Causal Inference
6 - Difference-in-Differences

Benjamin Elsner
benjamin.elsner@ucd.ie

Difference-in-Differences

Angrist & Pischke (2009, ch. 5)
Cunningham (2020, ch. 9)

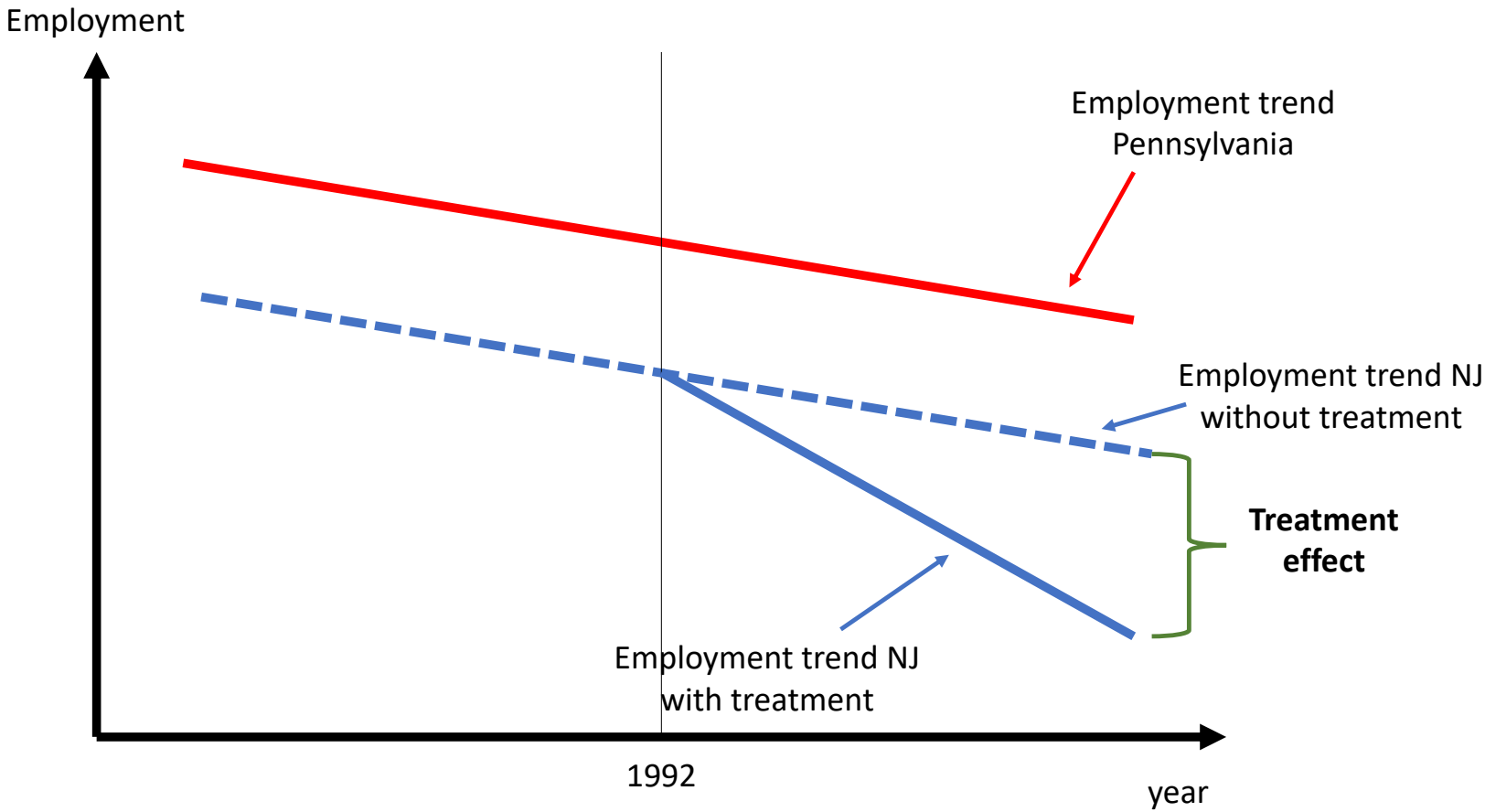
This Lecture

Builds upon the introduction to Diff-in-Diff in Econometrics I

Methodological refinements

- Diff-in-Diff with a continuous treatment
- Semi-parametric Diff-in-Diffs
- Event studies (Staggered adoption designs)

Classic Diff-in-Diff (Card & Krueger, 1994)



The Simple 2×2 Model

The **simple DiD** is a **comparison of two groups before and after**

$$\widehat{\delta}_{kU}^{2\times2} = \left(\overline{y}_k^{\text{post}(k)} - \overline{y}_k^{\text{pre}(k)}\right) - \left(\overline{y}_U^{\text{post}(k)} - \overline{y}_U^{\text{pre}(k)}\right)$$

k	Treated group
U	Untreated group
$pre(k)$	periods before group k was treated
$post(k)$	periods after group k was treated
$\widehat{\delta}_{kU}^{2\times2}$	ATT for group k

What does the estimated parameter $\widehat{\delta}_{kU}^{2\times2}$ map onto?

Simple DiD rewritten as a **conditional expectation**

$$\widehat{\delta}_{kU}^{2\times2} = \Big(E\big[Y_k \mid \text{Post}\big] - E\big[Y_k \mid \text{Pre}\big]\Big) - \Big(E\big[Y_U \mid \text{Post}\big] - E\big[Y_U \mid \text{Pre}\big]\Big)$$

One can show that (in potential outcomes notation), the estimated effect equals

$$\widehat{\delta}_{kU}^{2\times2} = \underbrace{E\big[Y_k^1 \mid \text{Post}\big] - E\big[Y_k^0 \mid \text{Post}\big]}_{\text{ATT}} + \underbrace{\Big[E\big[Y_k^0 \mid \text{Post}\big] - E\big[Y_k^0 \mid \text{Pre}\big]\Big] - \Big[E\big[Y_U^0 \mid \text{Post}\big] - E\big[Y_U^0 \mid \text{Pre}\big]\Big]}_{\text{Non-parallel trends bias in } 2\times2 \text{ case}}$$

Challenge of DiD: Parallel Trends Assumption

We can only obtain an **unbiased estimate of the ATT** if

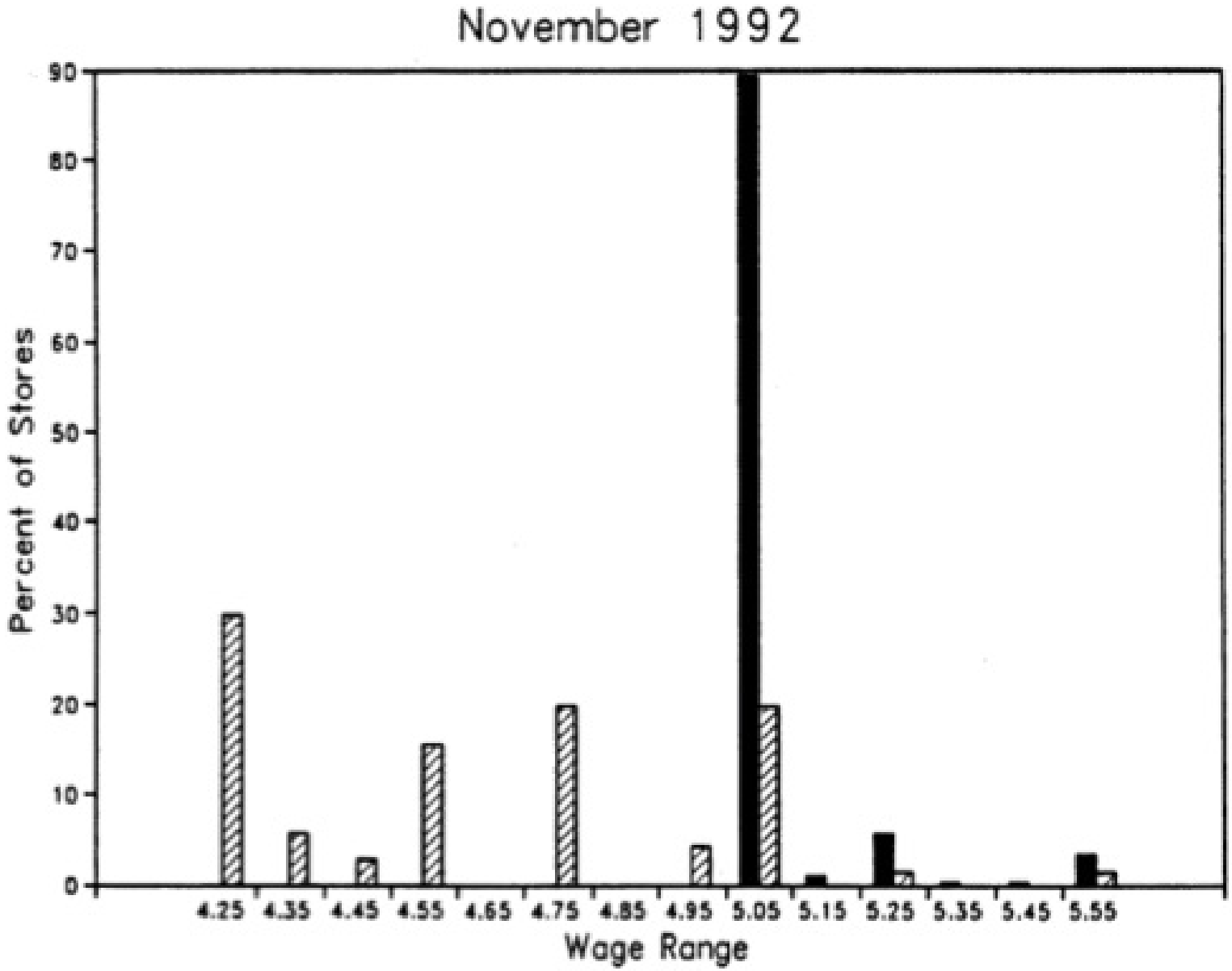
$$\underbrace{\Big[E\Big[Y_k^0 \mid \text{Post}\Big] - E\Big[Y_k^0 \mid \text{Pre}\Big]\Big] - \Big[E\Big[Y_U^0 \mid \text{Post}\Big] - E\Big[Y_U^0 \mid \text{Pre}\Big]\Big]}_{\text{Non-parallel trends bias in } 2 \times 2 \text{ case}} = 0$$

The **Parallel Trends Assumption** is an **identifying assumption**

- We cannot prove that it is true
- We don't observe the counterfactual outcome $Y_k^0 \mid \text{Post}$

The 2×2 DiD in Card & Krueger (1994)

The minimum wage in NJ bites



The 2 × 2 DiD in Card & Krueger (1994)

ATT of interest:

$$\widehat{\delta}^{2\times2}_{NJ,PA} = \underbrace{E\Big[Y^1_{NJ} \mid \text{Post}\Big] - E\Big[Y^0_{NJ} \mid \text{Post}\Big]}_{\text{ATT}}$$
$$+ \Big[\underbrace{E\Big[Y^0_{NJ} \mid \text{Post}\Big] - E\Big[Y^0_{NJ} \mid \text{Pre}\Big]\Big] - \Big[E\Big[Y^0_{PA} \mid \text{Post}\Big] - E\Big[Y^0_{PA} \mid \text{Pre}\Big]\Big]}_{\text{Non-parallel trends bias}}$$

With **constant state and time effects**, this **maps into the regression**

$$Y_{its} = \alpha + \gamma NJ_s + \lambda D_t + \delta(NJ \times D)_{st} + \varepsilon_{its}$$

The 2 × 2 DiD in Card & Krueger (1994)

Variable	Stores by state: 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)

Does $\widehat{\delta}^{2\times2}_{NJ,PA} = 2.76$ mean that the minimum wage raised employment?

Supporting Evidence: Parallel Trends Assumption

Visual inspection is key!

- If you can, plot the raw data (see the Mixtape for some examples)

More than two periods: **check if pre-trends are parallel**

- This is a **common diagnostic check**
- It is neither a necessary nor a sufficient conditions for parallel trends *after* treatment (Freyaldenhoven *et al.*, 2019; Kahn-Lang & Lang, 2020)

Diff-in-Diff with Multiple Groups and Periods

The basic model is often extended, allowing for **multiple groups and periods**,

$$Y_{igt} = \gamma_g + post_t + \delta(D_g \times post_t) + \varepsilon_{igt}$$

units i in groups g and periods t

γ_g	Group fixed effects
$post_t$	Dummy for post-treatment periods

This model still assumes that **all treated groups** are **treated at the same time**.

In many cases, each unit is its own group (e.g. state-level panels in the US) $\Rightarrow$ we will assume this from now on

More General: Two-way FE models

A more general DiD estimator also allows for **differential treatment timing** (*staggered adoption*).

