Causal Inference: Difference-in-differences

Ken Stiller

6th June 2022

Last week

- Panel Data
- First Difference Estimator
- Time Fixed Effects
- Entity/Unit Fixed Effects
- Two-way Fixed Effects: Time and Entity

Let's begin with a recap!

Two-way fixed effects: time and entity fixed effects

Imagine we are interested in evaluating a policy change D on the outcome Y on a panel of individuals through years:

$$Y_{i,t} = \alpha_i + \delta_t + \beta D_{i,t} + u_{i,t} \tag{1}$$

- What are we controlling for with α_i and δ_t?
- Can we estimate/control for the effect of the Covid pandemic as well in the same specification (assuming it occurred in years 2020 and 2021) ?
- Can we estimate/control for the effect of sex of the individual in the same specification (assuming it correlates to both D and Y) ?
- What is the leftover variation in $D_{i,t}$ used to estimate β?

Recap: Two-way fixed effects

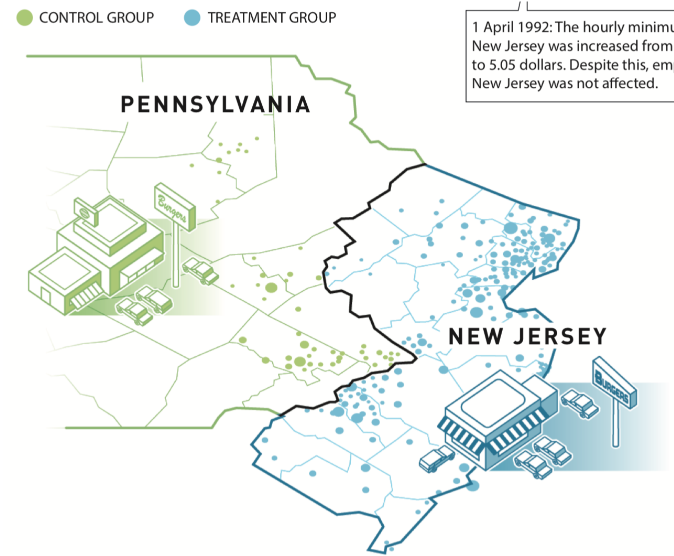
- α_i controls for anything that is time-invariant within a unit, but varies across units
- δ_t controls for anything that is time variant but invariant across units
- $D_{i,t}$ can be anything that varies with time and across units
  - This can be a continuous variable such as the number of women MPs.
  - This can also be a binary variable such as a policy change (ban of a newspaper, introduction of higher minimum wages). $\rightarrow$ lends itself to a Difference-in-Differences research design

Let's begin with a simple example of two time periods and two units.

Diff-in-diff example: Minimum wage laws and employment

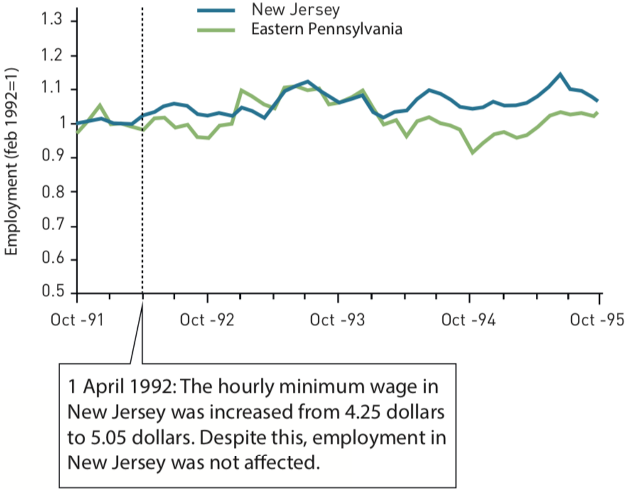
- Do higher minimum wages decrease low-wage employment?
- Card and Krueger (1994) exploit the change in New Jersey's 1992 minimum wage increase from \$4.25 to \$5.05 per hour to measure the effect of minimum wage on unemployment in the fast food industry
- New regulation only applies to NJ, which allows to have other States as control groups
- Compare employment in 410 fast-food restaurants in New Jersey and eastern Pennsylvania before and after the rise
- Survey data on wages and employment from two waves:
  - Wave 1: March 1992, one month before the minimum wage increase
  - Wave 2: December 1992, eight months after increase

Locations of Restaurants (Card and Krueger 2000)



1 April 1992: The hourly minimum wage in New Jersey was increased from 4.25 dollars to 5.05 dollars. Despite this, employment in New Jersey was not affected.

