Difference-in-Differences III: Staggered DID

Lingguo Cheng

Nanjing University

December 22, 2025

Outline

1. Staggered Adoption

2. Event-study under the setting of staggered adoption

3. Goodman-Bacon Decomposition

Staggered adoption or differential timing

- Staggered assignment of treatment across geographic units over time
- That is, geographic units or groups receive treatments at different points in time
- The DD strategy is thus divided into two types:
  - DD with a universal timing/adoption
  - DD with a differential timing/staggered adoptions
- We have a good understanding of the first type of DD; but we did not have as good an understanding of type II DD until Goodman-Bacon (2021)

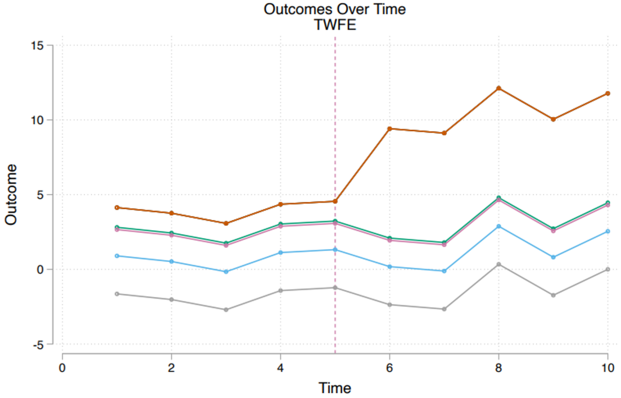
Twoway Fixed Effect DD estimation

- Suppose you have $s = 1, 2, \ldots$ and $t = 1, \ldots, T$ periods
- Treatment "turns on" at different times in different states.

$$Y_{st} = \beta^{FE} D_{st} + \alpha_s + \beta_t + \varepsilon_{st}$$

  - D_{st}, dummy variable set to 1 if the policy is enforced in state s during period t
  - α_s, state fixed effect (time invariant factor)
  - β_t, time fixed effect (time varying common factor)
- Twoway Fixed Effects: The current workhorse

How is the TWFE model like a DID?



- Group fixed effects: groups are different even before treatment. But group differences never change.
- Time fixed effects: trends are flexible, but exactly the same across groups.

Staggered adoption or differential timing makes things quite different

- Things become messy, actually very messy
- Multiple states adopt the treatment at different times, then
  - What is pre or post? No clear "pre" and "post"
  - What is control group or treatment group? No clear "control" or "treatment"
  - What if treatment effects are heterogeneous?
    - Across Units?
    - Over Time (phase-in effects, interactions with calendar time)
- How will you do the event study without the universal adoption time? E.g., how can you define pre or post, if state s never adopt the policy?

Outline

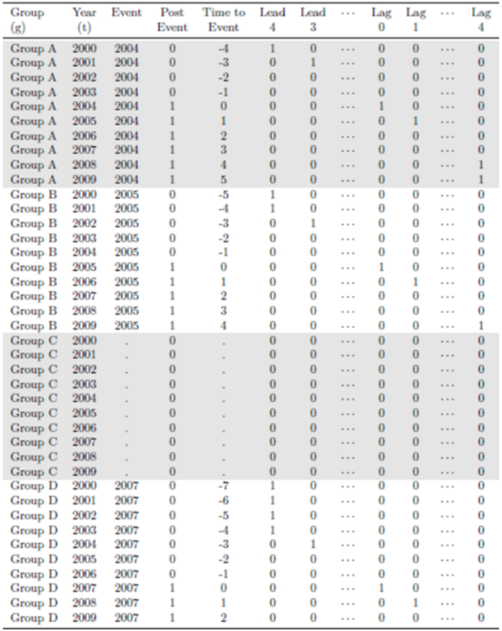
1. Staggered Adoption
2. Event-study under the setting of staggered adoption
3. Goodman-Bacon Decomposition

Event-study under the setting of staggered adoption

- Consider a panel covering a group, indexed as g and time periods t.
- We will denote as $Event_g$, a variable recording the period t in which the event is adopted in group g. The panel event study specification can be written as

$$y_{gt} = \alpha + \sum_{j=2}^{J} \beta_j (Lead\ j)_{gt} + \sum_{k=1}^{K} \gamma_k (Lag\ k)_{gt} + \mu_g + \lambda_t + X'_{gt}\Gamma + \varepsilon_{gt}$$

- Here μ_g and λ_t are group and time fixed effects, X_{gt} are (optionally) time-varying controls, and ε_{gt} is an unobserved error term. Leads and lags to the event of interest are defined as follows:
  - $(Lead\ J)_{gt} = \mathbb{1}[t \leq Event_g - J]$
  - $(Lead\ j)_{gt} = \mathbb{1}[t = Event_g - j]$ for $j \in \{1, \ldots, J-1\}$
  - $(Lag\ k)_{gt} = \mathbb{1}[t = Event_g + k]$ for $k \in \{1, \ldots, K-1\}$
  - $(Lag\ K)_{gt} = \mathbb{1}[t \geq Event_g + K]$