This is typically done through a **two-way fixed effect model**:

$$Y_{it} = \gamma_i + \gamma_t + \delta \, \mathbb{1}(t > t_i^*) + \varepsilon_{it}$$

γ_i	Unit fixed effects
γ_t	Period fixed effects
t_i^*	Period when treatment of unit i starts
$\mathbb{1}(t > t_i^*)$	Dummy $= 1$ if unit i and the period is post treatment

Assumption here: once treated, units remain treated

Flexible Two-way FE DiD (Event Study)

We can also incorporate **$\underline{j}$ leads and $\bar{j}$ lags before/after treatment kicks in (*event*)**

$$Y_{it} = \gamma_i + \gamma_t + \sum_{j=\underline{j}}^{\bar{j}} \delta_j D_{jit} + \varepsilon_{it},$$

The event dummies around the **event window** $[\underline{j}, \bar{j}]$, with $\underline{j} < 0$ are defined as

$$D_{jst} = \begin{cases} \mathbb{1}\left[t \leq \text{ Event }_s + j\right] & \text{if } j = \underline{j} \\ \mathbb{1}\left[t = \text{ Event }_s + j\right] & \text{if } \underline{j} < j < \bar{j} \\ \mathbb{1}\left[t \geq \text{ Event }_s + j\right] & \text{if } j = \bar{j} \end{cases}$$

Note: in the literature you find other specifications called “Event Study”; we will revisit this later

Flexible Two-way FE DiD: event window and binning

A recent paper by Schmidheiny & Siegloch (2020) clarifies two important aspects of event studies

1) The **choice of the event window** can **affect the estimates**

- The length of the event window affects the weights of the OLS estimator

2) Binning of observations **before the first lead and after the last lag?**

- The treatment dummies before the first lead are coded as 0
- Those after the last lag are coded as 1
- Binning is critical for causal identification

Some Comments on Two-way FE DiD

The identification of δ is based on **variation within groups over time**

This model is **extremely popular in applied research**

- 100s of papers on the roll-out of policies across US states
- Appeal of the model: identification assumption easier to defend because of unit fixed effects

But **there is a problem**: it is **not clear how δ (or δ_j) can be interpreted**

- Many researchers interpret it like the 2×2 DiD estimator
- As we will see later, this interpretation is often misleading

Event Study Example (Autor, 2003)

The effect of **EPL on temporary employment**

Exploits court rulings that happened in different states 1979-1995

- Identification comes from differential timing

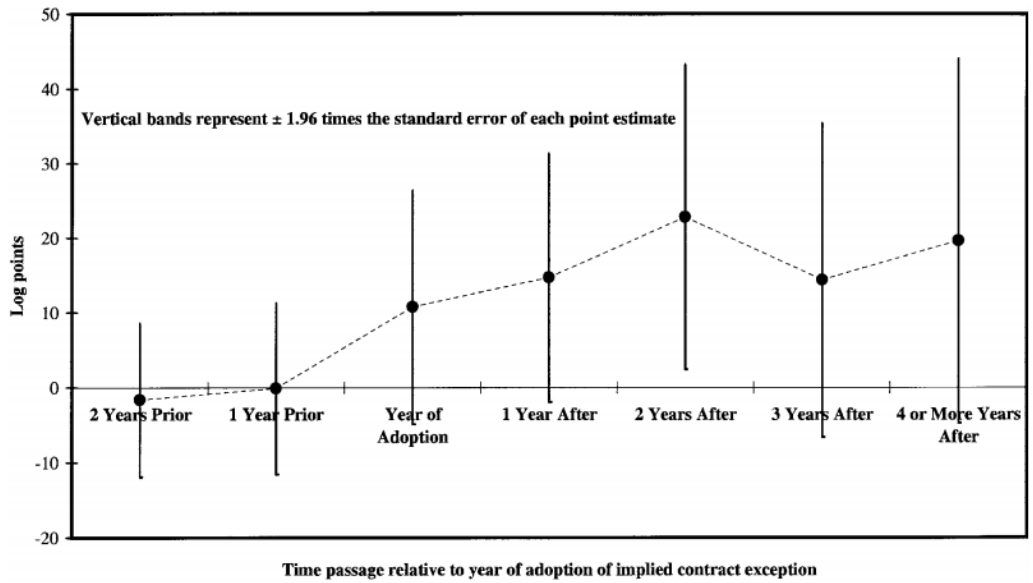


FIG. 3.—Estimated impact of implied contract exception on log state temporary help supply industry employment for years before, during, and after adoption, 1979–95.

DiD with a Continuous Treatment

Most textbook DiD examples are based on a **binary treatment**

But it is also possible to have a **continous treatment**

In this case the **treatment intensity differs between units**

Useful/important: can check for **parallel trends before treatment**

DiD with a Continuous Treatment

Duflo (2001) studies the effect of a **school construction program in Indonesia**

- Between 1973 and 1978 massive school construction program
- Number of schools/population varied across regions and time
- More school were built in areas with low enrolment rates
- Some cohorts were too old to benefit from the program

Identification

- Compare older and younger cohorts
- In areas with different treatment intensity

DiD with a Continuous Treatment

Basic regression in Duflo (2001): compares two cohorts (young and old)

$$S_{ijk} = c_1 + \alpha_{1j} + \beta_{1k} + (P_j * T_i)\gamma_1 + (\mathbf{C_j} * T_i)\delta_\mathbf{1} + \epsilon_{ijk}$$

i: individual; j: region of birth; k: cohort of birth

S_{ijk}	level of schooling
α_{1j}	region of birth FE
β_{1k}	birth cohort FE
P_j	program intensity
T_i	treatment dummy (young)
C_j	region-specific variables

DiD with a Continuous Treatment

Flexible DiD

$$S_{ijk} = c_1 + \alpha_{1j} + \beta_{1k} + \sum_{l=2}^{23}(P_j * d_{il})\gamma_{1l} + \sum_{l=2}^{23}(\mathbf{C_j} * d_{il})\delta_{\mathbf{1l}} + \epsilon_{ijk}$$

i: individual; j: region of birth; k: cohort of birth

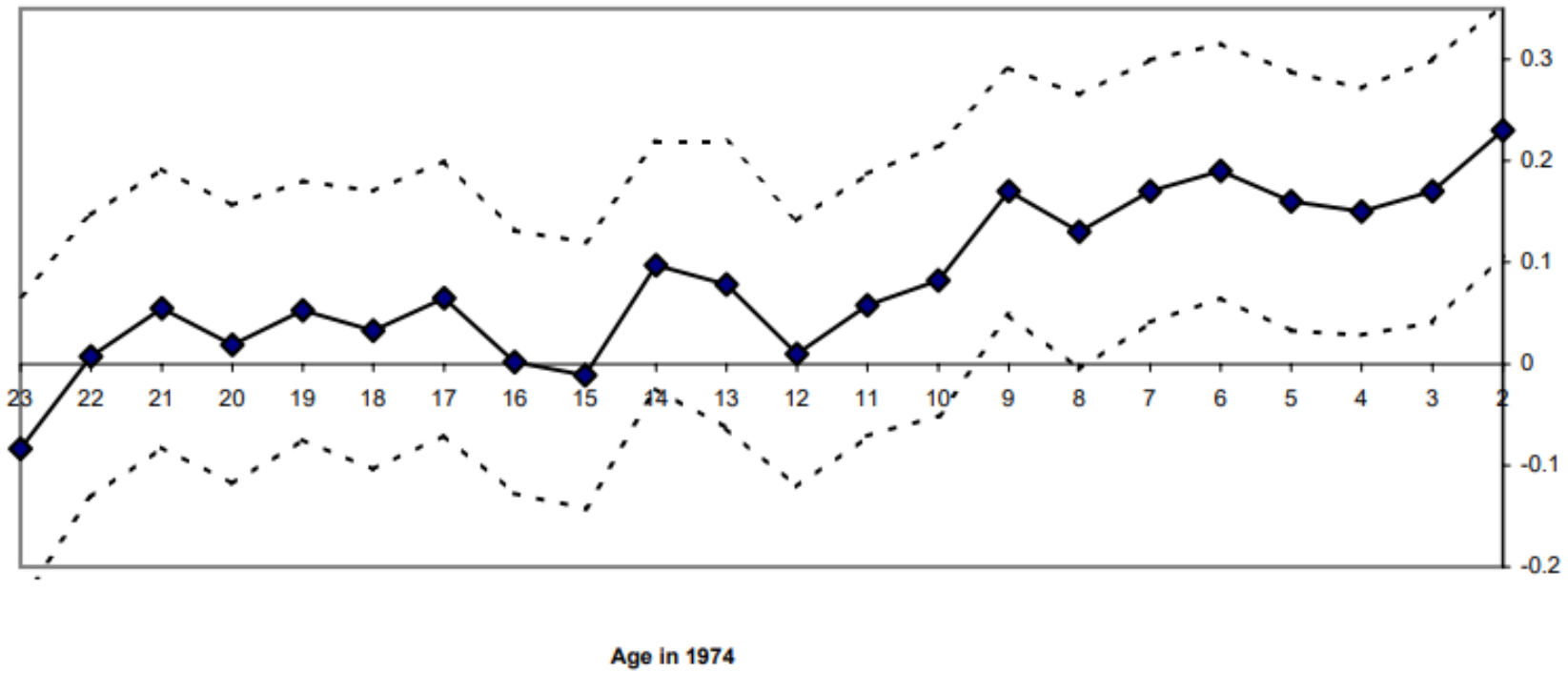
d_{il} age t dummies

Identification:

- Comes from across regions across cohorts (i.e. relative to older cohorts, younger cohorts in some regions had more schools to go to than in other regions)
- Assumption: *within a region*, the assignment of treatment across groups was as good as random

DiD with a Continuous Treatment

Main result in Duflo (2001)



Event Study Design: Lafortune *et al.* (2018)

Paper evaluates the school finance reforms (SFR) in the US in the 1990s

After court rulings, **multiple states changed their funding model**

- initially large heterogeneity in school finances
- 1980s SFR: **equity-based**; all students should get the same resources
- 1990s SFR: **adequacy rules**: more resources for low-income districts

They study the **effect of the 1990s school reforms** on **standardized test scores**

Identification

They use an **event study design**

- between 1990 and 2011 there were 64 reforms in 26 states
- they compare standardized test scores within states before and after the reform

Diff-in-diff logic: did test scores increase in states with a reform *relative* to states without a reform?

Idenfitication assumptions:

- in absence of the reform, test scores would have been the same in states with and without a reform
- the timing of the reform is as good as random

Simple Event Study Design

$$\theta_{st} = \delta_s + \kappa_t + 1(t > t_s^*)\beta^{jump} + \varepsilon_{st}$$

s: state; t: year

t_s^*	time of the event
θ_{st}	outcome (test scores, etc)
δ_s, κ_t	state and year effects

β^{jump} represents the **difference-in-differences estimator** (state s relative to all other states)

Parametric Event Study Design

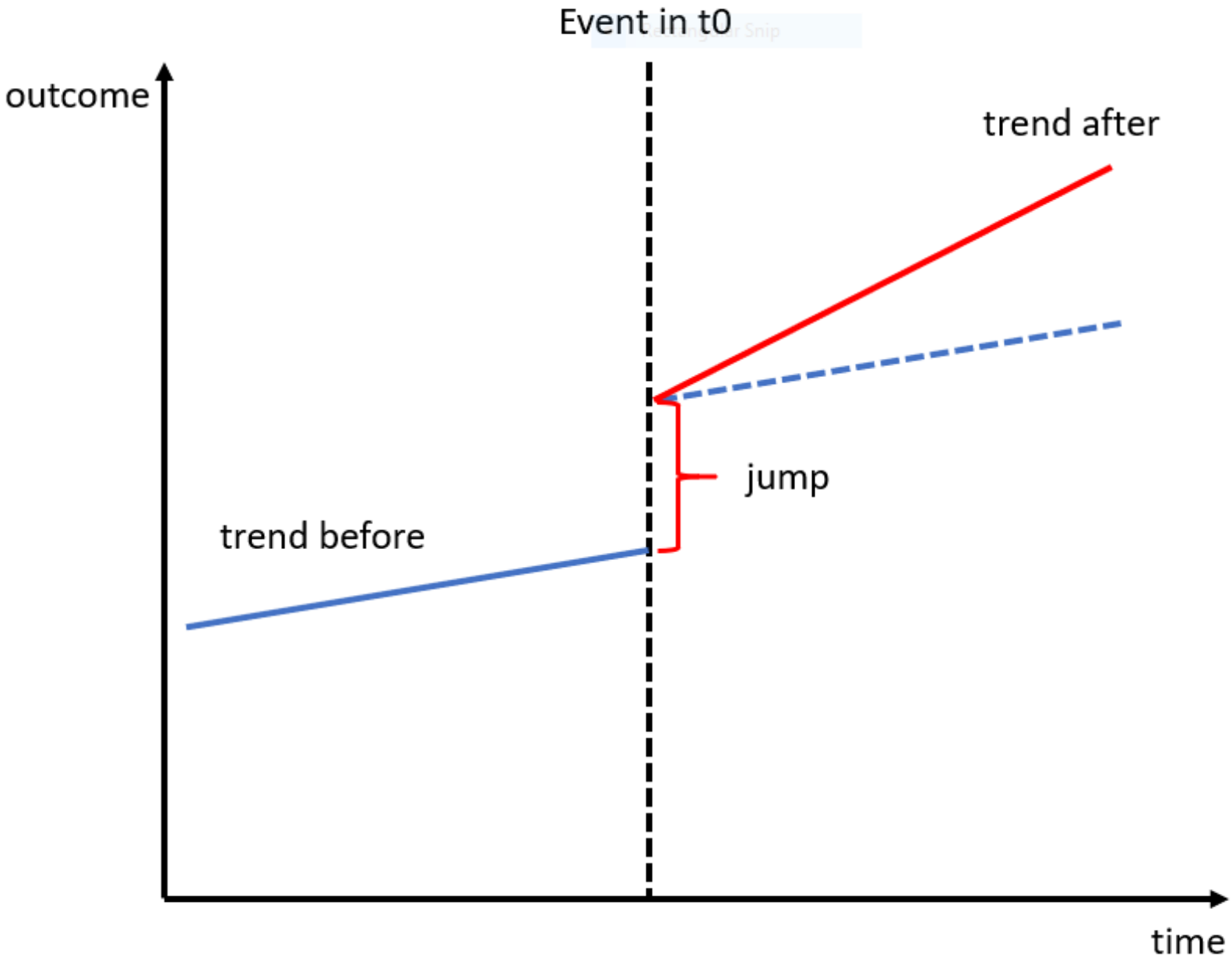
Lafortune *et al.* (2018) develop a useful parametric extension to the standard diff-in-diff model

They consider three parameters

- the **trend** in the outcome **before the event**
- the **jump at the time of the event**
- the **trend after the event**

$$\theta_{st} = \delta_s + \kappa_t + 1(t > t_s^*)\beta^{jump} + 1(t > t_s^*)(t - t_s^*)\beta^{phasein} + (t - t_s^*)\beta^{trend} + \varepsilon_{st}$$

Parametric Event Study Design



Non-parametric Event Study Design

They also estimate **(standard) non-parametric event study designs**

$$\theta_{st} = \delta_s + \kappa_t + \sum_{r=k_{min}, r\neq 0}^{k_{max}} 1(t = t_s^* + r)\beta_r + \varepsilon_{st}$$

The dummy variables $1(t = t_s^* + r)$ represent **leads and lags for each event**

The **coefficients of interest** are β_r

- difference in the outcome relative to the event year
- ...relative to states without an event

Parametric vs. Non-parametric Designs

The **non-parametric design** has the **advantage of being flexible**

- **no restriction of trends** before and after the event
- allows to study **dynamic adjustments**