Wages Before and After Rise in Minimum Wage



Sample Means: Minimum wage laws and employment

	Stores by state		
Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Is this a causal estimate? What selection bias is controlled for? What is remaining?

Difference-in-differences

- There are plenty of examples of **treatments that occur at a particular time.** We can see the world before the treatment is applied, and after. We want to know how much of the change in the world is due to that treatment.
- We are looking for how much more the treated group changed than the untreated group when going from before to after. The change in the untreated group represents how much change we would have expected in the treated group if no treatment had occurred. So any additional change beyond that amount must be the effect of the treatment.
- Identification assumption: while the treated and control groups may vary in their characteristics over time, the selection bias into treatment must be **time-invariant**. This is the parallel trends assumption.

Difference-in-differences, II

- **Simple case:** binary treatment, applied at one point in time (but not to everyone)
- **More general case:** general treatment, applied in any pattern
- Panel data requirements: multiple observations over time, with treatment varying within group or unit over time
- Estimation via a regression that controls for time period and group or unit **(fixed effects)**

Notation for time periods

Up to now:

- Potential outcomes: Y_{0i}, Y_{1i}
- Definition linking them: $\tau_i \equiv Y_{1i} - Y_{0i}$

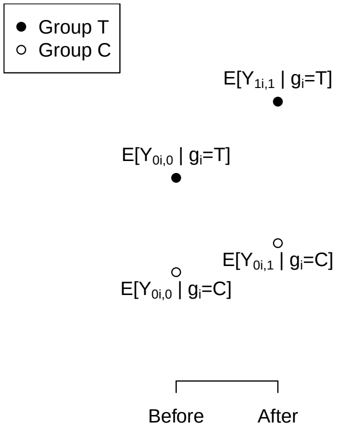
With two time periods:

- Potential outcomes: $Y_{0i,t}$, $Y_{1i,t}$ for $t \in \{0, 1\}$
- Definitions linking them:

		time period t	
		0	1
treatment condition d	0	$Y_{0i,0}$	$Y_{0i,1} = Y_{0i,0} + \lambda_i$
	1	$Y_{1i,0} = Y_{0i,0} + \tau_{i,0}$	$Y_{1i,1} = Y_{0i,0} + \lambda_i + \tau_{i,1}$

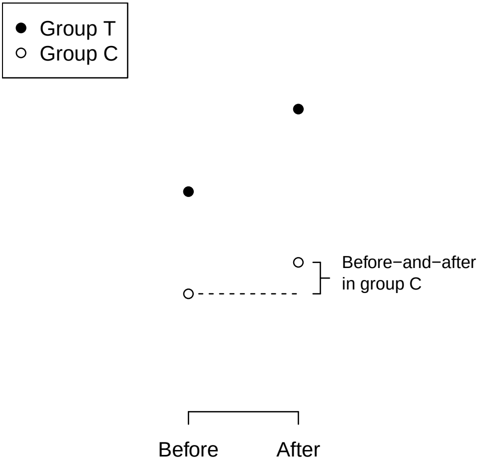
NB: This is notation, not an assumption.

