Difference-in-Difference

Mauricio Romero
(Based on Owen Ozier and Pamela Jakiela's notes)

Things that don't work

- Before vs. After comparisons
  - Compares individuals/communities before and after program
  - But does not control for time trends
- Treated vs. Untreated comparisons
  - Compares treated to those untreated
  - But does not control for selection — why didn't untreated get treated?

Two wrongs make a right (sometimes)

- Difference-in-Differences combines the (biased) pre vs. post and (biased) treated vs. non-treated comparisons
  - Sometimes this overcomes selection bias and time trends
- Basic idea: observe the (self-selected) treatment group and a (self-selected) comparison group before and after the program

$$\delta^{DD} = \left(\overline{Y}_{post}^{treated} - \overline{Y}_{pre}^{treated}\right) - \left(\overline{Y}_{post}^{comparison} - \overline{Y}_{pre}^{comparison}\right)$$

Two wrongs make a right (sometimes)

$$\delta^{DD} \quad = \quad \left(\overline{Y}^{treated}_{post} - \overline{Y}^{treated}_{pre}\right) - \left(\overline{Y}^{comparison}_{post} - \overline{Y}^{comparison}_{pre}\right)$$

- Intuitively
  - $\overline{Y}^{treated}_{post} - \overline{Y}^{treated}_{pre} =$ treatment effect + time trend
  - $\overline{Y}^{comparison}_{post} - \overline{Y}^{comparison}_{pre} =$ time trend
  - $\delta^{DD} =$ treatment effect

Two wrongs make a right

$$\begin{aligned}
\delta^{DD} &= \left(\overline{Y}^{treated}_{post} - \overline{Y}^{treated}_{pre}\right) - \left(\overline{Y}^{comparison}_{post} - \overline{Y}^{comparison}_{pre}\right) \\
&= \left(\overline{Y}^{treated}_{post} - \overline{Y}^{comparison}_{post}\right) - \left(\overline{Y}^{treated}_{pre} - \overline{Y}^{comparison}_{pre}\right)
\end{aligned}$$

- Intuitively II
  - $\overline{Y}^{treated}_{post} - \overline{Y}^{comparison}_{post} =$ treatment effect + selection bias
  - $\overline{Y}^{treated}_{pre} - \overline{Y}^{comparison}_{pre} =$ selection bias
  - $\delta^{DD} =$ treatment effect

Difference-in-Difference

Introduction

The simple 2×2

Regression Framework

Working example

Defending the Common Trends Assumption

Diff-in-Diff in a Panel Data Framework

Standard errors

The simple 2×2

	Treated	Comparison
Pre	$\overline{Y}^{Treated}_{Pre}$	$\overline{Y}^{Comparison}_{Pre}$
Post	$\overline{Y}^{Treated}_{Post}$	$\overline{Y}^{Comparison}_{Post}$

- Intuitively, diff-in-diff estimation is just a comparison of 4 cell-level means

- Only one cell is treated: Treatment×Post

Difference-in-Differences estimation

- Let δ denote the true impact of the program

$$\delta = \mathbb{E}[Y_{1i} | T_i = 1, t = \tau] - \mathbb{E}[Y_{0i} | T_i = 1, t = \tau]$$

- Assumption: δ does not depend on the time period (τ) or i's characteristics

Difference-in-Differences estimation

The **assumption** underlying difference-in-difference estimation boils down to:

- In the absence of the program, individual i's outcome at time t is given by

$$\mathbb{E}[Y_i | T_i = 0, t = \tau] = \gamma_i + \lambda_\tau$$

- Two implicit identifying assumptions
  1. Selection bias relates to fixed individuals characteristics (γ_i)
     - Selection bias does not change over time
  2. Time trend (δ_τ) same for treatment and comparison groups
     - Common/parallel trends assumption

Difference-in-Differences estimation

In the absence of the program, individual i's outcome at time t is given by

$$\mathbb{E}[Y_i | T_i = 0, t = \tau] = \gamma_i + \lambda_\tau$$

Thus

$$\begin{aligned}
\mathbb{E}[Y_{pre}^{comparison}] &= \mathbb{E}[Y_{i0} | T_i = 0, t = pre] = \mathbb{E}[\gamma_i | T_i = 0] + \mathbb{E}[\lambda_\tau | t = pre] \\
\mathbb{E}[Y_{post}^{comparison}] &= \mathbb{E}[Y_{i0} | T_i = 0, t = post] = \mathbb{E}[\gamma_i | T_i = 0] + \mathbb{E}[\lambda_\tau | t = post] \\
\mathbb{E}[Y_{pre}^{treated}] &= \mathbb{E}[Y_{i0} | T_i = 1, t = pre] = \mathbb{E}[\gamma_i | T_i = 1] + \mathbb{E}[\lambda_\tau | t = pre] \\
\mathbb{E}[Y_{post}^{treated}] &= \mathbb{E}[Y_{i1} | T_i = 1, t = pre] = \delta + \mathbb{E}[\gamma_i | T_i = 1] + \mathbb{E}[\lambda_\tau | t = post]
\end{aligned}$$