Disadvantage: non-parametric designs are demanding on the data

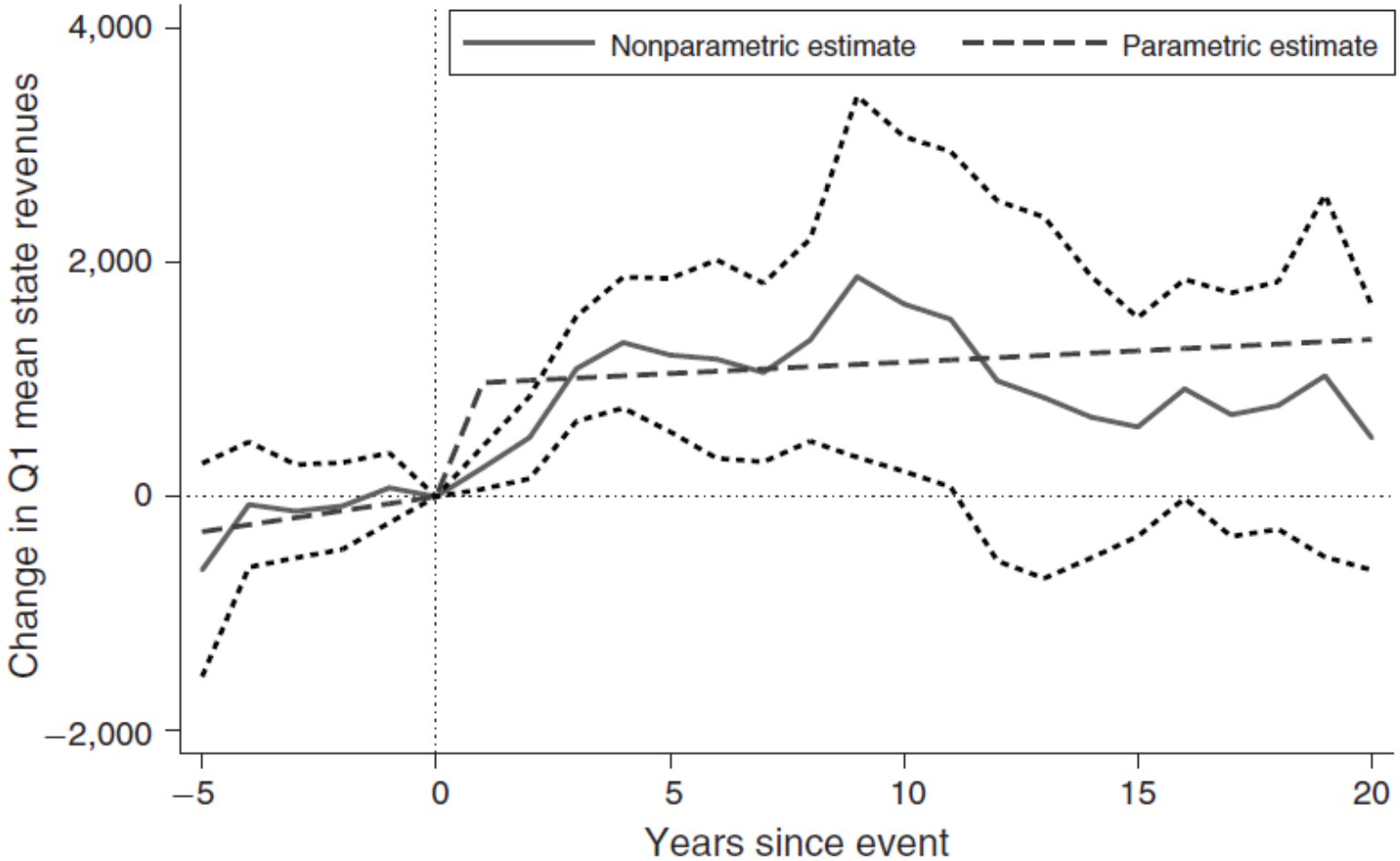
- need to estimate many lead and lag coefficients

Advantage of **stepwise parametric approach**

- only need to estimate three parameters
- easy to put a number on the effects
- enough to model trend after the event

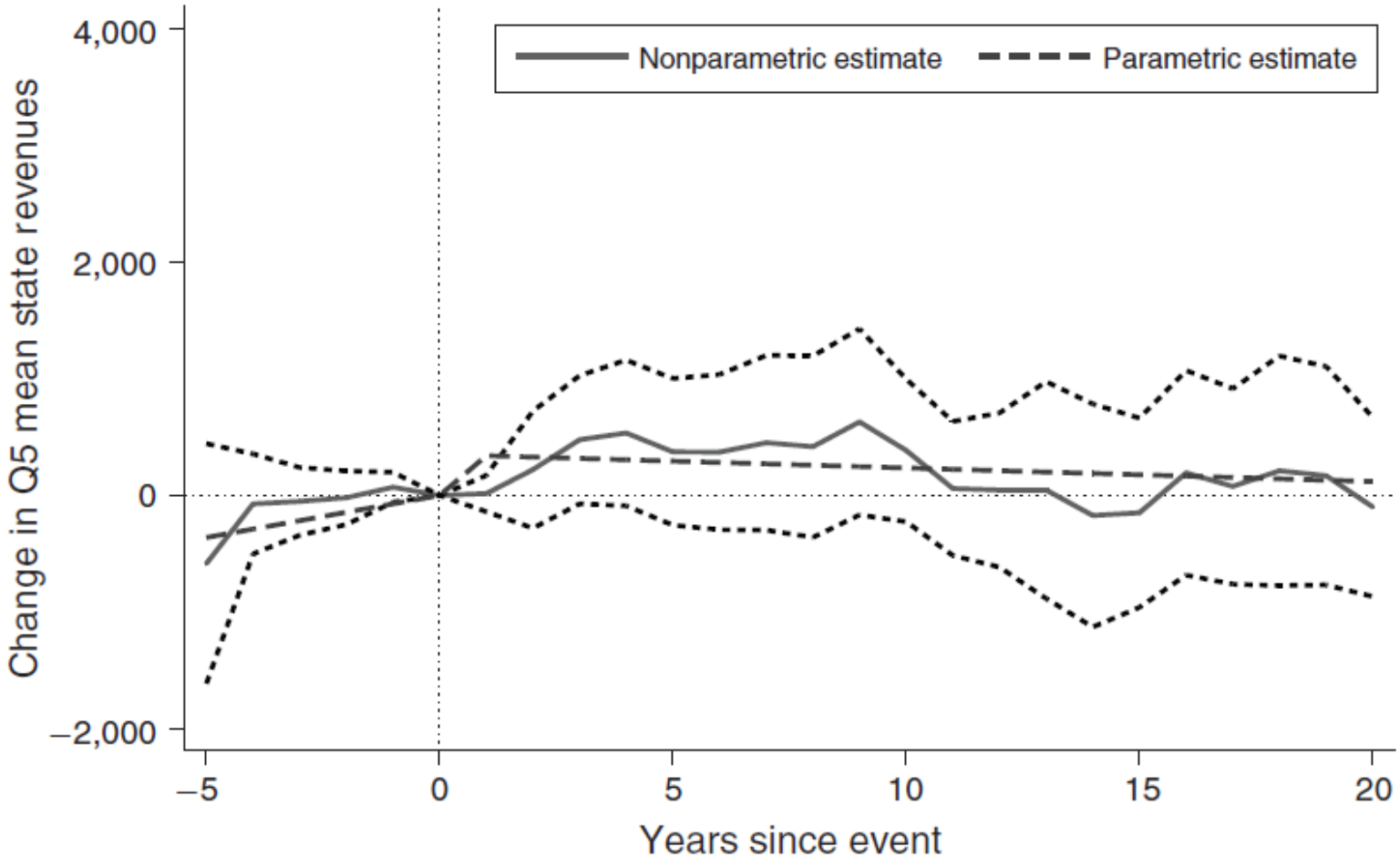
Effect on School Finances

Districts in lowest 20%



Effect on School Finances

Districts in highest 20%



Effect on Test Scores

The parametric results are useful in regression tables

TABLE 5—EVENT STUDY ESTIMATES OF EFFECTS OF SCHOOL FINANCE REFORMS ON STUDENT ACHIEVEMENT

	Slopes		Q1	Q5	Q1 − Q5	
	(1)	(2)	(3)	(4)	(5)	(6)
Post event × years elapsed	−0.011	−0.010	0.007	−0.001	0.008	0.013
	(0.004)	(0.003)	(0.003)	(0.003)	(0.004)	(0.006)
Trend	0.001					−0.006
	(0.003)					(0.005)
Post event	0.001					0.011
	(0.023)					(0.024)
Observations	1,498	1,498	1,509	1,506	1,504	1,504
p, total event effect = 0	0.02	0.01	0.02	0.69	0.04	0.07
State fixed effects	X	X	X	X	X	X
Subject-grade-year fixed effects	X	X	X	X	X	X

Lessons from Lafortune *et al.* (2018)

Clean and **intuitive event study design**

Shows the **usefulness of parametric and non-parametric methods**

Quantifies the policy effect $\Rightarrow$ **useful for cost-benefit analysis**

More on Supporting Evidence of the Identification Assumptions

Parallel pre-trends are one (commonly used) diagnostic test

It is also helpful to run **placebo tests**

- There should be no effect on units that are **plausibly unaffected by treatment**
- If there is an effect, this may indicate a violation of parallel trends
- Your estimator is probably picking up some underlying trends

Plausibly unaffected units can also form an additional control group in a **triple difference design (DDD)**

Triple Differences

Idea: **compare two difference-in-differences**

1. Units **plausibly affected** by treatment (some treated but not others)
2. Units **plausibly unaffected** by treatment (some "treated" but not others)

Minimum wage example (NJ & PA):

1. Workers in fast-food restaurants are **plausibly affected** by a MW change
2. Workers with a college degree are **plausibly unffected** by a MW change (under assumptions...)

$\Rightarrow$ allows us to **difference out state-specific shocks**

Triple Differences: Regression Model

$$Y_{ijt} = \alpha + \underbrace{\beta_1 \tau_t + \beta_2 \delta_j + \beta_3 D_i}_{\text{FE for period, treated DD and unit}} + \underbrace{\beta_4 (\delta \times \tau)_{jt} + \beta_5 (\tau \times D)_{ti} + \beta_6 (\delta \times D)_{ij}}_{\text{Two-way interactions}} + \underbrace{\beta_7 (\delta \times \tau \times D)_{ijt}}_{\text{Triple Difference}} + \varepsilon_{ijt}$$

In the MW example, β_7 represents the difference between

- Fast-food workers in NJ (treated) vs PA (non-treated) $\Rightarrow$ **Diff 1**
- Before and after the change $\Rightarrow$ **Diff 2**
- Relative to the differences 1 and 2 among unaffected workers $\Rightarrow$ **Diff 3**

DDD Example: Immigration and Native Labor Supply

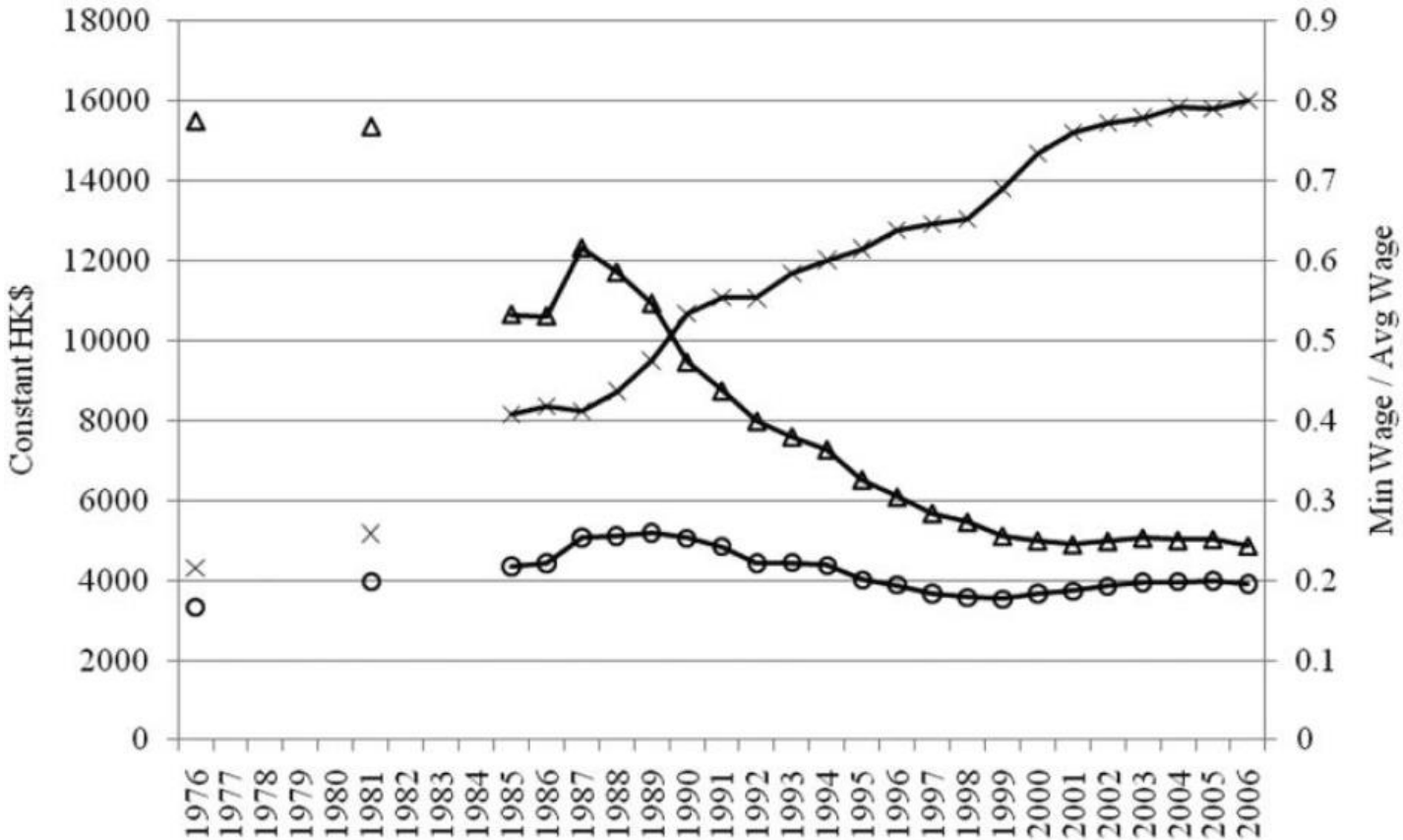
Cortés & Pan (2013): domestic outsourcing

Question: does (low-skilled) immigration affect native labor supply?

