- For Blacks, in 1979 the difference in unemployment rates between Miami and comparison cities was $8.3 - 10.3 = -2$.
- In 1981, this difference is $9.6 - 12.6 = -3$
- The DiD is equal to $-3 - (-2) = -1$: no evidence that the Mariel influx adversely affected the unemployment rate of Blacks

*Table 4.* Unemployment Rates of Individuals Age 16–61 in Miami and Four Comparison Cities, 1979–85. (Standard Errors in Parentheses)

| *Group* | *1979* | *1980* | *1981* | *1982* | *1983* | *1984* | *1985* |
|---|---|---|---|---|---|---|---|
| *Miami:* | | | | | | | |
| Whites | 5.1 | 2.5 | 3.9 | 5.2 | 6.7 | 3.6 | 4.9 |
| | (1.1) | (0.8) | (0.9) | (1.1) | (1.1) | (0.9) | (1.4) |
| Blacks | 8.3 | 5.6 | 9.6 | 16.0 | 18.4 | 14.2 | 7.8 |
| | (1.7) | (1.3) | (1.8) | (2.3) | (2.5) | (2.3) | (2.3) |
| Cubans | 5.3 | 7.2 | 10.1 | 10.8 | 13.1 | 7.7 | 5.5 |
| | (1.2) | (1.3) | (1.5) | (1.5) | (1.6) | (1.4) | (1.7) |
| Hispanics | 6.5 | 7.7 | 11.8 | 9.1 | 7.5 | 12.1 | 3.7 |
| | (2.3) | (2.2) | (3.0) | (2.5) | (2.1) | (2.4) | (1.9) |
| *Comparison Cities:* | | | | | | | |
| Whites | 4.4 | 4.4 | 4.3 | 6.8 | 6.9 | 5.4 | 4.9 |
| | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) | (0.3) | (0.4) |
| Blacks | 10.3 | 12.6 | 12.6 | 12.7 | 18.4 | 12.1 | 13.3 |
| | (0.8) | (0.9) | (0.9) | (0.9) | (1.1) | (0.9) | (1.3) |

# Outline

# Basic DiD assumptions

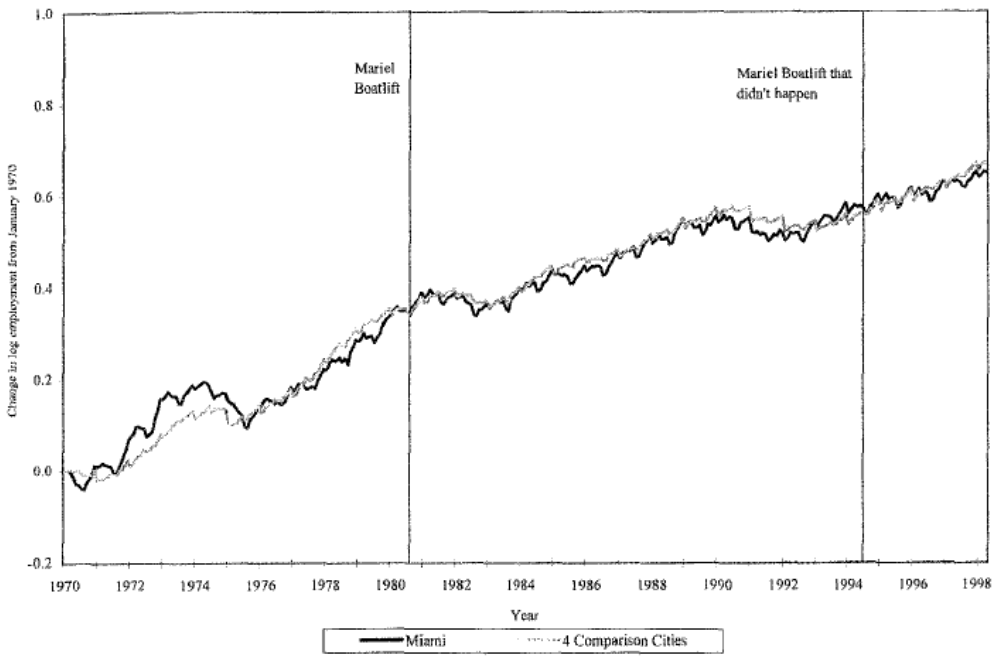
- A key assumption: the time-invariance assumption
  - In absence of treatment, unemployment rates in treatment and control group should have **parallel trends over time**
  - Requires careful choice of comparison groups whe possible