Treated/Untreated comparison

$$\begin{aligned}
\mathbb{E}[Y_{post}^{treated}] - \mathbb{E}[Y_{post}^{comparison}] &= \delta + \mathbb{E}[\gamma_i | T_i = 1] + \mathbb{E}[\lambda_\tau | t = post] - \\
&\quad \mathbb{E}[\gamma_i | T_i = 0 - \mathbb{E}[\lambda_\tau | t = post] \\
&= \delta + \underbrace{\mathbb{E}[\gamma_i | T_i = 1] - \mathbb{E}[\gamma_i | T_i = 0]}_{\text{selection bias}}
\end{aligned}$$

Post/Pre comparison

$$\begin{aligned}
\mathbb{E}[Y_{post}^{treated}] - \mathbb{E}[Y_{pre}^{treated}] &= \delta + \mathbb{E}[\gamma_i | T_i = 1] + \mathbb{E}[\lambda_\tau | t = post] - \\
&\quad \mathbb{E}[Y_{i0} | T_i = 1, t = pre] = \mathbb{E}[\gamma_i | T_i = 1] - \mathbb{E}[\lambda_\tau | t = pre] \\
&= \delta + \underbrace{\mathbb{E}[\lambda_\tau | t = post] - \mathbb{E}[\lambda_\tau | t = pre]}_{\text{time trend}}
\end{aligned}$$

Difference in Difference comparison

$$
\begin{aligned}
\delta^{DD} &= \left(\overline{Y}^{treated}_{post} - \overline{Y}^{treated}_{pre}\right) - \left(\overline{Y}^{comparison}_{post} - \overline{Y}^{comparison}_{pre}\right) \\
&= (\delta + \mathbb{E}[\gamma_i | T_i = 1] + \mathbb{E}[\lambda_\tau | t = post] - \mathbb{E}[\gamma_i | T_i = 1] - \mathbb{E}[\lambda_\tau | t = pre]) - \\
&\quad (\mathbb{E}[\gamma_i | T_i = 0] + \mathbb{E}[\lambda_\tau | t = post] - \mathbb{E}[\gamma_i | T_i = 0] - \mathbb{E}[\lambda_\tau | t = pre]) \\
&= (\delta + \mathbb{E}[\lambda_\tau | t = post] - \mathbb{E}[\lambda_\tau | t = pre]) - \\
&\quad (\mathbb{E}[\lambda_\tau | t = post] - \mathbb{E}[\lambda_\tau | t = pre]) \\
&= \delta
\end{aligned}
$$

Diff-in-Diff recovers the impact of the program on participants (if assumptions aren't violated)

Difference in Difference comparison

- Diff-in-Diff does not rely on assumption of homogeneous treatment effects
- When treatment effects are homogeneous, DD estimation yields average treatment effect on the treated (ATT)
- If not, it averages across treated units and over time
  - When impacts change over time (within treated units), DD estimate of treatment effect may depend on choice of evaluation window

Difference-in-Difference

Introduction

The simple 2×2

Regression Framework

Working example

Defending the Common Trends Assumption

Diff-in-Diff in a Panel Data Framework

Standard errors

Regression DD

- It's easy to calculate the standard errors
- We can control for other variables which may reduce the residual variance (and smaller standard errors)
- It's easy to include multiple periods (and varying treatment timing)
- We can study treatments with different treatment intensity

DD in a Regression Framework

To implement diff-in-diff in a regression framework, we estimate:

$$Y_{i,t} = \alpha + \beta T_i + \zeta Post_t + \delta \left(T_i * Post_t \right) + \varepsilon_{i,t}$$

where:

- $Post_i$ is an indicator equal to 1 if $t = 2$
- δ is the coefficient of interest (the treatment effect)
- $\alpha = E[\gamma_i | T_i = 0] + \lambda_1$: pre-program mean in comparison group
- $\beta = E[\gamma_i | T_i = 1] - E[\gamma_i | D_i = 0]$: selection bias
- $\zeta = \lambda_2 - \lambda_1$: time trend

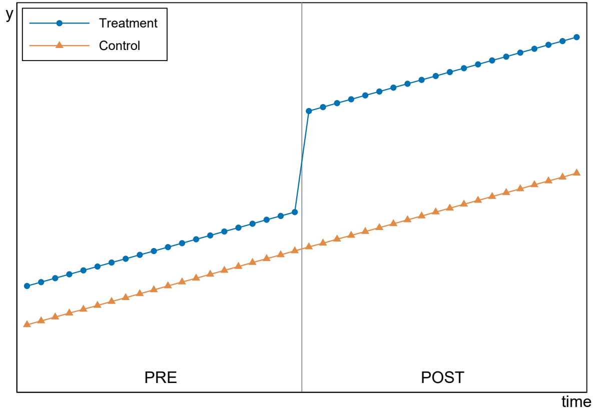
DD in a Regression Framework

- Another option is to use Two-Way Fixed Effects (TWFE)
- With more than two periods of data using TWFE can increase statistical power