- Recruitment policy of foreign domestic workers
- in Hong Kong in the 1970s
- High inflow of FDWs, mainly from Philippines and Thailand
- Taiwan serves as “control country”

DDD Example: Immigration and Native Labor Supply

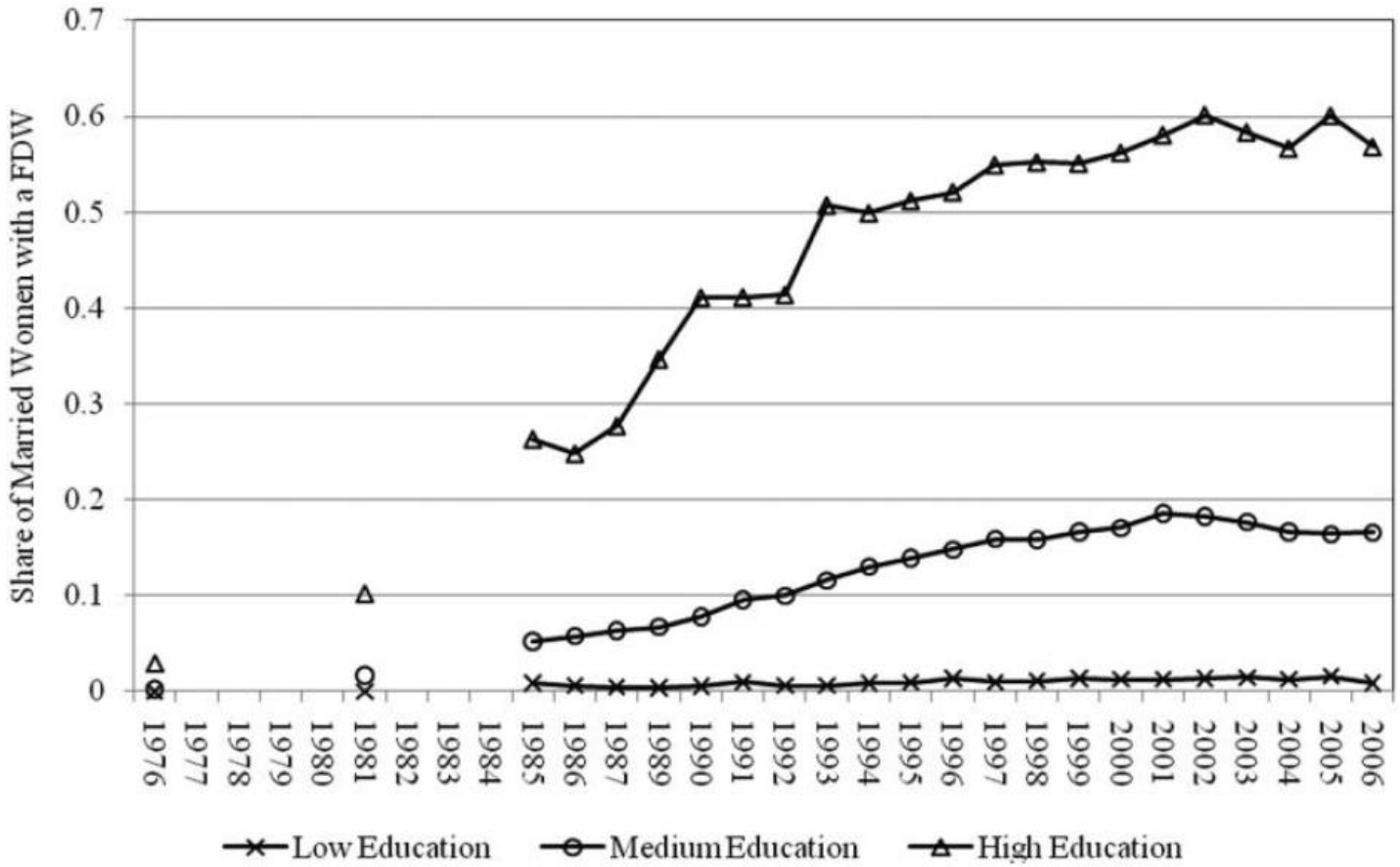
Cortés & Pan (JOLE, 2013): FDW became cheaper



X: average wage, triangle: relative price of FDW

DDD Example: Immigration and Native Labor Supply

Cortés & Pan (JOLE, 2013): increase in FDW



DDD Example: Immigration and Native Labor Supply

Identification through triple differences

- Difference 1: women with and without school-age children
- Difference 2: before and after the reform
- Difference 3: difference in 1 and 2 between Hong Kong (treated) and Taiwan (non-treated)

DDD Example: Immigration and Native Labor Supply

Cortés & Pan (JOLE, 2013): domestic outsourcing

	All Women			
	(1)	(2)	(3)	(4)
HK × child05 (base period 1978–84)	−.035***	−.042***	−.053***	−.062***
	(.007)	(.004)	(.005)	(.011)
	[.042]	[.019]	[.023]	[.034]
HK × child05 × p85–87	.017	.043***	.034**	.030*
	(.012)	(.012)	(.013)	(.018)
	[.014]	[.020]	[.018]	[.030]
HK × child05 × p89–93	.034*	.056***	.064***	.040***
	(.018)	(.015)	(.017)	(.014)
	[.028]	[.031]	[.017]	[.030]
HK × child05 × p94–98	.110***	.114***	.135***	.076***
	(.009)	(.005)	(.008)	(.012)
	[.040]	[.035]	[.016]	[.038]
HK × child05 × p99–02	.126***	.112***	.134***	.129***
	(.007)	(.004)	(.006)	(.012)
	[.058]	[.020]	[.006]	[.036]
HK × child05 × p03–06	.099***	.076***	.085***	.116***
	(.011)	(.010)	(.012)	(.016)
	[.068]	[.015]	[.013]	[.031]

Triple Differences

Cortés & Pan (2013) is a good **example for the use of triple differences**

- DiD not invalid because of differential time trends between treated and control units
- DDD allows them to difference out these time trends

General advice:

- Not advisable: DDD is the main result
- Much harder to interpret compared to DD
- Best to show DD first and use DDD as a robustness test
- (Probably) not suitable for staggered adoption designs

Conceptual Problem with DiD: What Do We Compare with What?

The canonical DiD model is a **two-way FE regression**

$$y_{it} = \alpha_{i\cdot} + \alpha_{\cdot t} + \beta^{DD} D_{it} + e_{it}$$

$D_{it} = 1$ if a unit has been treated in period t or before

Textbook DiD: units are **treated at the same time**

But when the **treatment timing differs between units**...

- As is the case in many studies?
- **What do we compare** with what?

Conceptual Problem with DiD: What Do We Compare with What?

Big problem: researchers often **interpret β^{DD} just like in the 2×2 case**

- This interpretation is **akin to the ATT**

Several recent papers highlight these **problems and provide solutions**

- Goodman-Bacon (2018), Callaway & Sant'Anna (2020), de Chaisemartin & D'Haultfoeuille (2019), Abraham & Sun (forthcoming), ...
- **General problem I**: β^{DD} is a **(very strange) weighted average**, s.t. $\beta^{DD} \neq ATT$
- **General problem II**: **Treatment effects may be heterogeneous** across groups and over time

The Bacon Decomposition

Goodman-Bacon (2018) proposes a **decomposition** of β^{DD} for the **staggered adoption design**

Paper points to several **challenges** for estimation and interpretation

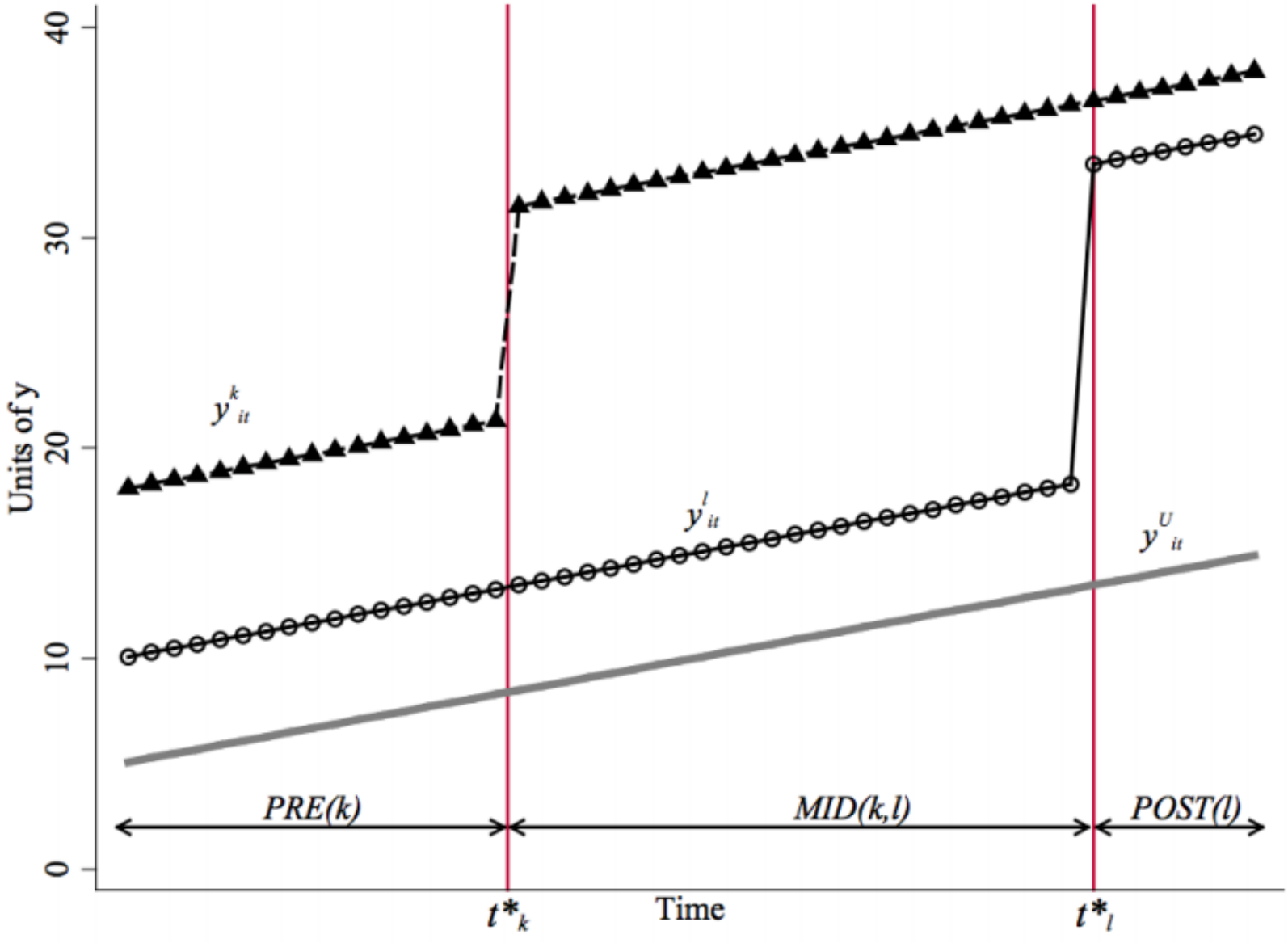
- Late adopters are a control group for early adopters
- But early adopters are also a control group for late adopters
- Heterogeneous treatment effects may lead to severe bias
- Estimate depends on many factors: variation in treatment, group sizes, etc

Decomposition shows:

- β^{DD} is a (strange) weighted average of 2×2 comparisons
- We can only estimate the ATT under restrictive assumptions