Two groups, two time periods

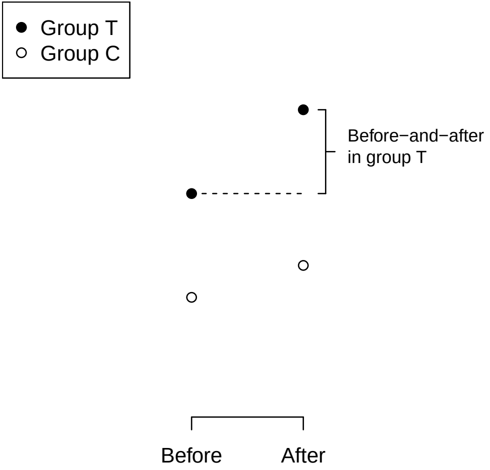


Where g_i denote i's group (treatment or control). For example, $E[Y_{1i,1} \mid g_i = T]$ is the average potential outcome under treatment in period 1 for units in group T.

Before-and-after in group C



Before-and-after in group T



Before-and-after in group C

After-minus-before in group C is

$$E[Y_{0i,1} \mid g_i = C] - E[Y_{0i,0} \mid g_i = C]$$

We use the definitions above to restate in terms of the time trend:

$$\begin{aligned}
&= E[Y_{0i,0} + \lambda_i \mid g_i = C] - E[Y_{0i,0} \mid g_i = C] \\
&= E[\lambda_i \mid g_i = C] + E[Y_{0i,0} \mid g_i = C] - E[Y_{0i,0} \mid g_i = C] \\
&= E[\lambda_i \mid g_i = C] \\
&= \text{Time trend in group C}
\end{aligned}$$

Before-and-after in group T

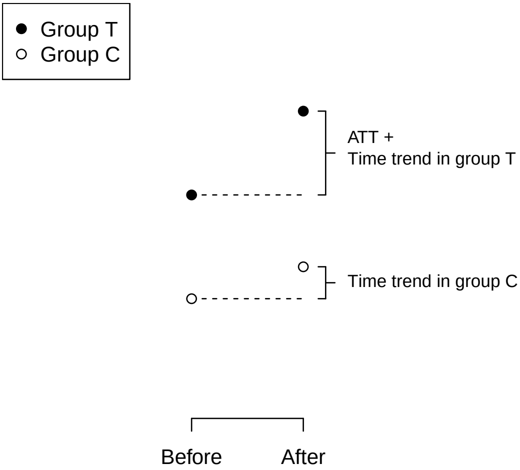
After-minus-before in group T is

$$E[Y_{1i,1} \mid g_i = T] - E[Y_{0i,0} \mid g_i = T]$$

We use the definitions above to restate in terms of time trend and ATE:

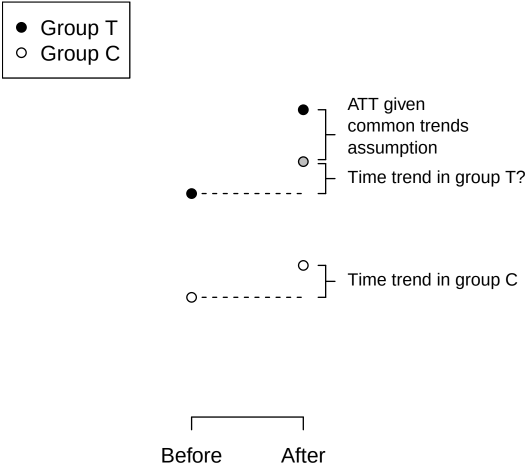
$$\begin{aligned}
&= E[Y_{0i,0} + \lambda_i + \tau_{i,1} \mid g_i = T] - E[Y_{0i,0} \mid g_i = T] \\
&= E[\lambda_i \mid g_i = T] + E[\tau_{i,1} \mid g_i = T] + E[Y_{0i,0} \mid g_i = T] - E[Y_{0i,0} \mid g_i = T \\
&= E[\lambda_i \mid g_i = T] + E[\tau_{i,1} \mid g_i = T] \\
&= \text{Time trend in group T} + \text{ATE in group T}
\end{aligned}$$

Before-and-after in both groups

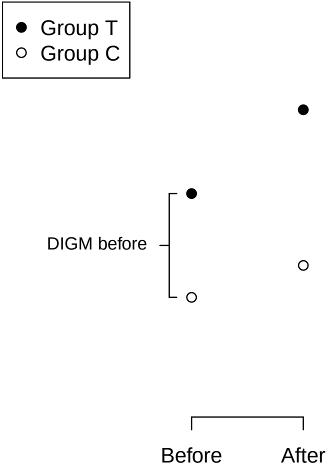


Identification assumption for ATT: common trend in group T and C.