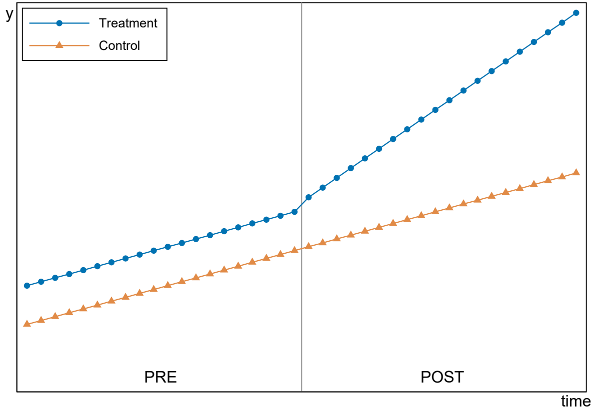
$$Y_{i,t} = \alpha + \eta_i + \nu_t + \gamma T_{i,t} + \varepsilon_{i,t}$$

- η_i unit fixed effects (replaces the $Post_t$ dummy)
- ν_t time fixed effects (replaces the T_i dummy)

DD in a Regression Framework



DD in a Regression Framework



DD in a Regression Framework

Event study framework includes dummies for each post-treatment period:

$$Y_{i,t} = \alpha + \eta_i + \nu_t + \gamma_1 T1_{i,t} + \gamma_2 T2_{i,t} + \gamma_3 T3_{i,t} + \ldots + \varepsilon_{i,t}$$

When treatment intensity is a continuous variable:

$$Y_{i,t} = \alpha + \beta Intensity_i + \zeta Post_t + \delta\left(Intensity_i * Post_t\right) + \varepsilon_{i,t}$$

Difference-in-Difference

Seguro Popular

American Economic Journal: Economic Policy 2014, 6(4): 71–99
http://dx.doi.org/10.1257/pol.6.4.71

The Trade-Offs of Welfare Policies in Labor Markets with Informal Jobs: The Case of the "Seguro Popular" Program in Mexico†

By Mariano Bosch and Raymundo M. Campos-Vazquez*

In 2002, the Mexican government began an effort to improve health access to the 50 million uninsured in Mexico, a program known as Seguro Popular (SP). The SP offered virtually free health insurance to informal workers, altering the incentives to operate in the formal economy. We find that the SP program had a negative effect on the number of employers and employees formally registered in small and medium firms (up to 50 employees). Our results suggest that the positive gains of expanding health coverage should be weighed against the implications of the reallocation of labor away from the formal sector. (*JEL* E26, I13, I18, I38, J46, O15, O17)

Seguro Popular

- Mexico's current social protection system was born in 1943.
    - Formal Sector workers and their families are part of the social protection system (IMSS/ISSSTE)
    - Informal sector workers are uninsured
- By 2000, the inequalities in this system were apparent.
    - Nearly 50 % of the Mexican population ($\sim$ 47 million) was uninsured
- World Health Organization ranked Mexico 144/191 in fairness of health care
- The Mexican Ministry of Health estimated that 10 to 20% of the population, suffered catastrophic and impoverishing health care expenses every year

Seguro popular

- The Sistema de Protección Social en Salud, System for Social Protection in Health (SPS), was designed in the early 2000s to address some of these issues
- A key component of this reform was the Seguro Popular program.
    - Passed into law in 2004 as a modification of the existing General Health Law, the program actually began with a pilot phase in 5 states in 2002
    - Provide health insurance to the 50 million uninsured
- States and municipalities offered virtually free health insurance to informal workers altering the incentives for workers and firms to operate in the formal/registered economy

Identification strategy

- Take advantage of the **staggered implementation** of the program across municipalities

Seguro Popular

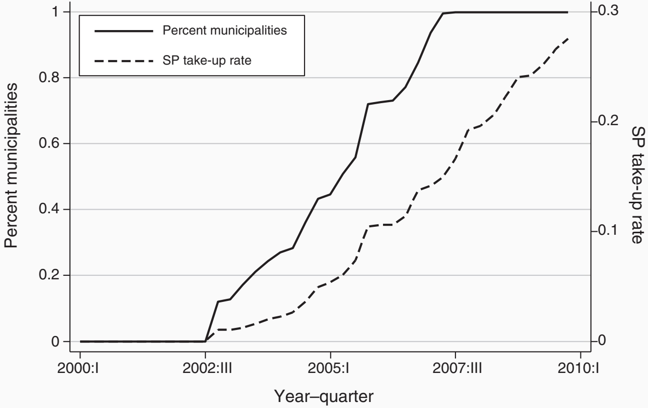


Figure 2. Share of Covered Municipalities and Population: 2000–2009

Notes: The figure shows the share of municipalities treated (left *y*-axis) and the SP take-up rate as a percentage of total population (right *y*-axis). Number of beneficiaries obtained from the administrative records of SP and population from the 2000 Population census and 2005 population count.