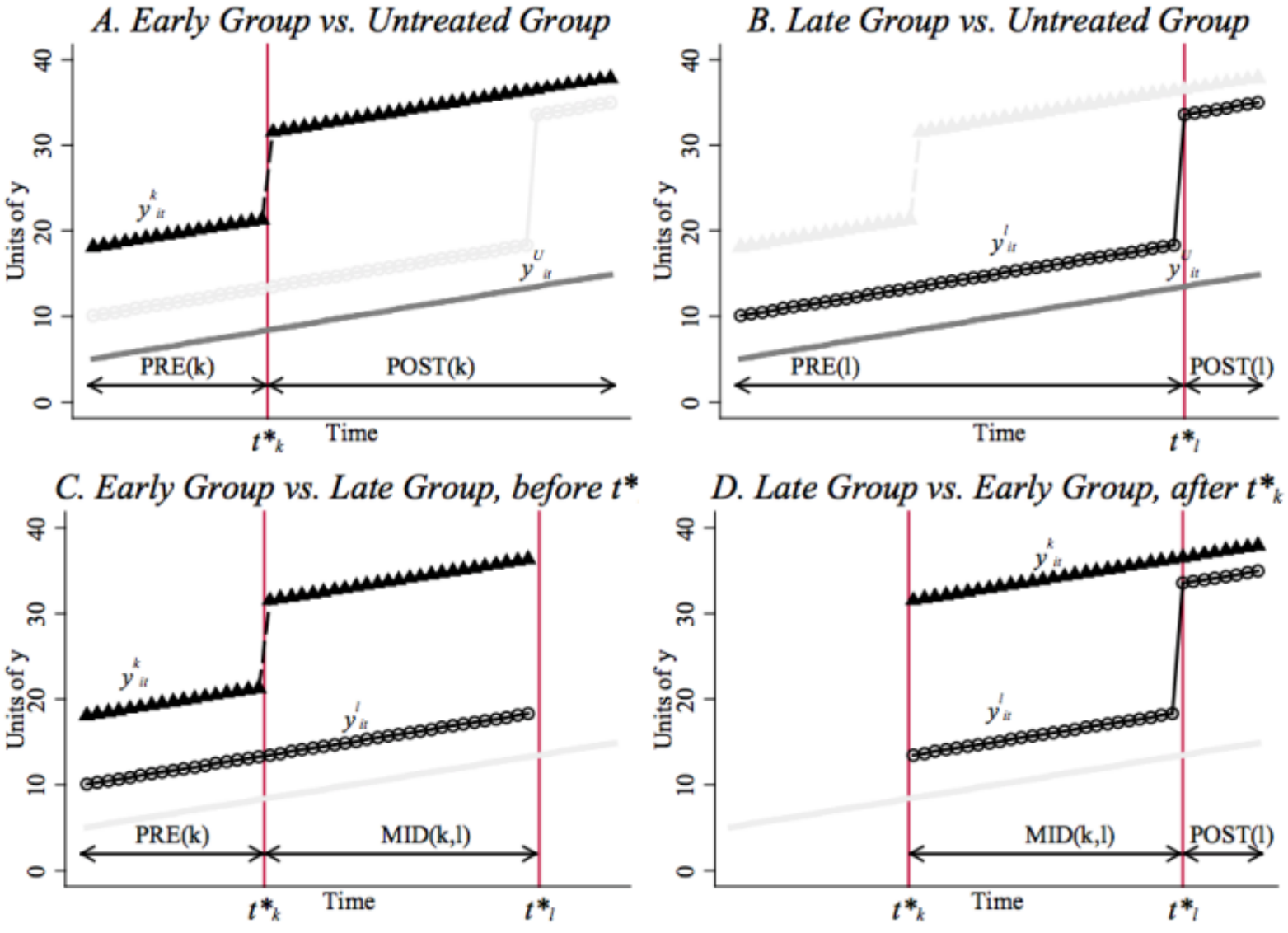
Difference in Treatment Timing

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



2 × 2 Comparisons of DiDs

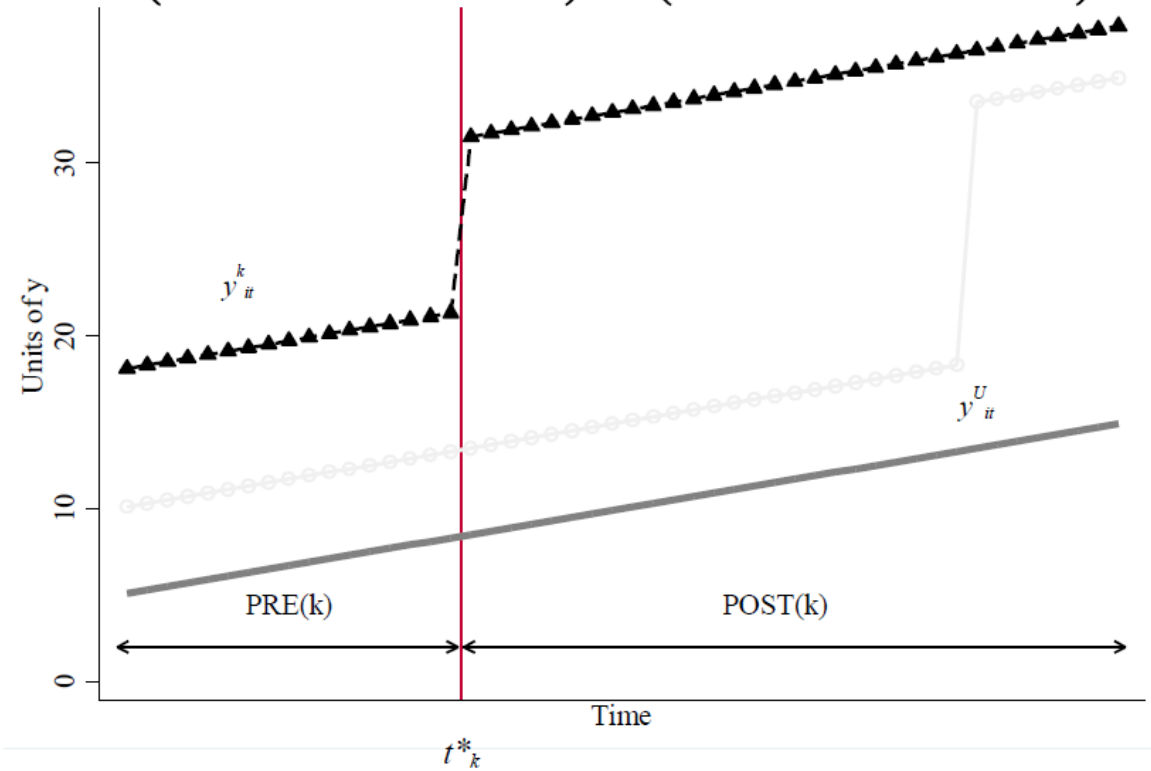
Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



Treated vs. Untreated I

$$\widehat{\beta}_{kU}^{2x2}$$

$$\hat{\beta}_{kU}^{2x2} = \left(\bar{y}_k^{POST(k)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_U^{POST(k)} - \bar{y}_U^{PRE(k)}\right)$$



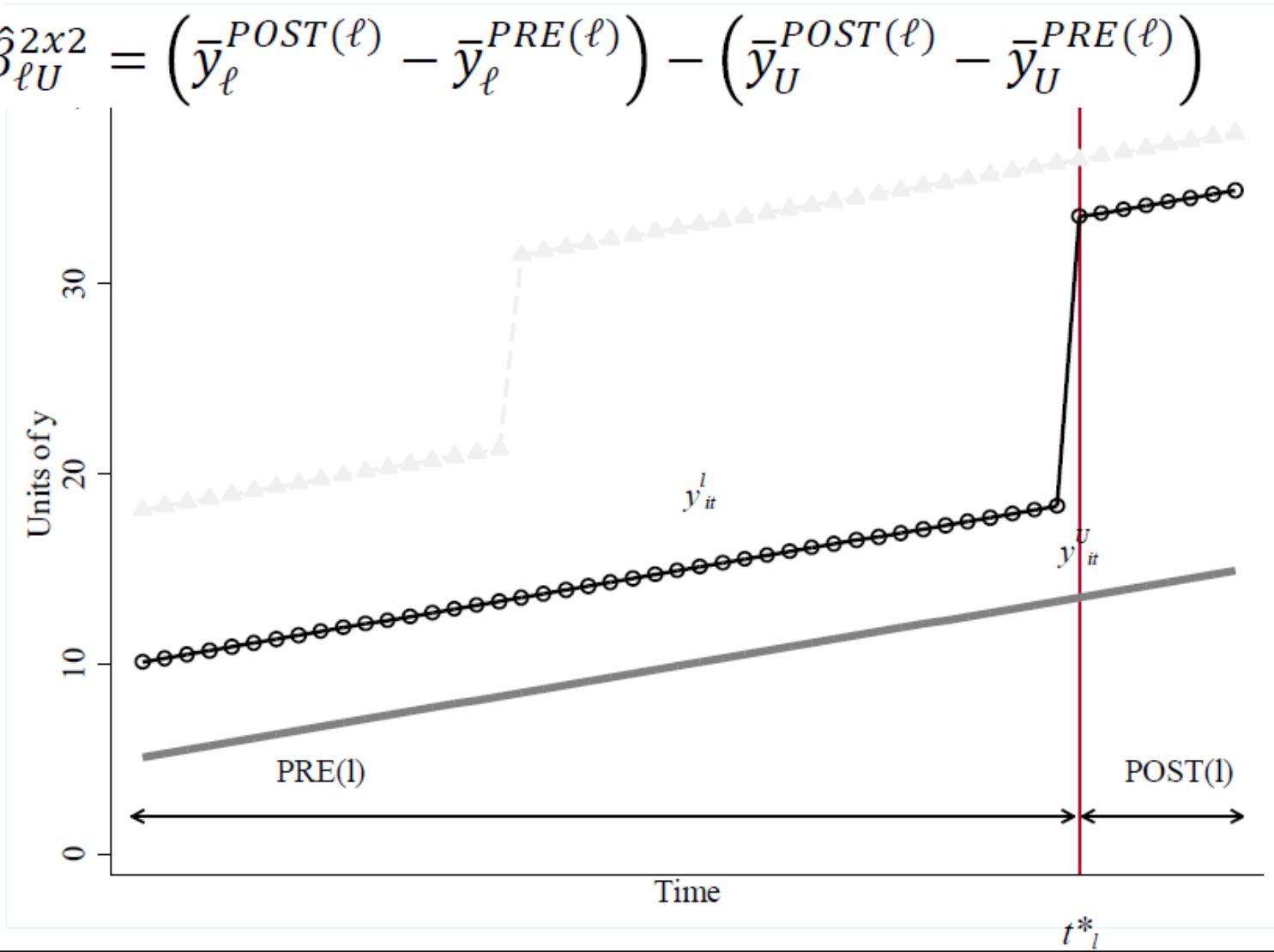
Each **treatment group considered separately**:

- identified by **canonical DiD model**
- Compares units over the entire sample period

Treated vs. Untreated II

$$\widehat{\boldsymbol{\beta}}_{\ell U}^{2x2}$$

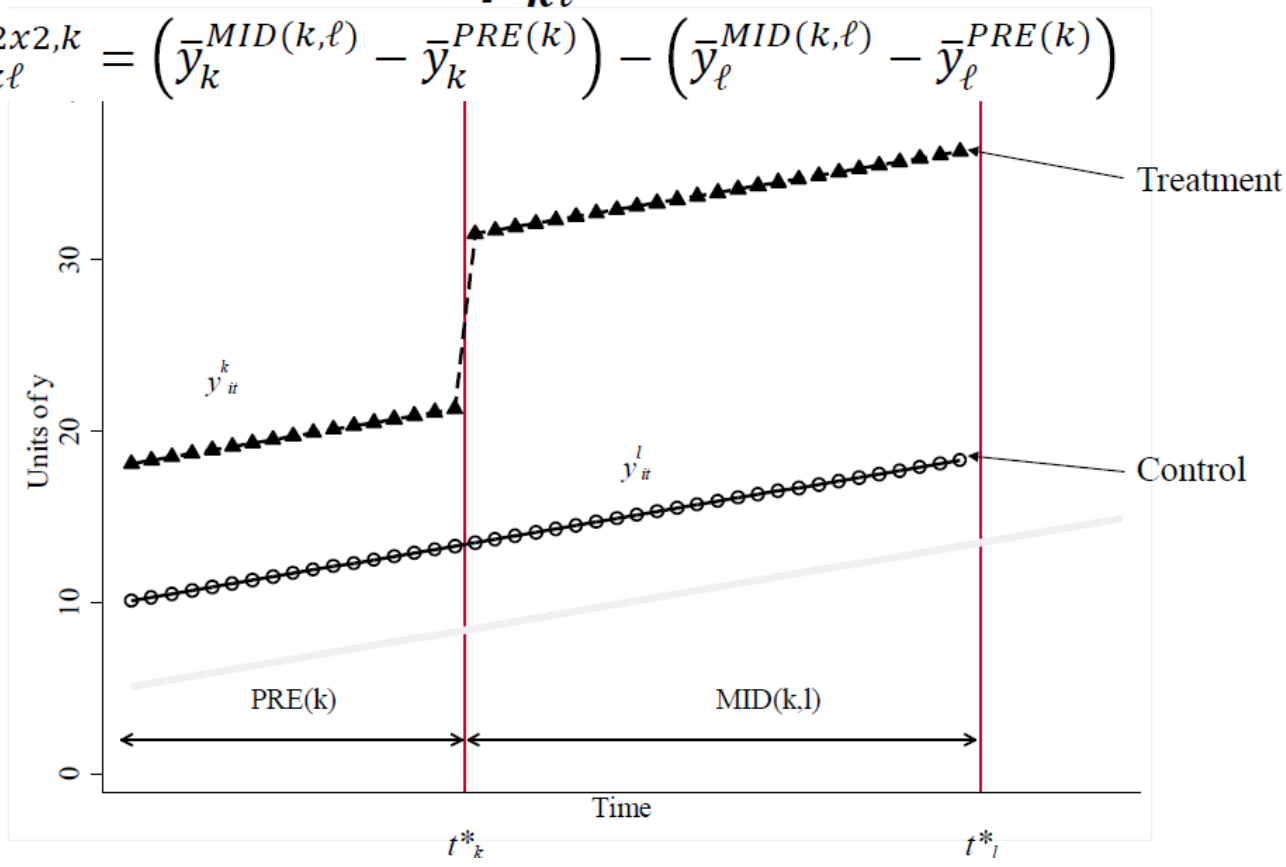
$$\hat{\beta}_{\ell U}^{2x2} = \left(\bar{y}_{\ell}^{POST(\ell)} - \bar{y}_{\ell}^{PRE(\ell)}\right) - \left(\bar{y}_{U}^{POST(\ell)} - \bar{y}_{U}^{PRE(\ell)}\right)$$



Treated at Different Times I

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{k\ell}}^{\boldsymbol{2x2,k}}$$

$$\hat{\beta}_{k\ell}^{2x2,k} = \left(\bar{y}_k^{MID(k,\ell)} - \bar{y}_k^{PRE(k)}\right) - \left(\bar{y}_\ell^{MID(k,\ell)} - \bar{y}_\ell^{PRE(k)}\right)$$