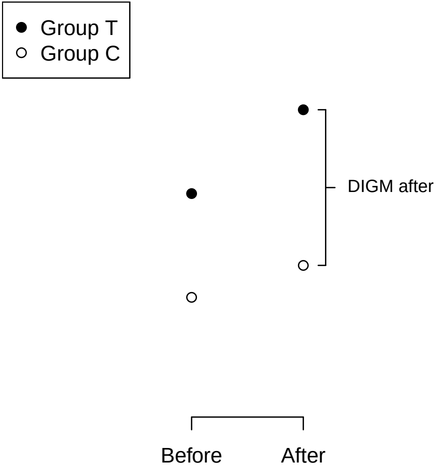
ATT given common trend assumption



Difference in Group Means (DIGM) before



Difference in Group Means (DIGM) after



Difference in Group Means (DIGM) after

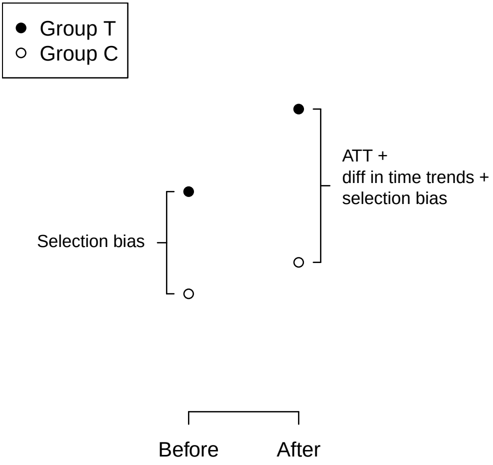
The DIGM at time 1 is

$$E[Y_{1i,1} \mid g_i = T] - E[Y_{0i,1} \mid g_i = C]$$

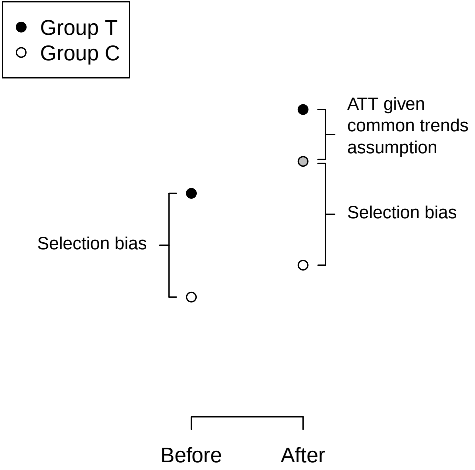
We use the definitions above to restate in terms of time trend, selection bias, and ATE:

$= \mathrm{E}[\mathrm{Y}_{0i,0} + \lambda_i + \tau_{i,1} \mid g_i = T] \quad - \quad E[Y_{0i,0} + \lambda_i \mid g_i = C]$
$= \mathrm{E}[\mathrm{Y}_{0i,0} \mid g_i = T] + E[\lambda_i \mid g_i = T] + E[\tau_{i,1} \mid g_i = T] - E[Y_{0i,0} \mid g_i = C] - E[\lambda_i \mid g_i = C]$
$= \mathrm{E}[\mathrm{Y}_{0i,0} \mid g_i = T] - E[Y_{0i,0} \mid g_i = C] + E[\lambda_i \mid g_i = T] - E[\lambda_i \mid g_i = C] + E[\tau_{i,1} \mid g_i = T]$
= Selection bias + Time trend in group T - Time trend in group C + ATE in group T

Both Difference in Group Means (DIGM)



ATT given common trends assumption



Can the common trends assumption be tested?

No. But common trends in several pre-treatment periods is suggestive.