# Basic DiD assumptions

- A key assumption: the time-invariance assumption
  - In absence of treatment, unemployment rates in treatment and control group should have **parallel trends over time**
  - Requires careful choice of comparison groups whe possible
- Other important assumptions
  - Constant treatment effect
  - Macro shocks have the same effect in the treatment and control group
  - No anticipation of the treatment

# The variation in employment rates in Miami and comparison cities over the period 1970-1998 (Card, 1990)

- Before the 1980 Mariel boatlift employment rates in Miami and in comparison cities have parallel trends
- After this event, there is no significant change in trends: no significant effect



# Basic DiD formulation

- We estimate the following linear probabilitymodel

$$Y_{it} = \alpha + \delta_g D_{it} + \delta_t P_{it} + \tau D_{it} * P_{it} + \epsilon_{it}$$

- The interaction term between the treatment dummy $D_{it}$ and the post-event dummy $P_{it}$ captures the effect of interest
- The OLS estimate of $\tau$ is the DiD estimator of the ATT.

# Basic DiD formulation

- We estimate the following linear probabilitymodel

$$Y_{it} = \alpha + \delta_g D_{it} + \delta_t P_{it} + \tau D_{it} * P_{it} + \epsilon_{it}$$

- The interaction term between the treatment dummy $D_{it}$ and the post-event dummy $P_{it}$ captures the effect of interest
- The OLS estimate of $\tau$ is the DiD estimator of the ATT.

- DiD allows to remove unobservable unit-specific effects constant over time and common time effects.
- BUT time-variant unobserved heterogeneity possibly remains unidentified.

# Outline

# Multiple groups and periods

- DiD can accomodate multiple groups and periods
- In the Card example, group dummies $G_{it}$ could be defined by different cities ($g = 1$: Houston, $g = 2$: Los Angeles, ... $g = 5$: Miami)
- Time dummies $P_{it}$ could be 1979, 1981, 1982 ... 1985

# Multiple groups and periods

- DiD can accomodate multiple groups and periods
- In the Card example, group dummies $G_{it}$ could be defined by different cities ($g = 1$: Houston, $g = 2$: Los Angeles, ... $g = 5$: Miami)
- Time dummies $P_{it}$ could be 1979, 1981, 1982 ... 1985

- The corresponding regression would be:

$$Y_{it} = \alpha + \sum_{g=2}^{5} \delta_g 1(G_{it} = g) + \sum_{t=1981}^{1985} \delta_j 1(j = t)$$

$$+ \sum_{t=1981}^{1985} \tau_t 1(G_{it} = 5) * 1(j = t) + \epsilon_{it}$$

# Multiple groups and periods

- Multiple groups allow to do some **"placebo" tests**
- If at least two groups not receiving treatment are available a placebo test consists in estimating the DiD on them and checking whether this is not significantly different from zero
- Very useful to check whether any other change occurred over the period of interest could have contributed to the effect

# Multiple groups and periods

- In our example, consider this regression

$$Y_{it} = \alpha + \sum_{g=2}^{5} \delta_g 1(G_{it} = g) + \sum_{t=1981}^{1985} \delta_j 1(j = t)$$

$$+ \sum_{t=1981}^{1985} \sum_{g=2}^{4} \rho_{gt} 1(G_{it} = g) * 1(j = t) + \epsilon_{it}$$

- In that case, the different coefficients $\rho_{gt}$ should not be significant
- We should not observe a significant evolution in the difference in unemployment rates before and after the Mariel boatlift between Houston and Los Angeles or between Houston and Atlanta

# Multiple levels

- DiD helps control for a time-invariant unobserved factor by differencing out this factor over time
- If access to data at different levels is available, other types of unobserved factors can be differenced out.
- For the case of an additional level (besides groups and time) the method is known as a "triple difference" or DiDiD: an extension of DiD.

# Multiple levels: an example

- A public policy is implemented in 2019 at the **state** level in Pennsylvania ($G_{it} = 1$) for **elementary** school children ($E_i = 1$).
  - The state of Florida ($G_{it} = 0$) can be used as comparison group
  - AND we can also look at **middle** school students ($E_i = 0$) as another comparison group

# Multiple levels: an example

- A public policy is implemented in 2019 at the **state** level in Pennsylvania ($G_{it} = 1$) for **elementary** school children ($E_i = 1$).
  - The state of Florida ($G_{it} = 0$) can be used as comparison group
  - AND we can also look at **middle** school students ($E_i = 0$) as another comparison group

- The regression model can be written as:

$$Y_{it} = \alpha + \delta_g 1(G_{it} = 1) + \delta_t 1(t > 2019) + \delta_e 1(E_{it} = 1)$$

$$+\delta_{gt} 1(G_{it} = 1) * 1(t > 2019) + \delta_{te} 1(t > 2019) * 1(E_{it} = 1) + \delta_{ge} 1(G_{it} = 1) * 1(E_{it} = 1)$$

$$+\tau 1(G_{it} = 1) * 1(t > 2019) * 1(E_{it} = 1) + \epsilon_{it}$$

# Multiple levels: the DiDiD estimator

- The advantage of the DiDiD model is that it also controls for factors that vary by school level
- The DiDiD estimator, i.e. the ATT denoted by $\tau$ is identified from the interaction of the group-, time- and school-level dummies.