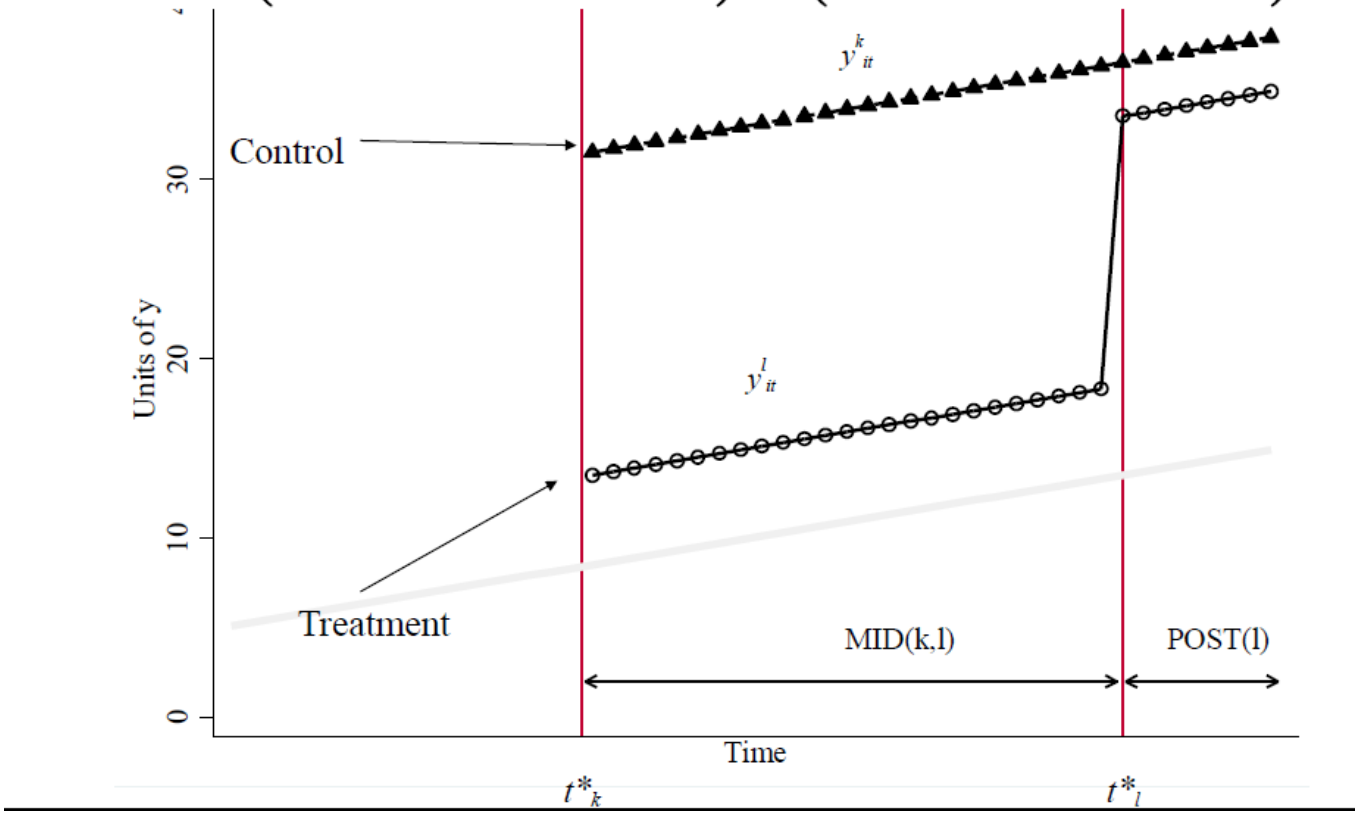
No untreated group

- identified by **difference in timing only**
- think about it as a **level shift**

Treated at Different Times II

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{2x2,\ell}}_{\boldsymbol{k\ell}}$$

$$\hat{\beta}^{2x2,\ell}_{k\ell} = \left(\bar{y}^{POST(\ell)}_{\ell} - \bar{y}^{MID(k,\ell)}_{\ell}\right) - \left(\bar{y}^{POST(\ell)}_{k} - \bar{y}^{MID(k,\ell)}_{k}\right)$$



Late vs. early after the early group has been treated

All the 2×2 Pairs

$$\text{treated vs. untreated } \widehat{\delta}^{2\times2}_{kU} = \left(\overline{y}^{\text{post}(k)}_{k} - \overline{y}^{\text{pre}(k)}_{k}\right) - \left(\overline{y}^{\text{post}(k)}_{U} - \overline{y}^{\text{pre}(k)}_{U}\right)$$

$$\text{early vs. late } \widehat{\delta}^{2\times2}_{kl} = \left(\overline{y}^{mid(k,l)}_{k} - \overline{y}^{\text{pre}(k)}_{k}\right) - \left(\overline{y}^{mid(k,l)}_{l} - \overline{y}^{\text{pre}(k)}_{l}\right)$$

$$\text{late vs. early } \widehat{\delta}^{2\times2}_{lk} = \left(\overline{y}^{\text{post}(l)}_{l} - \overline{y}^{mid(k,l)}_{l}\right) - \left(\overline{y}^{\text{post}(l)}_{k} - \overline{y}^{mid(k,l)}_{k}\right)$$

The Decomposition Theorem

Goodman-Bacon (2018) shows that the **two-way FE model is a weighted average of all the 2×2 DiDs**

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}^{2\times2}_{kU} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}^{2\times2,k}_{kl} + (1 - \mu_{kl}) \widehat{\delta}^{2\times2,l}_{kl} \right]$$

s_{kU}	weight of treated vs. untreated group
s_{kl}	weight of early vs. late adopters
μ_{kl}	relative weight of comparison early-late vs. late-early

A Look at the Weights

$$s_{ku} = \frac{n_k n_u \overline{D}_k (1 - \overline{D}_k)}{\widehat{Var}(\tilde{D}_{it})}$$

$$s_{kl} = \frac{n_k n_l (\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))}{\widehat{Var}(\tilde{D}_{it})}$$

$$\mu_{kl} = \frac{1 - \overline{D}_k}{1 - (\overline{D}_k - \overline{D}_l)}$$

n_k, n_u, n_l	sample sizes of each 2×2 pair
$\overline{D}_k(1 - \overline{D}_k)$	Within-2×2-group variance in treatment
$(\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l))$	Within-2×2-group variance in treatment
μ_{kl}	Share of time spent under treatment early vs. late
$\widehat{Var}(\tilde{D}_{it})$	Overall variance in treatment (conditional on FE)

It's all about Weights

The weights depend on the **time spent under treatment** $\bar{D}$

- $\bar{D}(1 - \bar{D})$ is maximized at $\bar{D}^* = 0.5$

This has **profound implications for the interpretation of** $\widehat{\delta}^{DD}$

- Units that are treated early or late receive very little weight in the estimation
- The estimate depends on the sample period...
- ...add more data points before or after, and $\bar{D}(1 - \bar{D})$ will change!

It's all about Weights

Interpreting the **weights** is **less clear for earlier-later comparisons**

The **treatment variance** is $V = (\overline{D}_k - \overline{D}_l)(1 - (\overline{D}_k - \overline{D}_l)$

- the earlier group is under treatment for $\overline{D}_k$ periods
- the late group is under treatment for $\overline{D}_l$ periods

Numerical example: $\overline{D}_k = 67\%$, $\overline{D}_l = 15\%$

- $V = 0.52 \times 0.48 = 0.2496$
- This is close to the maximum variance of $0.5^2 = 0.25$

Your (oh-so-simple and transparent) DD estimator gives the **greatest weight** to groups whose treatment periods are **50% of the sample period apart**...

Lessons from Decomposition Theorem, so far

δ^{DD} **depends on the weights** for three groups

- Treated vs. untreated
- Early vs. late
- Late vs. early (this is less obvious)

Greater weight will be given to pairs with

- big groups (i.e. many observations)
- groups that are treated closer to the middle of the sampling period
- and treated groups whose treatment periods are half the sample period apart

Often times a **few cases dominate** in the estimation of δ^{DD}

What Parameter are We Estimating?

Average Treatment Effect on the Treated for timing group k for year τ

$$ATT_k(\tau) = E\Big[Y_{it}^1 - Y_{it}^0 \mid k, t = \tau\Big]$$

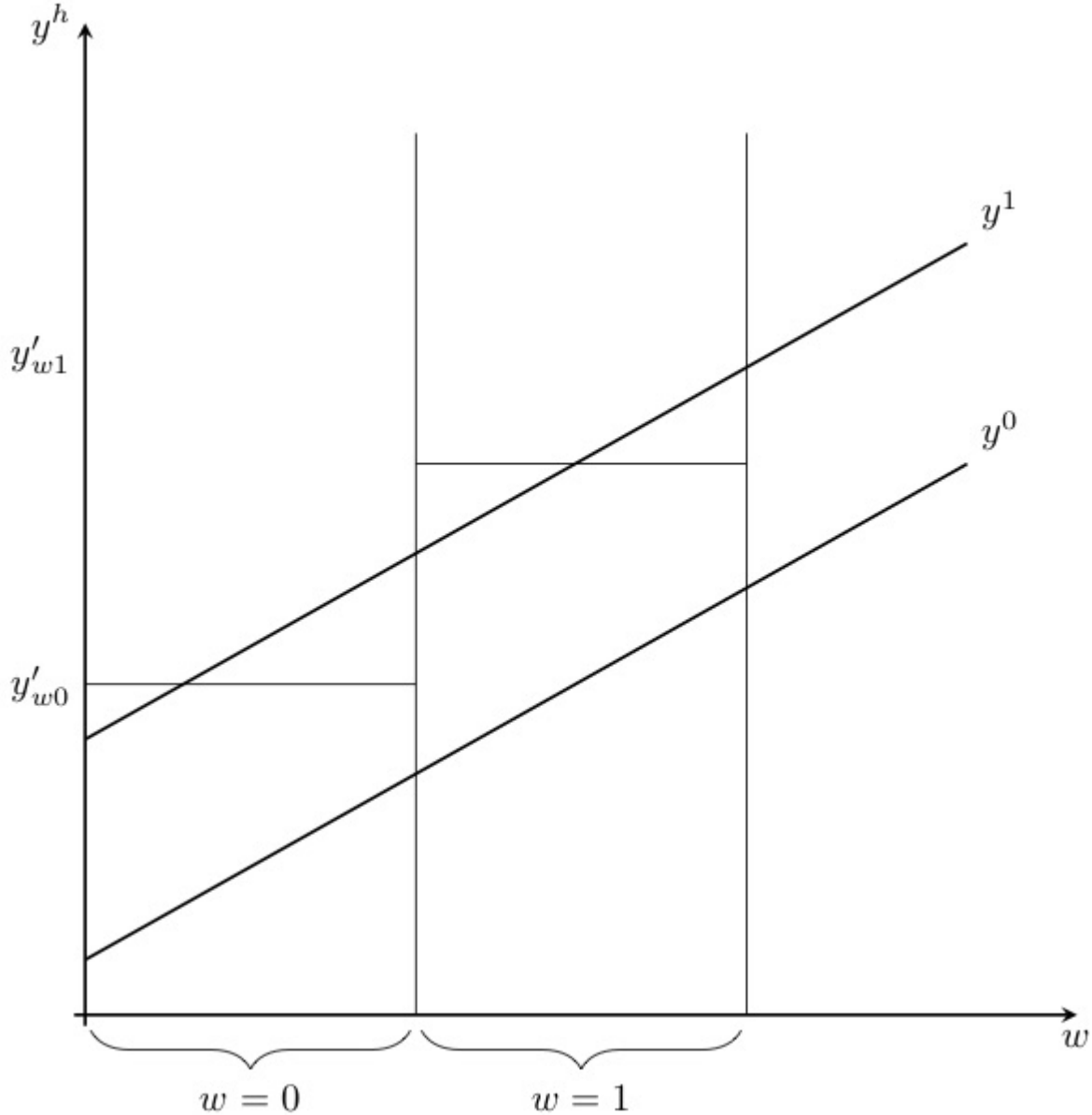
Consider the ATT for a time window W

$$ATT_k(\tau) = E\Big[Y_{it}^1 - Y_{it}^0 \mid k, \tau \in W\Big]$$

Difference over time in average potential outcomes

$$\Delta Y_k^h(W_1, W_0) = E\Big[Y_{it}^h \mid k, W_1\Big] - E\Big[Y_{it}^h \mid k, W_0\Big]$$

Average Potential Outcomes with a Trend



Source: mixtape ch. 9

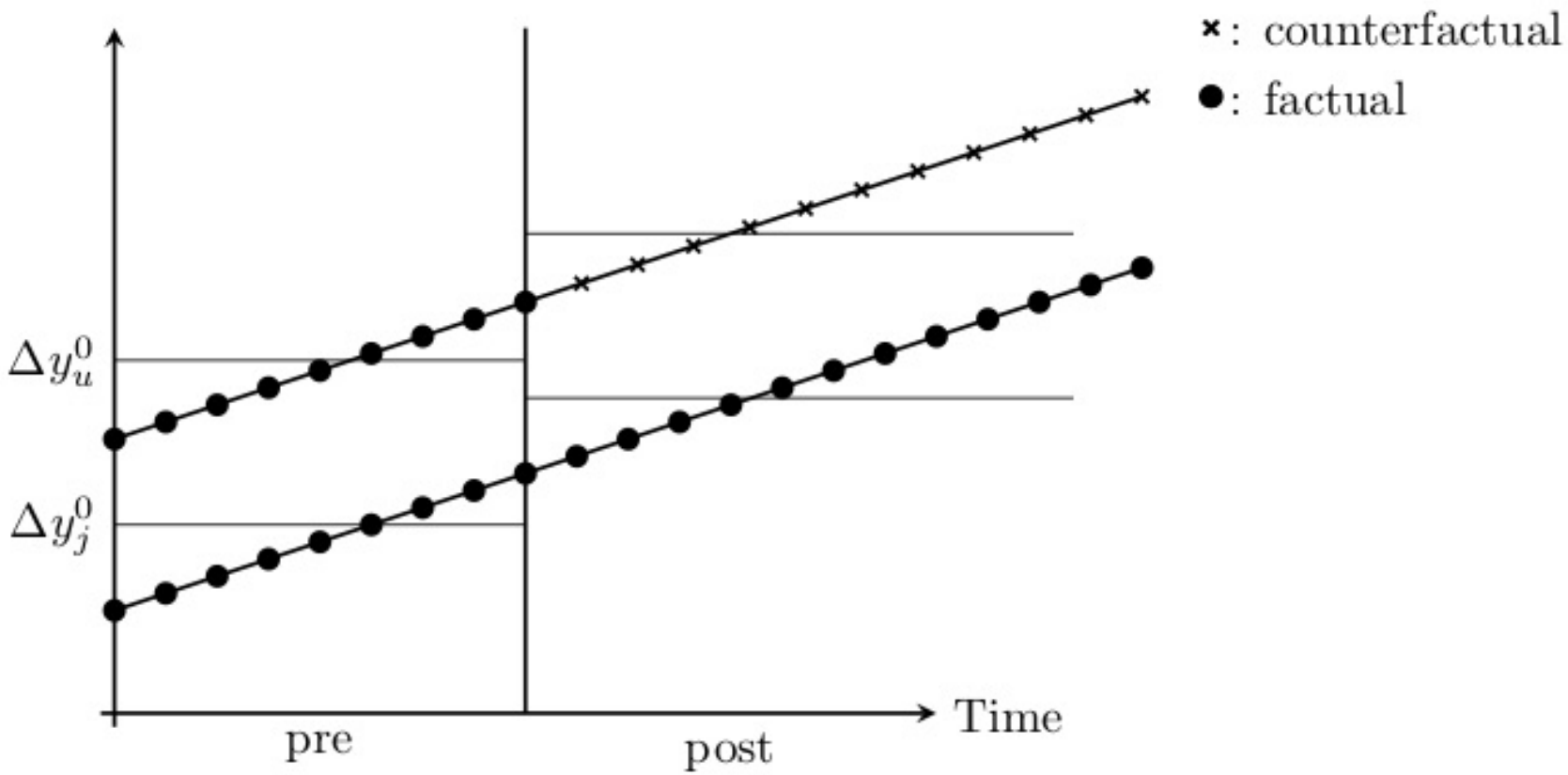
Remember the ATT in a 2×2 Pair

$$\widehat{\delta}^{2\times2}_{kU} = \underbrace{E\Big[Y^1_j \mid \text{Post}\Big] - E\Big[Y^0_j \mid \text{Post}\Big]}_{\text{ATT}}$$
$$+ \Bigg[\underbrace{E\Big[Y^0_j \mid \text{Post}\Big] - E\Big[Y^0_j \mid \text{Pre}\Big]\Big] - \Big[E\Big[Y^0_U \mid \text{Post}\Big] - E\Big[Y^0_U \mid \text{Pre}\Big]\Big]}_{\text{Non-parallel trends bias in } 2\times2 \text{ case}}\Bigg]$$