Dinas et al (2018) on political impact of refugees

- **Question**: Did the influx of refugees in Greece increase support for the right-wing Golden Dawn party in 2015?
- **Treatment**: Large number of refugees arriving in locality
- **Outcome**: Golden Dawn vote share in locality

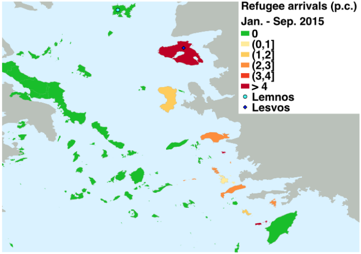
To consider:

- What about a cross-sectional approach? What covariates might help?
- How can we use variation over time in a diff-in-diff?

Dinas et al on the Golden Dawn (2)

Islands that received lots of refugees may vote differently even without the refugee influx.

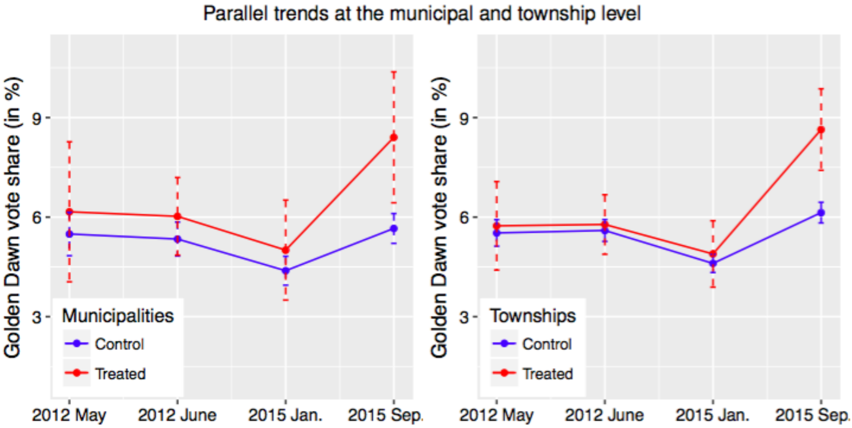
Maybe that difference is constant over time.



Common trends assumption: if they had not received refugees, islands that did receive refugee would have seen the same **change** in support for Golden Dawn as other islands.

To consider: are these other islands really *untreated*?

Dinas et al on the Golden Dawn (3)



Diff-in-diff implementation: method 1

Method 1: group-period interactions

- data structure: two rows for each municipality (elections of Jan. 2015, Sept. 2015)
- `evertr`: 1 for municipalities that received refugees
- `post`: 1 for election after the influx
- `gdper`: support for Golden Dawn

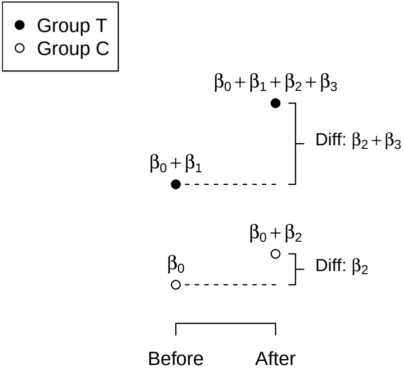
municipality	evertr	post	gdper
Αίγινας	0	0	6.363300
Αίγινας	0	1	7.617789
Αγίου Βασιλείου	0	0	2.714932
Αγίου Βασιλείου	0	1	3.694069
Αγίου Ευστρατίου	0	0	4.878048
Αγίου Ευστρατίου	0	1	5.988024
Αγίου Νικολάου	0	0	3.159049
Αγίου Νικολάου	0	1	4.604597
Αγαθονησίου	1	0	3.278688
Αγαθονησίου	1	1	5.000000
Αγκιστρίου	0	0	6.129032
Αγκιστρίου	0	1	9.981852
Αλοννήσου	0	0	5.727377
Αλοννήσου	0	1	5.976096

Estimating the linear regression:

$$\text{gdper}_{mt} = \beta_0 + \beta_1 \text{evertr}_m + \beta_2 \text{post}_t + \beta_3 \text{evertr}_m \times \text{post}_t + u_{mt}$$

Interpretation of coefficients using method 1

$$\text{gdper}_{mt} = \beta_0 + \beta_1 \text{evertr}_m + \beta_2 \text{post}_t + \beta_3 \text{evertr}_m \times \text{post}_t + u_{mt}$$



Diff-in-diff implementation: method 2

Method 2: unit & time dummies and treatment indicator

We have controlled for group differences with a group dummy.