Data

- Data from the Instituto Mexicano de Seguro Social (IMSS) records for the **entire universe of municipalities** in Mexico from 2000 to 2009

- Merge with the **administrative records** of Seguro Popular by municipality

The Common Trends Assumption

- The key assumption for any DD strategy is that the outcome in treatment and control group would follow the same time trend in the absence of the treatment
  - This doesn't mean that they have to have the same mean of the outcome
- Alternatively, the assumptions underlying diff-in-diff estimation:
  - Selection bias relates to fixed characteristics of individuals (γ_i)
  - Time trend (λ_t) same for treatment and control groups
- These assumptions cannot be tested directly — we have to trust!
  - As with any identification strategy, it is important to think carefully about whether it checks out both intuitively and econometrically

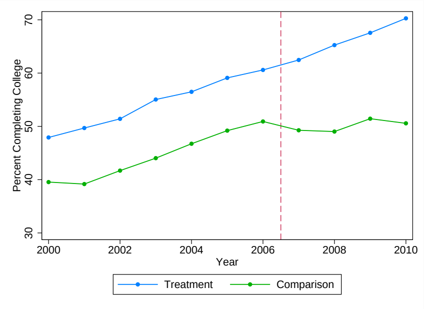
Losing parallel trends

- If parallel trends doesn't hold, then ATT is not identified
- But, regardless of whether ATT is identified, OLS always estimates the same thing
- OLS uses the slope of the control group to estimate the DD parameter, which is only unbiased if that slope is the correct counterfactual for the treatment

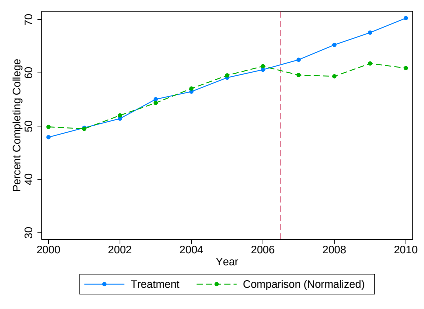
Parallel leads, not trends

- Parallel trends cannot be directly verified because technically one of the parallel trends is an unobserved counterfactual
- But one often will check using pre-treatment data to show that the trends had been the same prior to treatment
- But, even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias)

The Common Trends Assumption

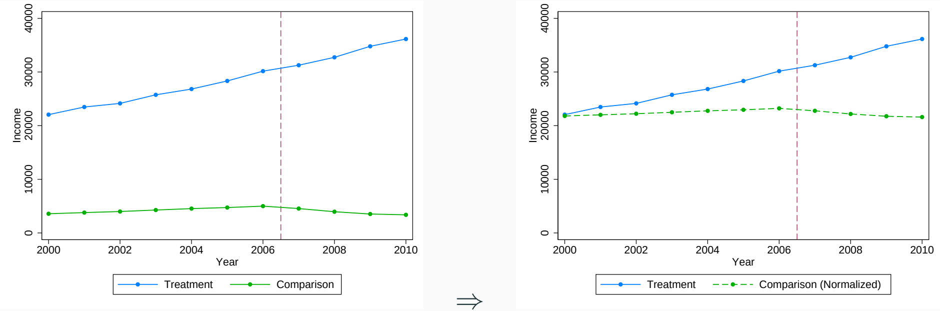


$\Rightarrow$



Sometimes, the common trends assumption is clearly OK

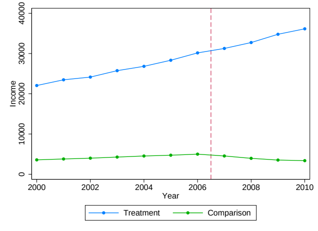
The Common Trends Assumption



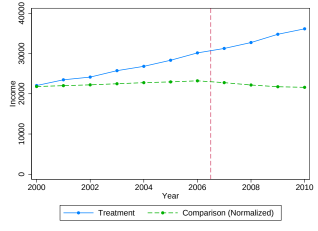
Other times, the common trends assumption is fairly clearly violated