Group (g)	Year (t)	Event	Post Event	Time to Event	Lead 4	Lead 3	···	Lag 0	Lag 1	···	Lag 4
Group A	2000	2004	0	-4	1	0	···	0	0	···	0
Group A	2001	2004	0	-3	0	1	···	0	0	···	0
Group A	2002	2004	0	-2	0	0	···	0	0	···	0
Group A	2003	2004	0	-1	0	0	···	0	0	···	0
Group A	2004	2004	1	0	0	0	···	1	0	···	0
Group A	2005	2004	1	1	0	0	···	0	1	···	0
Group A	2006	2004	1	2	0	0	···	0	0	···	0
Group A	2007	2004	1	3	0	0	···	0	0	···	0
Group A	2008	2004	1	4	0	0	···	0	0	···	1
Group A	2009	2004	1	5	0	0	···	0	0	···	1
Group B	2000	2005	0	-5	1	0	···	0	0	···	0
Group B	2001	2005	0	-4	1	0	···	0	0	···	0
Group B	2002	2005	0	-3	0	1	···	0	0	···	0
Group B	2003	2005	0	-2	0	0	···	0	0	···	0
Group B	2004	2005	0	-1	0	0	···	0	0	···	0
Group B	2005	2005	1	0	0	0	···	1	0	···	0
Group B	2006	2005	1	1	0	0	···	0	1	···	0
Group B	2007	2005	1	2	0	0	···	0	0	···	0
Group B	2008	2005	1	3	0	0	···	0	0	···	0
Group B	2009	2005	1	4	0	0	···	0	0	···	1
Group C	2000	.	0	.	0	0	···	0	0	···	0
Group C	2001	.	0	.	0	0	···	0	0	···	0
Group C	2002	.	0	.	0	0	···	0	0	···	0
Group C	2003	.	0	.	0	0	···	0	0	···	0
Group C	2004	.	0	.	0	0	···	0	0	···	0
Group C	2005	.	0	.	0	0	···	0	0	···	0
Group C	2006	.	0	.	0	0	···	0	0	···	0
Group C	2007	.	0	.	0	0	···	0	0	···	0
Group C	2008	.	0	.	0	0	···	0	0	···	0
Group C	2009	.	0	.	0	0	···	0	0	···	0
Group D	2000	2007	0	-7	1	0	···	0	0	···	0
Group D	2001	2007	0	-6	1	0	···	0	0	···	0
Group D	2002	2007	0	-5	1	0	···	0	0	···	0
Group D	2003	2007	0	-4	1	0	···	0	0	···	0
Group D	2004	2007	0	-3	0	1	···	0	0	···	0
Group D	2005	2007	0	-2	0	0	···	0	0	···	0
Group D	2006	2007	0	-1	0	0	···	0	0	···	0
Group D	2007	2007	1	0	0	0	···	1	0	···	0
Group D	2008	2007	1	1	0	0	···	0	1	···	0
Group D	2009	2007	1	2	0	0	···	0	0	···	0

Table 1: A Stylized Example

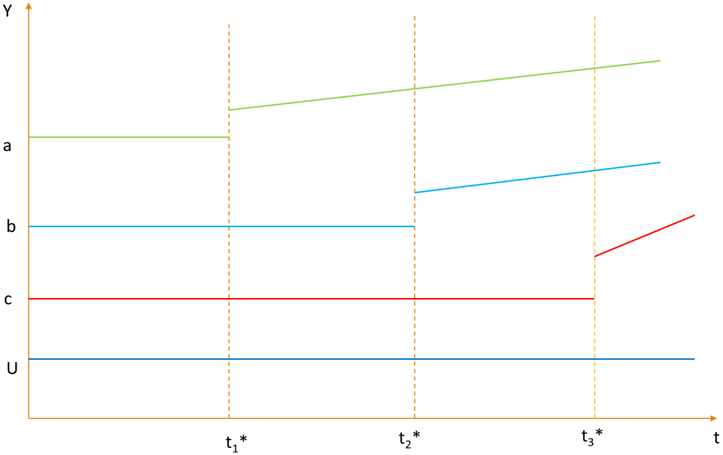
- Illustration through the Excel file
- What is the maximum K? J?