What about using *municipality* dummies instead?

municipality	evertr	election	treatment	gdper
Αίγινας	0	May12	0	7.9822884
Αίγινας	0	June12	0	7.2771678
Αίγινας	0	Jan15	0	6.3633003
Αίγινας	0	Sept15	0	7.6177893
Αγίου Βασιλείου	0	May12	0	2.5829175
Αγίου Βασιλείου	0	June12	0	4.2843981
Αγίου Βασιλείου	0	Jan15	0	2.7149322
Αγίου Βασιλείου	0	Sept15	0	3.6940687
Αγίου Ευστρατίου	0	May12	0	4.9549551
Αγίου Ευστρατίου	0	June12	0	4.7619047
Αγίου Ευστρατίου	0	Jan15	0	4.8780484
Αγίου Ευστρατίου	0	Sept15	0	5.9880238
Αγίου Νικολάου	0	May12	0	2.8652139
Αγίου Νικολάου	0	June12	0	3.0493212
Αγίου Νικολάου	0	Jan15	0	3.1590488
Αγίου Νικολάου	0	Sept15	0	4.6045966
Αγαθονησίου	1	May12	0	3.5714288
Αγαθονησίου	1	June12	0	4.6875000
Αγαθονησίου	1	Jan15	0	3.2786884
Αγαθονησίου	1	Sept15	1	5.0000000

Estimate the regression:

$$\text{gdper}_{mt} = \beta_1 \text{treatment}_{mt} + \alpha_m + \delta_t + u_{mt}$$

Diff-in-diff implementation: method 2

Method 2: unit & time dummies and a treatment indicator

Regression output:

```
Call:
lm(formula = gdper ~ treatment + as.factor(election) + as.factor(muni) -
    1, data = d[use, ])

Residuals:
    Min      1Q  Median      3Q     Max
-4.5855 -0.5236 -0.0003  0.4404  6.9990

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
treatment                           2.0788     0.3948   5.265 2.79e-07 ***
as.factor(election)Sept15           7.7566     0.5635  13.764  < 2e-16 ***
as.factor(election)Jan15            6.4612     0.5624  11.488  < 2e-16 ***
as.factor(election)June12           7.4365     0.5624  13.222  < 2e-16 ***
as.factor(election)May12            7.5862     0.5624  13.489  < 2e-16 ***
as.factor(muni)Αγίου Βασιλείου     -3.9911     0.7829  -5.098 6.33e-07 ***
as.factor(muni)Αγίου Ευστρατίου    -2.1644     0.7829  -2.765 0.006078 **
as.factor(muni)Αγίου Νικολάου      -3.8906     0.7829  -4.969 1.17e-06 ***
as.factor(muni)Αγαθονησίου         -3.6954     0.7891  -4.683 4.41e-06 ***
as.factor(muni)Αγκιστρίου           4.2533     0.7829   5.433 1.20e-07 ***
as.factor(muni)Αλοννήσου           -2.1973     0.7829  -2.807 0.005357 **
as.factor(muni)Αμαρίου             -4.5633     0.7829  -5.828 1.53e-08 ***
```

[result clipped]

Diff-in-diff implementation: group dummy or unit dummies?

Unit dummies produce lower standard errors, so why not always use them instead of **group dummies**?

Basic diff-in-diff can be done in two kinds of data:

- panel data: same units at several points in time
- repeated cross-section: may not be same units

Cannot use unit dummies with repeated cross-section.

Panel difference-in-difference

$$\mathrm{y}_{it} = \beta_1 \mathrm{treatment}_{it} + \alpha_i + \delta_t + u_{it}$$

Key points:

- β_1 estimated based on **variation in treatment over time within units**
- the only relevant confounders **vary with treatment over time within units**

Panel DiD regression as the **“within”** estimator.