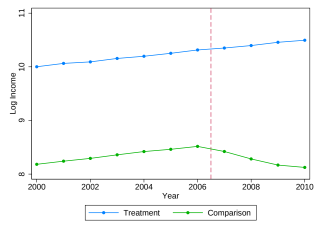
The Common Trends Assumption

Or is it? DD is robust to transformations of the outcome variable

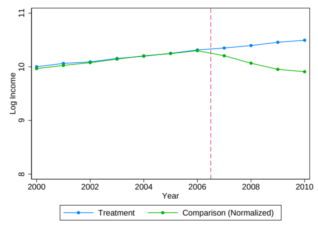


⇒





⇒



Defending the Common Trends Assumption

Three approaches:

1. A compelling graph
2. A falsification test or, analogously, a direct test in panel data
3. Controlling for time trends directly
    - Drawback: identification comes from functional form assumption

Defending the Common Trends Assumption

Three approaches:

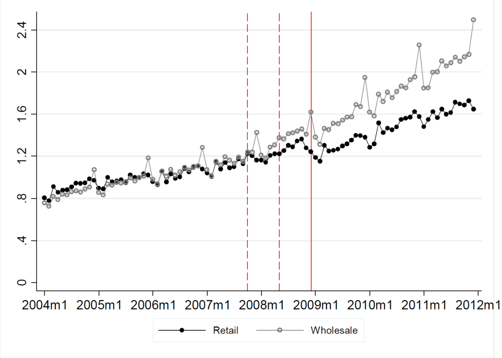
1. A compelling graph
2. A falsification test or, analogously, a direct test in panel data
3. Controlling for time trends directly
    - Drawback: identification comes from functional form assumption

None of these approaches are possible with two periods of data

Approach #1: DD Porn

Figure 4: Compliance Effect – Retail vs. Wholesale

a. Raw data: reported revenue changes



Source: Naritomi (2015)

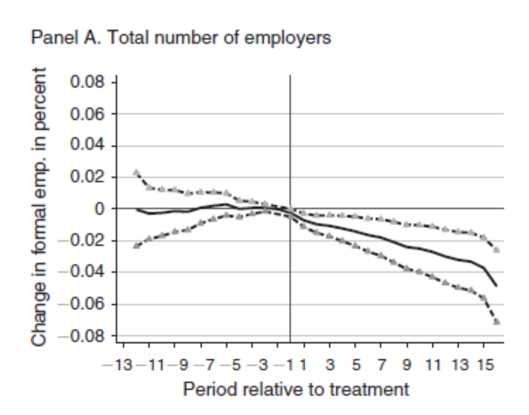
Event study regression

- Including leads into the DD model is an easy way to analyze pre-treatment trends
- Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

$$Y_{its} = \gamma_s + \lambda_t + \sum_{\tau=-1}^{-q} \gamma_\tau D_{s\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{s\tau} + x_{ist} + \varepsilon_{ist}$$

  - Treatment occurs in year 0
  - Includes q leads or anticipatory effects
  - Includes m leads or post treatment effects

Approach #3: Event Study



Variation in Treatment Timing

Example: municipalities introduced Seguro Popular at different times

Fixed Effects Estimates of β^{DD}

$$Y_{it} = \alpha_i + \gamma_t + \beta^{DD} D_{it} + \varepsilon_{ti}$$

Fixed Effects Estimates of β^{DD}

$$\underbrace{Y_{it} = \alpha_i}_{\text{unit fixed effects}} + \underbrace{\gamma_t}_{\text{time fixed effects}} + \beta^{DD} \underbrace{D_{it}}_{\text{treatment dummy}} + \varepsilon_{ti}$$

Fixed Effects Estimates of β^{DD}

What exactly is β^{DD}?

$$\underbrace{Y_{it} = \alpha_i}_{\text{unit fixed effects}} + \underbrace{\gamma_t}_{\text{time fixed effects}} + \beta^{DD} \underbrace{D_{it}}_{\text{treatment dummy}} + \varepsilon_{ti}$$

Fixed Effects Estimates of β^{DD}

Frisch-Waugh (1933): Two-way fixed effects regression is equivalent to univariate regression:

$$\tilde{Y}_{it} = \tilde{D}_{it} + \zeta_{ti}$$

where

$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i - \left(\bar{Y}_t - \bar{\bar{Y}}\right)$$

and

$$\tilde{D}_{it} = D_{it} - \bar{D}_i - \left(\bar{D}_t - \bar{\bar{D}}\right)$$

Fixed Effects Estimates of β^{DD}

Frisch-Waugh (1933): Two-way fixed effects regression is equivalent to univariate regression:

$$\tilde{Y}_{it} = \tilde{D}_{it} + \zeta_{ti}$$

where

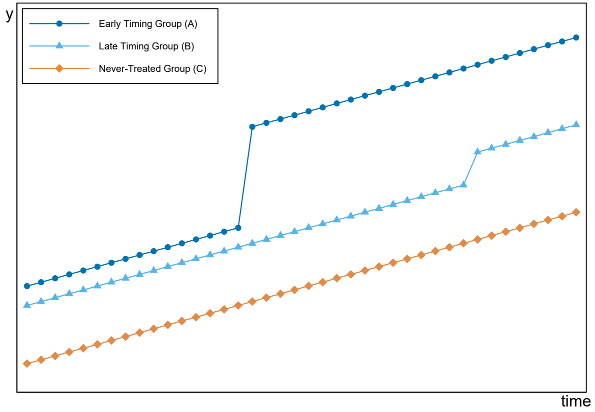
$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i - \left(\bar{Y}_t - \bar{\bar{Y}}\right)$$

and

$$\tilde{D}_{it} = D_{it} - \bar{D}_i - \left(\bar{D}_t - \bar{\bar{D}}\right)$$

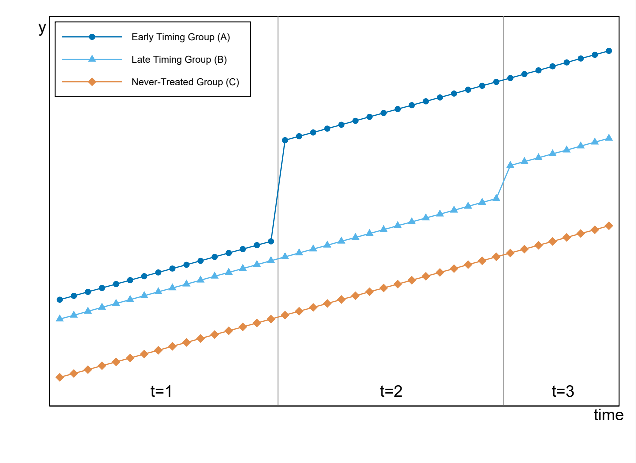
Which is cool, but doesn't really tell us what the estimand is

Decomposition into Timing Groups



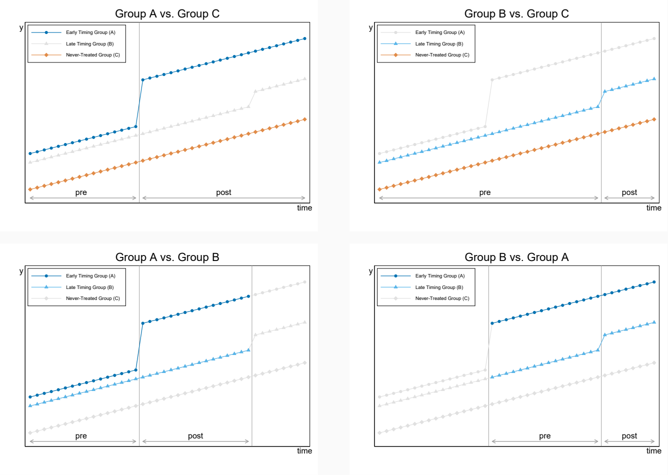
Goodman-Bacon (2019): panel with variation in treatment timing can be decomposed into **timing groups** reflecting observed onset of treatment

Decomposition into Timing Groups

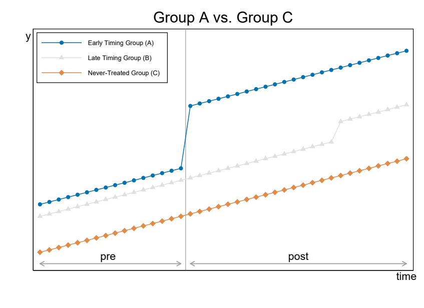


Example: with three timing groups (one of which is never treated), we can construct three timing windows (pre, middle, post or $t = 1, 2, 3$)

Decomposition into Standard 2×2 DDs



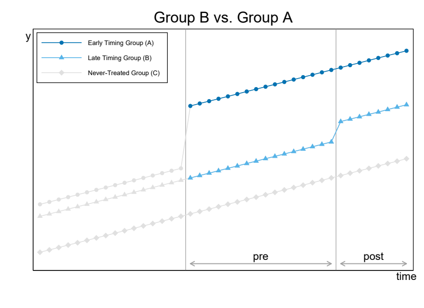
Decomposition into Standard 2×2 DDs



We know the DD estimate of the treatment effect for each timing group:

$$\widehat{\beta}_{AC}^{DD} = \left(\bar{Y}_A^{POST} - \bar{Y}_C^{POST}\right) - \left(\bar{Y}_A^{PRE} - \bar{Y}_C^{PRE}\right)$$
$$= \left(\bar{Y}_A^{t=2,3} - \bar{Y}_C^{t=2,3}\right) - \left(\bar{Y}_A^{t=1} - \bar{Y}y_C^{t=1}\right)$$