Or, more compact,

$$\widehat{\delta}^{2\times2}_{kU} = ATT_{\text{Post},j} + \underbrace{\Delta Y^0_{\text{Post,Pre},j} - \Delta Y^0_{\text{Post,Pre},U}}_{\text{Selection bias}}$$

Counterfactual in a 2×2 Pair



Source: mixtape ch. 9

ATT in Early-vs-late Pairs

Here the late adopters are a **counterfactual for early adopters**

$$\widehat{\delta}_{kl}^{2\times2} = ATT_k(MID) + \Delta Y_k^0(MID, \text{Pre}) - \Delta Y_l^0(MID, \text{Pre})$$

- Trends need to be parallel until both groups are treated
- Parallel trends bias: $\Delta Y_k^0(MID, \text{Pre}) - \Delta Y_l^0(MID, \text{Pre})$

ATT in Late-vs-Early Pairs

This is where things get tricky! Early adopters are a **counterfactual for late adopters**

$$\widehat{\delta}_{lk}^{2\times 2} = ATT_{l,\text{Post}(l)} + \underbrace{\Delta Y_l^0(\text{Post}(l), MID) - \Delta Y_k^0(\text{Post}(l), MID)}_{\text{Parallel-trends bias}} - \underbrace{(ATT_k(\text{Post}) - ATT_k(Mid))}_{\text{Heterogeneity in time bias}}$$

- Trends need to be parallel from the time the early adopter has been treated
- But that is not enough. The treatment effect for the early adopter needs to be constant over time

Heterogeneous Treatment Effects

There are two types of heterogeneous treatment effects:

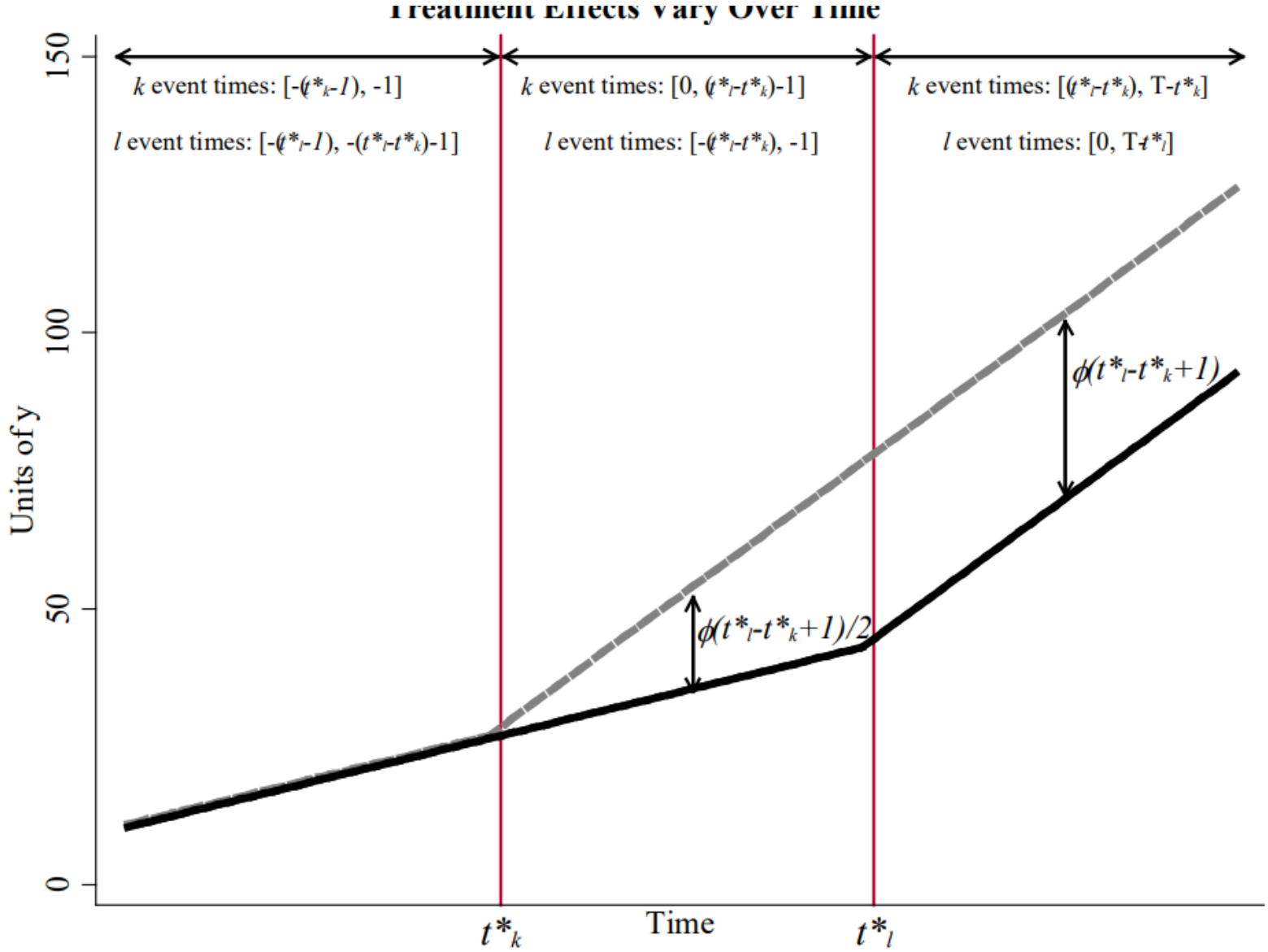
Heterogeneous effects across groups

- The difference in potential outcomes differs across groups
- In other words, the same treatment would lead to different responses in different groups/units

Heterogeneous effects within groups over time

- Need to see this relative to a counterfactual time path
- The difference between the actual path and the counterfactual changes over time
- Example: treatment pushes units onto a different time trend

(Within-group) Treatment Effects Vary over Time



Source: Goodman-Bacon (2018)

(Within-group) Treatment Effects Vary over Time

Previous graph:

- Treatment pushes the early adopters onto a **different time trend**
- Early adopters after treatment are a control group for late adopters
- Late adopters get the wrong counterfactual

This is **not a violation of the parallel trends assumption**

- After treatment, both are on the same time trend

But the **bias due to time-varying treatment effects** can be severe

- In this example, the estimated treatment effect is smaller than the true treatment effect
- ...it could even be negative (despite the true effect being positive)

Building Blocks of the Decomposition Theorem

$$\widehat{\delta}^{DD} = \sum_{k \neq U} s_{kU} \widehat{\delta}^{2\times2}_{kU} + \sum_{k \neq U} \sum_{l > k} s_{kl} \left[\mu_{kl} \widehat{\delta}^{2\times2,k}_{kl} + (1 - \mu_{kl}) \widehat{\delta}^{2\times2,l}_{kl} \right]$$

1) Weights: recall that s_{kU}, s_{kl}, μ_{kl} depend on the variation in treatment within a 2×2 pair and the pair's sample size

2) 2×2 **pairs** ($|$ $\Rightarrow$ intermediate periods)

$$\begin{aligned}
\widehat{\delta}^{2\times2}_{kU} &= ATT_k(\text{Post}) + \Delta Y^0_l(\text{Post}, \text{Pre}) - \Delta Y^0_U(\text{Post}, \text{Pre}) \\
\widehat{\delta}^{2\times2,k}_{kl} &= ATT_k(|) + \Delta Y^0_l(|, \text{Pre}) - \Delta Y^0_l(|, \text{Pre}) \\
\widehat{\delta}^{2\times2,l}_{lk} &= ATT_l\,\text{Post}(l) + \Delta Y^0_l(\text{Post}(l), |) - \Delta Y^0_k(\text{Post}(l), |) \\
&\quad - (ATT_k(\text{Post}) - ATT_k(|))
\end{aligned}$$

What Parameter are We Estimating?

For the canonical model, it can be shown that

$$\underset{N\to\infty}{\text{plim}}\hat{\delta}^{DD} = \beta^{DD} = VWATT + VWCT - \Delta ATT$$

VWATT	Variance-weighted ATT
VWCT	Variance-weighted common trend
ΔATT	Change in a treatment effects (within groups) over time

To identify $VWATT$, we need to assume (and justify) why $VWCT = \Delta ATT = 0$

The Variance-Weighted ATT

Ideally, we want to estimate the ATT. The VWATT is the next best alternative...

$$VWATT = \sum_{k \neq U} \sigma_{kU} ATT_k(\text{Post}(k)) + \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \Big[\mu_{kl} ATT_k(|) + (1 - \mu_{kl}) ATT_l(POST(l)) \Big]$$

The $VWATT = ATT$ if the **ATTs are the same for each pair**

Otherwise we **identify a weighted average**

- That's what regression does: it places more weight on groups with more variance in the treatment
- But the *VWATT* can be far away from the ATT if some groups carry a heavy weight

Variance-Weighted Common Trends

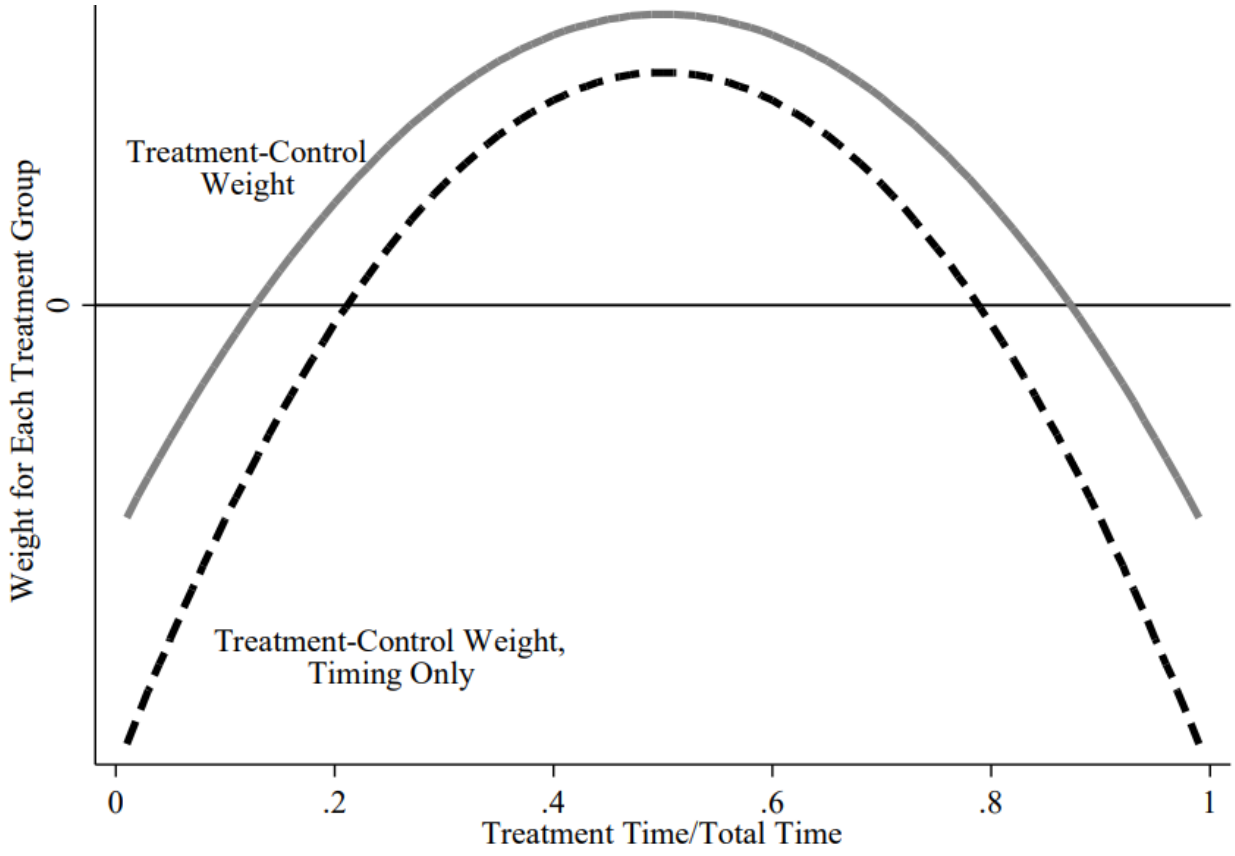
In the **staggered adoption design**, the **common trends assumption** $VWCT = 0$ is more complicated than in the simple 2×2 DiD

$$
\begin{aligned}
VWCT = \sum_{k \neq U} \sigma_{kU} \Big[& \Delta Y_k^0(\text{Post}(k), \text{Pre}) - \Delta Y_U^0(\text{Post}(k), \text{Pre}) \Big] \\
& + \sum_{k \neq U} \sum_{l > k} \sigma_{kl} \Big[\mu_{kl} \{ \Delta Y_k^0(Mid, \text{Pre}(k)) - \Delta Y_l^0(|, \text{Pre}(k)) \} \\
& + (1 - \mu_{kl}) \{ \Delta Y_l^0(\text{Post}(l), |) - \Delta Y_k^0(\text{Post}(l), |) \} \Big]
\end{aligned}
$$

Things to note here:

- For the identification of the *VWATT* we do not require parallel trends in each pair
- The weights are the same as for the *VWATT*