Explaining panel DiD findings

Suppose the **data generating process (DGP)** is

$$Y_{it} = \beta_1 D_{it} + \eta \mathbf{X}_t + \zeta \mathbf{U}_i + \psi \mathbf{V}_{it} + \omega_{it}$$

- $\mathbf{X}_t$ are time-specific variables that affect outcomes for all units the same way (e.g. national economic indicators),
- $\mathbf{U}_i$ are unit-specific variables that are constant over time (e.g. urban/rural character),
- $\mathbf{V}_{it}$ are variables that may vary within units over time (e.g. presence of ambitious council member, local economic situation), and
- ω_{it} is random noise.

In panel-DiD analysis where we estimate $Y_{it} = \beta_1 D_{it} + \alpha_i + \delta_t + \epsilon_{it}$,

- time dummies (δ_t) control for all $\mathbf{X}_t$
- unit dummies (α_i) control for all $\mathbf{U}_i$

so the only possible confounders are $\mathbf{V}_{it}$.

Testing the parallel trends assumption

While we cannot explicitly test the common trends assumptions, we can test for parallel trends in several pre-treatment periods. Regression equation was

$$\mathrm{y}_{it} = \beta_1 \mathrm{treatment}_{it} + \alpha_i + \delta_t + u_{it}$$

but consider adding unit-specific linear time trends:

$$\mathrm{y}_{it} = \sum_{k=-3}^{3} \beta_k \mathrm{treatment}_{i,t+k} + \alpha_i + \delta_t + u_{it}$$

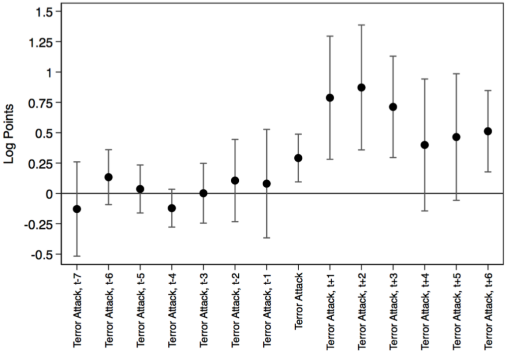
To implement include **lags** and **leads** of treatment: (needs at least 3 years in the pre-period for every unit)

Testing assumptions in panel DiD

Common practice is to visualise the parallel trends plot.

Figure: Ivandic, Kirchmaier and Machin, 2021

Figure 5: Daily Islamophobic Hate Crime and Terror Attacks, Seven Days Leads and Lags, in Logs



DiD with time-varying treatment

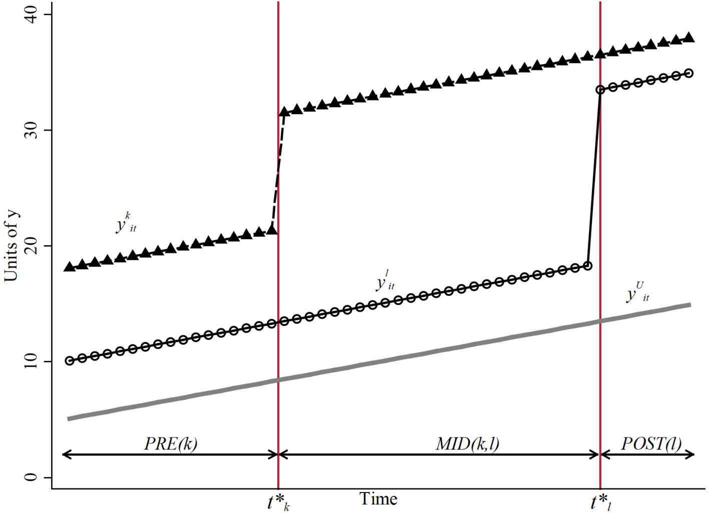
- So far we were considering a case where the treatment was administered in one period
- We can extend this allow for multiple periods and **variation in treatment timing** (staggered timing)
- However, the coefficients from standard TWFE models may not represent a straightforward weighted average of unit-level treatment effects when treatment effects are allowed to be heterogeneous across time or units.
- This is the topic of the recent two-way fixed effects literature