# Multiple levels: the DiDiD estimator

- The advantage of the DiDiD model is that it also controls for factors that vary by school level
- The DiDiD estimator, i.e. the ATT denoted by $\tau$ is identified from the interaction of the group-, time- and school-level dummies.
- This allows to implement placebo tests
  - Ex: we should see no significant effect of this policy in Pennsylvania before and after 2019 for **middle-school** children

# Outline

# Combining DiD with matching models

- A key assumption in DiD: common trends before the treatment occurs between treated and comparison groups
  - To make this more plausible, assume that this holds CONDITIONALLY on a set of pre-treatment characteristics
  - An adaptation of the Conditional Independence Assumption to the DiD case

# Combining DiD with matching models

- A key assumption in DiD: common trends before the treatment occurs between treated and comparison groups
  - To make this more plausible, assume that this holds CONDITIONALLY on a set of pre-treatment characteristics
  - An adaptation of the Conditional Independence Assumption to the DiD case

- Allows to control for observed heterogeneity between treated and control groups through standard matching methods (cf lecture 2 and 3) in two steps
  1. Computing weights associated to each control observation according to their closeness to treated ones in terms of characteristics (basic principle of matching)
  2. Computing the DiD estimator using the weights obtained in the first step

# Combining DiD with matching models

- A method widely used in DiD models
  - Combining DiD with propensity-score based matching methods (Heckman et al., 1998, 1999)
  - Semi-parametric DiD estimator: DiD + IPW matching (Abadie, 2005)

# Combining DiD with matching models

- A method widely used in DiD models
  - Combining DiD with propensity-score based matching methods (Heckman et al., 1998, 1999)
  - Semi-parametric DiD estimator: DiD + IPW matching (Abadie, 2005)

- Allows to control for observed heterogeneity between treated and control groups through standard matching methods (cf lecture 2 and 3) in two steps
  1. Computing weights associated to each control observation according to their closeness to treated ones in terms of characteristics (basic principle of matching)
  2. Computing the DiD estimator using the weights obtained in the first step

# Outline

# Be cautious about standard errors

- When using DiD models, be cautious whether outcomes are correlated within groups or over time
- Standard errors not accounting for this correlation may be misleading

# Be cautious about standard errors

- DiD consists of groups (ex: cities in the Card's study)
  - Outcomes are expected to be correlated WITHIN groups
  - In the Card's study, employment outcomes are correlated within cities due to potential shocks at the city-level

# Be cautious about standard errors

- DiD consists of groups (ex: cities in the Card's study)
  - Outcomes are expected to be correlated WITHIN groups
  - In the Card's study, employment outcomes are correlated within cities due to potential shocks at the city-level

- Correlation within groups is often modeled using a group structure of the residual $\epsilon_i$ (omitting time-dimension for the moment)

$$\epsilon_i = e_g + \eta_{ig}$$

  where $e_g$ is a component specific to each group $g$
- This error structure may increase standard errors sharply with respect to conventional standard errors (Moulton, 1986; Angrist and Pischke, 2008).

# Be cautious about standard errors

- The intra-class correlation coefficient can be written as:

$$\rho = \frac{\sigma_e^2}{\sigma_e^2 + \sigma_\eta^2}$$

- where $\sigma_e^2$ is the variance of $e_g$ and $\sigma_\eta^2$ is the variance of $\eta_{ig}$

# Be cautious about standard errors

- The intra-class correlation coefficient can be written as:

$$\rho = \frac{\sigma_e^2}{\sigma_e^2 + \sigma_\eta^2}$$

- where $\sigma_e^2$ is the variance of $e_g$ and $\sigma_\eta^2$ is the variance of $\eta_{ig}$
- The intra-class correlation coefficient $\rho_x$ of one covariate $x_{ig}$ that varies both at the individual level and for different group sizes $n_g$ can be written as:

$$\rho_x = \frac{\sum_g \sum_{i \neq k} (x_{ig} - \overline{x})(x_{kg} - \overline{x})}{V(x_{ig}) \sum_g n_g (n_g - 1)}$$