More on Weights



Earliest and latest adopters are mainly in the control group

Observations treated in the middle are over-represented in the treatment group

Heterogeneous Treatment Effects over Time

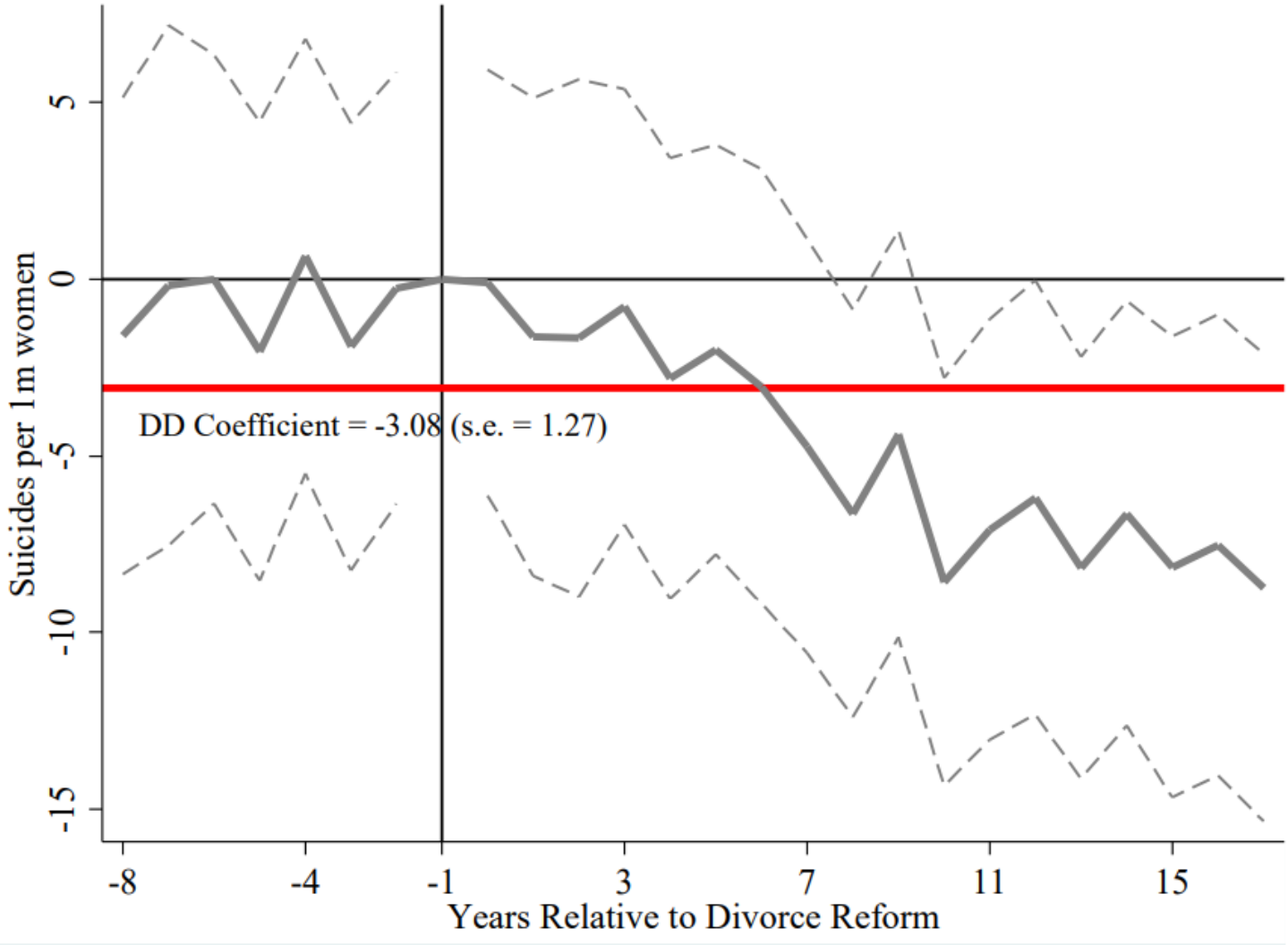
$$\Delta ATT = \sum_{k \neq U} \sum_{l > k} (1 - \mu_{kl}) \Big[ATT_k(\text{Post}(l) - ATT_k(\,|\,)) \Big]$$

ΔATT is a **source of bias** from a change in the ATT within a group over time

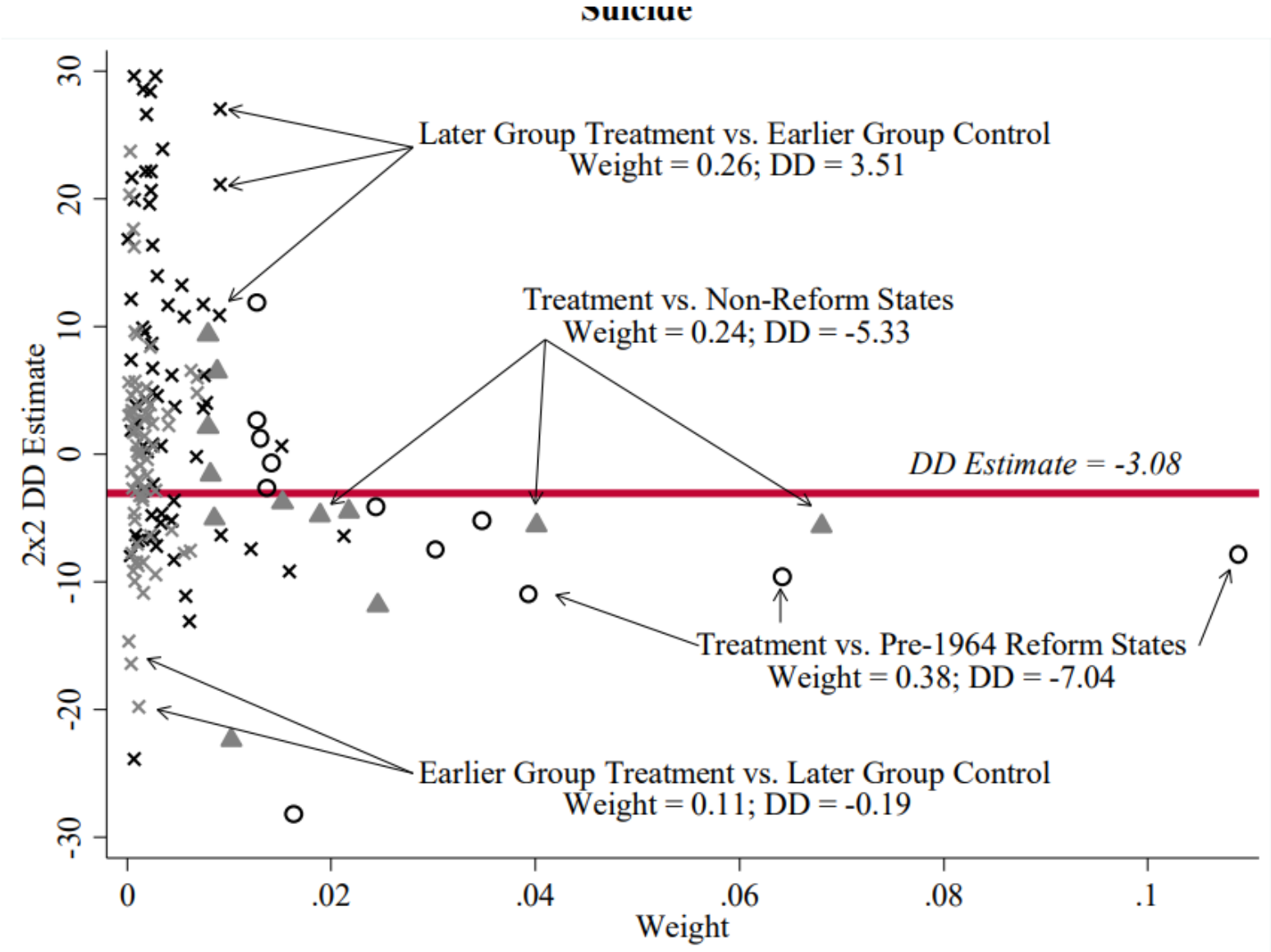
- This bias comes from the later vs. earlier comparison
- It appears whenever treatment leads to more than a level shift (!!)

Application: Effect of Unilateral Divorce

Stevenson & Wolfers (2006): divorce law reforms in 37 states during 70s/80s



Where Does the Effect Come from?



Lessons from Goodman-Bacon (2018)

DiD, while seemingly intuitive and transparent, is actually **not that easy**

As any FE regression, the **estimand is a variance-weighted average**

- It does not reflect the ATT
- Differential treatment timing adds a layer of complexity
- Thus, the DiD is not easy to interpret

What to do?

- Check the weights of the 2×2 pairs
- Corroborate *VWCT* through balancing tests
- Use a different estimator altogether

Group-Time ATT

As shown by Goodman-Bacon (2018), a **major source of bias** are comparisons of **late vs. early adopters**

- This bias cannot be eliminated through standard regression-based methods

? come up with an elegant non-parametric solution

- Idea: estimate the ATT separately for each group and time
- Use as control group only groups that have not yet been treated
- Aggregate the group-time ATTs into a (weighted) ATT

Group-Time ATT

$$ATT(g,t) = E[Y_t^1 - Y_t^0 | G_g = 1]$$

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

C	Indicator for never-treated group
G_g	Indicators for groups treated at different times

Propensity score $p_g(X) = P(G_g = 1 | X, G_g + C = 1)$

Building Blocks of the Group-Time ATT

$(Y_t - Y_{g-1})$: **Long differences between outcomes** in period t and the period before group g was treated

$$\left(\frac{G_g}{E[G_g]} - \frac{\frac{p_g(X)C}{1-p_g(X)}}{E\left[\frac{p_g(X)C}{1-p_g(X)}\right]} \right)$$

The expression in parentheses is a **weighting function** to balance the treated and control group on covariates

- Control units with similar characteristics to the treated groups are getting more weight

Further steps Callaway & Sant'Anna (2020)

Can **aggregate the** $ATT(g,t)$ across time and groups

- This will allow for the estimation of more interesting parameters

One can also **use this estimator to look at pre-trends**

- In TWFE models, these are inconsistently estimated (Borusyak & Jaravel, 2016; Abraham & Sun, forthcoming)

Inference is done through **bootstrapping**

Quo Vadis Diff-in-Diff?

DiD is often seen as a very **transparent** research design

- **It is not!** Especially not in the TWFE model
- There are many potential sources of bias
- The interpretation is often difficult
- And the identified parameters are not policy-relevant

Friends tell their friends not to use DiD? Not quite

- New methods help us to overcome many problems

Quo Vadis Diff-in-Diff?

Some recent **papers with methodological advances**

- Testing for parallel pre-trends (Freyaldenhoven *et al.*, 2019; Rambachan & Roth, 2020)
- Estimating dynamic treatment effects (Borusyak & Jaravel, 2016; Abraham & Sun, forthcoming)
- Re-weighting to recover relevant parameters (Callaway & Sant'Anna, 2020; de Chaisemartin & D'Haultfoeuille, 2019)
- Adjusting inference for (failed) pre-tests (Roth, 2019)
- Machine learning meets DiD (Athey *et al.*, 2018)

Conclusion: Important to stay up to date!

References I

Abraham, Sarah, & Sun, Liyang. forthcoming. Estimating Dynamic Treatment Effects in Event Studies With Heterogeneous Treatment Effects. *Journal of Econometrics*.

Angrist, Joshua, & Pischke, Jörn-Steffen. 2009. *Mostly Harmless Econometrics - An Empiricist's Companion*. Princeton University Press.

Athey, Susan, Bayati, Mohsen, Doudchenko, Nikolay, Imbens, Guido, & Khosravi, Khashayar. 2018. Matrix Completion Methods for Causal Panel Data Models. Oct.

Autor, David. 2003. Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing. *Journal of Labor Economics*, **21**(1), 1–42.

Borusyak, Kirill, & Jaravel, Xavier. 2016. Revisiting Event Study Designs. *Harvard University, mimeo*.

Callaway, Brantly, & Sant'Anna, Pedro H. C. 2020. Difference-in-Differences with Multiple Time Periods. *Journal of Econometrics*.

Card, David, & Krueger, Alan. 1994. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review*, **84**(4), 772–93.

Cortés, Patricia, & Pan, Jessica. 2013. Outsourcing Household Production: Foreign Domestic Workers and Native Labor Supply in Hong Kong. *Journal of Labor Economics*, **31**(2), 327–371.

Cunningham, Scott. 2020. *Causal Inference: The Mixtape*. Yale University Press.

de Chaisemartin, Clement, & D'Haultfoeuille, Xavier. 2019. Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects. *NBER Working Papers*, **25904**(May).

Duflo, Esther. 2001. Schooling And Labor Market Consequences Of School Construction In Indonesia: Evidence From An Unusual Policy Experiment. *American Economic Review*, **91**(4), 795–813.

Freyaldenhoven, Simon, Hansen, Christian, & Shapiro, Jesse M. 2019. Pre-event Trends in the Panel Event-Study Design. *American Economic Review*, **109**(9), 3307–38.

Goodman-Bacon, Andrew. 2018 (September). *Difference-in-Differences with Variation in Treatment Timing*. Working Paper 25018. National Bureau of Economic Research.

Kahn-Lang, Ariella, & Lang, Kevin. 2020. The Promise and Pitfalls of Differences-in-Differences: Reflections on 16 and Pregnant and Other Applications. *Journal of Business & Economic Statistics*, **38**(3), 613–620.

References II

Lafortune, Julien, Rothstein, Jesse, & Schanzenbach, Diane Whitmore. 2018. School Finance Reform and the Distribution of Student Achievement. *American Economic Journal: Applied Economics*, **10**(2), 1–26.

Rambachan, Ashesh, & Roth, Jonathan. 2020. An Honest Approach to Parallel Trends. *Harvard University, mimeo*.

Roth, Jonathan. 2019. Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends.

Schmidheiny, Kurt, & Siegloch, Sebastian. 2020. On Event Studies and Distributed-Lags in Two-way Fixed Effects Models: Identification, Equivalence, and Generalization.

Stevenson, Betsey, & Wolfers, Justin. 2006. Bargaining in the Shadow of the Law: Divorce Laws and Family Distress*. *The Quarterly Journal of Economics*, **121**(1), 267–288.