Decomposition into Standard 2×2 DDs



We know the DD estimate of the treatment effect for each timing group:

$$\widehat{\beta}_{BA}^{DD} = \left(\bar{Y}_B^{POST} - \bar{Y}_A^{POST}\right) - \left(\bar{Y}_B^{PRE} - \bar{Y}_A^{PRE}\right)$$
$$= \left(\bar{Y}_B^{t=3} - \bar{Y}_A^{t=3}\right) - \left(\bar{Y}_B^{t=2} - \bar{Y}y_A^{t=2}\right)$$

DD Decomposition Theorem (aka D^3 Theorem)

Theorem
Consider a data set comprising K timing groups ordered by the time at which they first receive treatment and a maximum of one never-treated group, U. The OLS estimate from a two-way fixed effects regression is:

$$\widehat{\beta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\beta}_{kU}^{DD} + \sum_{k \neq U} \sum_{j > k} \left[s_{kj} \widehat{\beta}_{kj}^{DD} + s_{jk} \widehat{\beta}_{jk}^{DD} \right]$$

In other words, the DD estimate from a two-way fixed effects regression is a weighted average of the (well-understood) 2×2 DD estimates

DD Decomposition Theorem (aka D^3 Theorem)

Weights depend on sample size, variance of treatment within each DD:

$$s_{kU} = \left[\frac{(n_k + n_U)^2}{\widehat{V}^{\tilde{D}}}\right] \underbrace{n_{kU}\left(1 - n_{kU}\right)\bar{D}_k(1 - \bar{D}_k)}_{\widehat{Var}^{\tilde{D}}_{kU}}$$

$$s_{kj} = \left[\frac{\left(\left(n_k + n_j\right)\left(1 - \bar{D}_j\right)\right)^2}{\widehat{V}^{\tilde{D}}}\right] \underbrace{n_{kj}(1 - n_{kj})\left(\frac{\bar{D}_k - \bar{D}_j}{1 - \bar{D}_j}\right)\left(\frac{1 - \bar{D}_k}{1 - \bar{D}_j}\right)}_{\widehat{Var}^{\tilde{D}}_{kj}}$$

$$s_{jk} = \left[\frac{\left(\left(n_k + n_j\right)\bar{D}_k\right)^2}{\widehat{V}^{\tilde{D}}}\right] \underbrace{n_{kj}(1 - n_{kj})\frac{\bar{D}_j}{\bar{D}_k}\left(\frac{\bar{D}_k - \bar{D}_j}{\bar{D}_k}\right)}_{\widehat{Var}^{\tilde{D}}_{jk}}$$

where n_k is the proportion of the sample in group k, $n_{kj} = n_k/(n_k + n_j)$, and $\bar{D}_k$ is the fraction of sample periods in which k is treated

DD Decomposition Theorem (aka D^3 Theorem)

Weights depend on sample size, variance of treatment within each DD:

$$s_{kU} = \left[\frac{(n_k + n_U)^2}{\widehat{V}^{\tilde{D}}}\right] \underbrace{n_{kU}\,(1 - n_{kU})\,\bar{D}_k(1 - \bar{D}_k)}_{\widehat{Var}^{\tilde{D}}_{kU}}$$

$$s_{kj} = \left[\frac{\left((n_k + n_j)\,(1 - \bar{D}_j)\right)^2}{\widehat{V}^{\tilde{D}}}\right] \underbrace{n_{kj}(1 - n_{kj})\left(\frac{\bar{D}_k - \bar{D}_j}{1 - \bar{D}_j}\right)\left(\frac{1 - \bar{D}_k}{1 - \bar{D}_j}\right)}_{\widehat{Var}^{\tilde{D}}_{kj}}$$

$$s_{jk} = \left[\frac{\left((n_k + n_j)\,\bar{D}_k\right)^2}{\widehat{V}^{\tilde{D}}}\right] \underbrace{n_{kj}(1 - n_{kj})\frac{\bar{D}_j}{\bar{D}_k}\left(\frac{\bar{D}_k - \bar{D}_j}{\bar{D}_k}\right)}_{\widehat{Var}^{\tilde{D}}_{jk}}$$

where n_k is the proportion of the sample in group k, $n_{kj} = n_k/(n_k + n_j)$, and $\bar{D}_k$ is the fraction of sample periods in which k is treated

Implications of the D^3 Theorem

1. When treatment effects are homogeneous, $\widehat{\beta}^{DD}$ is the ATE
2. When treatment effects are heterogeneous across units (not time), $\widehat{\beta}^{DD}$ is a variance-weighted treatment effect that is not the ATE (as usual with OLS)
   - $\Rightarrow$ Weights on timing groups are sums of s_{kU}, s_{kj} terms
3. When treatment effects change over time, $\widehat{\beta}^{DD}$ is biased
   - $\Rightarrow$ Changes in treatment effect bias DD coefficient
   - $\Rightarrow$ Event study, stacked DD more appropriate