# Be cautious about standard errors

- Let $V_c(\hat{\beta})$ be the conventional OLS variance formula for the regression coefficient and let $V(\hat{\beta})$ be the correct sampling variance given the error structure.
- Moulton (1986) shows that:

$$\frac{V(\hat{\beta})}{V_c(\hat{\beta})} = 1 + [\frac{V(n_g)}{\overline{n}} + \overline{n} - 1]\rho_x\rho$$

# Be cautious about standard errors

- Let $V_c(\hat{\beta})$ be the conventional OLS variance formula for the regression coefficient and let $V(\hat{\beta})$ be the correct sampling variance given the error structure.
- Moulton (1986) shows that:

$$\frac{V(\hat{\beta})}{V_c(\hat{\beta})} = 1 + [\frac{V(n_g)}{\overline{n}} + \overline{n} - 1]\rho_x\rho$$

- The Moulton factor is the square root of this ratio
- It measures the extent to which conventional OLS standard errors are underestimated.

# An example for computing the Moulton factor

- The Tennessee STAR experiment has consisted in randomly assigning students to classes of different sizes.
  - The goal: estimating the effect of reducing class size on students' test scores.
  - Issue: Standard errors have to account for the intra-class correlation coefficient of residuals $\rho$.

# An example for computing the Moulton factor

- The Tennessee STAR experiment has consisted in randomly assigning students to classes of different sizes.
  - The goal: estimating the effect of reducing class size on students' test scores.
  - Issue: Standard errors have to account for the intra-class correlation coefficient of residuals $\rho$.

- Here, $\rho_x = 1$: class size does not vary within class
- $V(n_g) = 17.1$: variance in class sizes
- $\rho = 0.31$ and the average class size $\overline{n} = 19.4$
- The Moulton factor is:

$$\sqrt{\frac{V(\hat{\beta})}{V_c(\hat{\beta})}} = 1 + [\frac{17.1}{19.4} + 19.4 - 1] * 0.31 \approx \sqrt{7} \approx 2.65$$

# Addressing the issue of intra-class correlation

- Computing Moulton factor and correct the OLS standard errors to account for intra-class correlation

- Use clustered standard errors (cf lecture 2)

$$(X'X)^{-1}(\frac{n_g}{(n_g-1)}\frac{N-1}{(N-K)}\sum X_g\hat{\epsilon}_g\hat{\epsilon}'_g X'_g)(X'X)^{-1}$$

  - where $\hat{\epsilon}_g$ is the vector of residuals for units of the group $g$ and $X_g$ is the vector of covariates for this units.
  - Not recommended in case of too few clusters

# Addressing the issue of intra-class correlation

- Computing Moulton factor and correct the OLS standard errors to account for intra-class correlation

- Use clustered standard errors (cf lecture 2)

$$(X'X)^{-1}(\frac{n_g}{(n_g-1)}\frac{N-1}{(N-K)}\sum X_g\hat{\epsilon}_g\hat{\epsilon}'_g X'_g)(X'X)^{-1}$$

  - where $\hat{\epsilon}_g$ is the vector of residuals for units of the group $g$ and $X_g$ is the vector of covariates for this units.
  - Not recommended in case of too few clusters

- Block-bootstrap: re-sampling **groups** and not only individuals

# The additional issue of serial correlation in DiD model

- Serial correlation: characteristics or outcomes for one observation could be correlated over time
  - Ex : regional shocks are highly serially correlated
  - Additional issue for statistical inference (Bertrand et al., 2004)

- In the Card example, if serial correlation is ignored, clustering by state x year is enough to correct for intra-state and correlation
- BUT the assumption that a state-year shock is serially uncorrelated is not plausible

# The additional issue of serial correlation in DiD model

- Serial correlation: characteristics or outcomes for one observation could be correlated over time
  - Ex : regional shocks are highly serially correlated
  - Additional issue for statistical inference (Bertrand et al., 2004)

- In the Card example, if serial correlation is ignored, clustering by state x year is enough to correct for intra-state and correlation
- BUT the assumption that a state-year shock is serially uncorrelated is not plausible

- One solution: to cluster only by state (and not by state x year)

# The additional issue of serial correlation in DiD model

- Clustering only by state raises another issue: a small number of clusters
- A rule of thumb (Angrist and Pischke, 2008): a need for at least 50 clusters.
- BUT what if the number of clusters is smaller?

- One solution: using a Bias-Reduced Linearization (Bell and McCaffrey, 2002)
  - Adapted to the DiD framework by Pustejovsky and Tipton (2016)
- Correcting the bias and using a degrees of freedom correction for Wald tests (implemented with R)