Some recent working papers and literature overviews

- Imai, Kosuke and In Song Kim, “On the use of two-way fixed effects regression models for causal inference with panel data,” Political Analysis, 2021, 29 (3), 405–415.
- de Chaisemartin, Clement and d’Haultfoeuille, Xavier. (2021). ”Two-Way Fixed Effects and Differences-in-Differences with Heterogeneous Treatment Effects: A Survey”. Available at SSRN
- Goodman-Bacon, Andrew. (2021). ”Difference-in-differences with variation in treatment timing.” Journal of Econometrics, Forthcoming
- Roth, J., Sant’Anna, P. H., Bilinski, A., & Poe, J. (2022). What’s Trending in Difference-in-Differences? A Synthesis of the Recent Econometrics Literature. arXiv preprint arXiv:2201.01194.

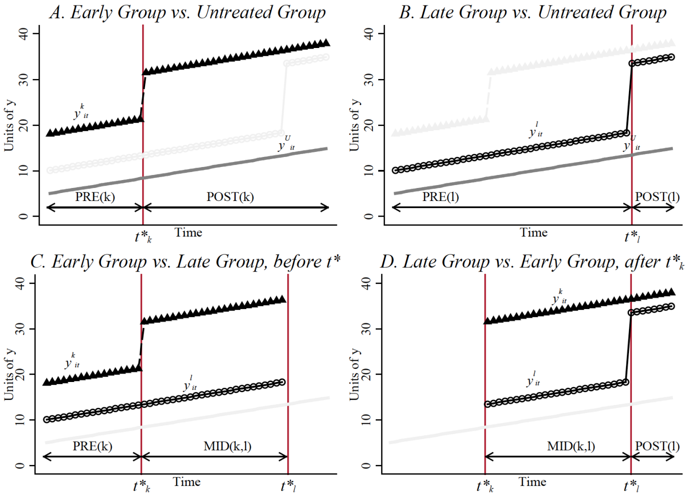
Goodman-Bacon's illustration

Figure 1. Difference-in-Differences with Variation in Treatment Timing: Three Groups



Goodman-Bacon's illustration (2)

Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



Key insights from the Goodman-Bacon paper

- Two-way fixed effect (TWFE) estimate is a weighted average of all possible two-by-two diff-in-diff comparisons
- Weights depend on variance in treatment variable $\rightarrow$ units treated near middle of panel get most weight
- TWFE estimate is sum of
  - Variance-weighted ATTs (across 2×2 diff-in-diffs)
  - Bias due to (variance-weighted) violations of common trends (across 2×2 diff-in-diffs)
  - Bias due to accumulation in treatment effects, because already-treated units act as controls for late-treated cohorts

Insights from advances in TWFE literature

- In short, TWFE regressions make both “clean” comparisons between treated and not-yet- treated units as well as “forbidden” comparisons between units who are both already-treated. When treatment effects are heterogeneous, these “forbidden” comparisons potentially lead to severe drawbacks such as TWFE coefficients having the opposite sign of all individual-level treatment effects due to “negative weighting” problems.
- A common theme is that these new estimators isolate **“clean” comparisons between treated and not-yet-treated groups**, and then aggregate them using user-specified weights to estimate a target parameter of economic interest.

Other considerations around the validity of DiD estimator

- Anticipation effects: We assume that the treatment has no causal effect before its implementation (no anticipation).
- Non-parallel dynamics: Often treatments/programs are targeted based on pre-existing differences in outcomes
  - "Ashenfelter dip": participants in training programs often experience a dip in earnings just before they enter the program (that may be why they participate). Since wages have a natural tendency to mean reversion, comparing wages of participants and non-participants using DiD leads to an upward biased estimate of the program effect
  - Non-parallel dynamics of the outcome variable depends on unobservables
- Long-term effects versus reliability: Parallel trends assumption for DiD is more likely to hold over a shorter time-window. In the long-run, many other things may happen that could confound the effect of the treatment.