Implications of the D^3 Theorem

DD in a potential outcomes framework assuming common trends:

$$Y_{it} = \begin{cases} Y_{0,it} \text{ if } D_{it} = 0 \\ Y_{0,it} + \delta_{it} \text{ if } D_{it} = 1 \end{cases}$$

Implications of the D^3 Theorem

DD in a potential outcomes framework assuming common trends:

$$Y_{it} = \begin{cases} Y_{0,it} \text{ if } D_{it} = 0 \\ Y_{0,it} + \delta_{it} \text{ if } D_{it} = 1 \end{cases}$$

$\widehat{\beta}_{kU}^{DD}$ and $\widehat{\beta}_{kj}^{DD}$ (where $k < j$) are familiar, but $\widehat{\beta}_{jk}^{DD}$ is different:

$$\begin{aligned}
\widehat{\beta}_{jk}^{DD} &= \bar{Y}_{0,j}^{POST} + \bar{\delta}_{j}^{POST} - \left(\bar{Y}_{0,k}^{POST} + \bar{\delta}_{k}^{POST}\right) - \left[\bar{Y}_{0,j}^{PRE} - \left(\bar{Y}_{0,k}^{PRE} + \bar{\delta}_{k}^{PRE}\right)\right] \\
&= \bar{\delta}_{j}^{POST} + \underbrace{\left[\left(\bar{Y}_{0,j}^{POST} - \bar{Y}_{0,k}^{POST}\right) - \left(\bar{Y}_{0,j}^{PRE} - \bar{Y}_{0,k}^{PRE}\right)\right]}_{\text{common trends}} + \underbrace{\left(\bar{\delta}_{k}^{PRE} - \bar{\delta}_{k}^{POST}\right)}_{\Delta\delta_k}
\end{aligned}$$

Weights discussion

- Think about what causes the treatment variance to be as big as possible. Let's think about the s_{ku} weights.

    1. $\overline{D} = 0.1$. Then $0.1 \times 0.9 = 0.09$
    2. $\overline{D} = 0.4$. Then $0.4 \times 0.6 = 0.24$
    3. $\overline{D} = 0.5$. Then $0.5 \times 0.5 = 0.25$

- What's this mean? The weight on treatment variance is maximized for *groups treated in middle of the panel*

More weights discussion

- But what about the "treated on treated" weights? What's this $\overline{D}_k - \overline{D}_l$ business about?
- Well, same principle as before - when the difference between treatment variance is close to 0.5, those 2×2s are given the greatest weight
- For instance, say $t_k^* = 0.15$ and $t_l^* = 0.67$. Then $\overline{D}_k - \overline{D}_l = 0.52$. And thus $0.52 \times 0.48 = 0.2496$.

TWFE and centralities

- Groups in the middle of the panel weight up their respective 2×2s via the variance weighting
- But when looking at treated to treated comparisons, when differences in timing have a spacing of around 1/2, those also weight up the respective 2s2s via variance weighting
- But there's no theoretical reason why should prefer this as it's just a weighting procedure being determined by how we drew the panel
- This is the first thing about TWFE that should give us pause, as not all estimators do this

Standard errors in DD strategies

- Many paper using DD strategies use data from many years – not just 1 pre and 1 post period

- The variables of interest in many of these setups only vary at a group level (say a state level) and outcome variables are often serially correlated

- As Bertrand, Duflo and Mullainathan (2004) point out, conventional standard errors often severely understate the standard deviation of the estimators – standard errors are biased downward (i.e., too small, over reject)

Standard errors in DD – practical solutions

- Bertrand, Duflo and Mullainathan propose the following solutions:

  1. Block bootstrapping standard errors (if you analyze states the block should be the states and you would sample whole states with replacement for bootstrapping)

  2. Clustering standard errors at the group level

  3. Aggregating the data at the group level

DD Robustness

- Very common for readers and others to request a variety of "robustness checks" from a DD design
- Think of these as along the same lines as the leads and lags we already discussed
  - Event study (already discussed)
  - Falsification test using data for alternative control group
  - Falsification test using alternative "placebo" outcome that should not be affected by the treatment

Takeaways

1. Stack the 2×2 DDs to asses common trends (visually)

    $\Rightarrow$ Trends should look similar before and after treatment
    $\Rightarrow$ Treatment effect should be a level shift, no a trend break
    $\Rightarrow$ How much weight is placed on problematic timing groups?

2. Plot the relationship between the 2×2 DD estimates, weights

    $\Rightarrow$ No heterogeneity? No problems!
    $\Rightarrow$ Heterogeneity across units is an object of interest

Concluding remarks on DD

- Chances are you are going to write more papers using DD than any other design
- Goodman-Bacon (2018, 2019) is *worth your time* so that you know what you are estimating
- De Chaisemartin & D'Haultfoeuille (2020) and Callaway & Sant'ann (2019) are also *worth your time* if you decide to run a diff-in-diff