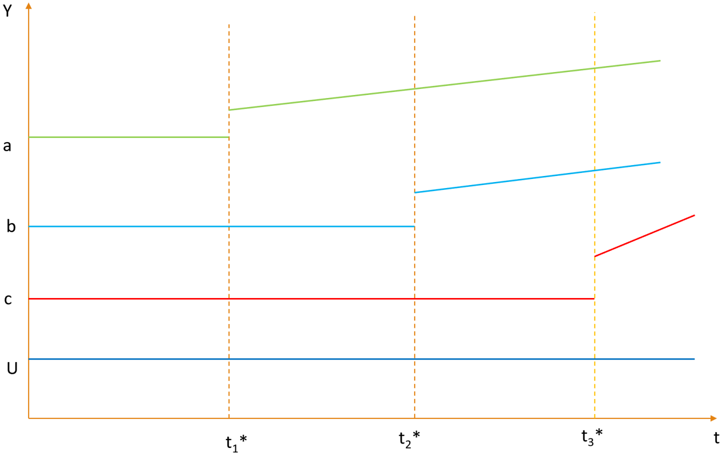
Outline

1. Staggered Adoption
2. Event-study under the setting of staggered adoption
3. Goodman-Bacon Decomposition

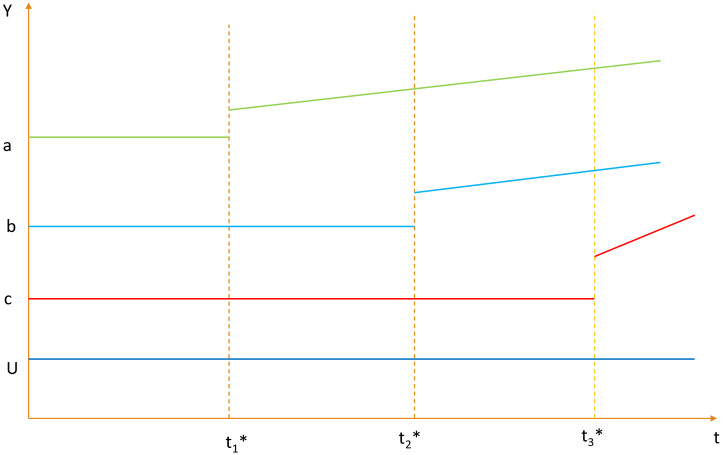
Goodman-Bacon decomposition: Goodman-Bacon (2019)

- What does the usually adopted TWFE DD estimation mean?
- The punchline of the Bacon decomposition theorem is that the two-way fixed effects estimator is a weighted average of all potential 2×2 DD estimates where weights are both based on group sizes and variance in treatment.


Y
a
b
c
U
t_1*
t_2*
t_3*
t



a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U



a to b	b to a	c to a
a to c	b to c	c to b
a to U	b to U	c to U

There are 9 "2×2" DD!

Goodman-Bacon decomposition: Goodman Bacon (2019)

- **A simple example.**
  - Assume that there are three groups: an early treatment group (k), a group treated later (l), and a group that is never treated (U). Groups k and l are similar in that they are both treated but they differ in that k is treated earlier than l.
  - Say there are 5 periods, and k is treated in period 2. Then it spends 80% of its time under treatment, or 0.8. But let's say l is treated in period 4. Then it spends 40% of its time treated, or 0.4. I represent this time spent in treatment for a group as D_k=0.8 and D_l=0.4.

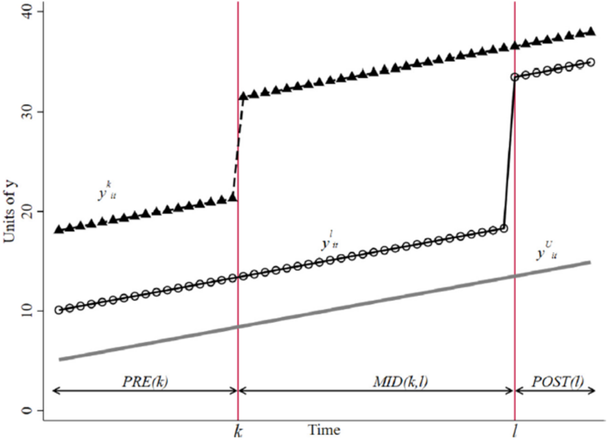
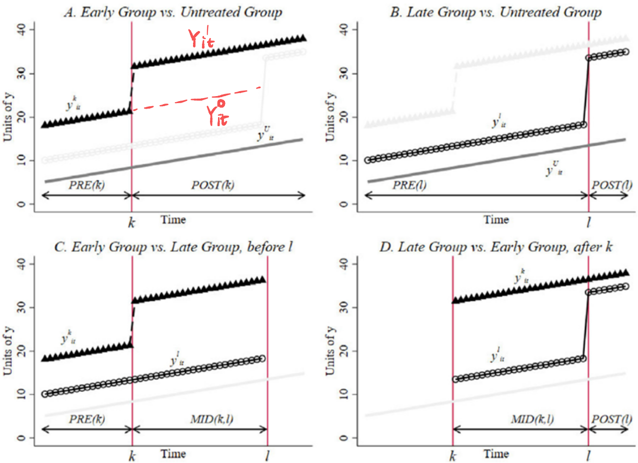


Fig. 1. Difference-in-Differences with variation in treatment Timing: Three groups. Notes: The figure plots outcomes in three timing groups: an untreated group, U; an early treatment group, k, which receives a binary treatment at $k = \frac{34}{100}T$; and a late treatment group, ℓ, which receives the binary treatment at $\ell = \frac{85}{100}T$. The x-axis notes the three sub-periods: the pre-period for timing group k, $[1, \mathrm{k} - 1]$, denoted by $PRE(k)$; the middle period when timing group k is treated and timing group ℓ is not, $[k, \ell - 1]$, denoted by $MID(k, \ell)$; and the post-period for timing group ℓ, $[\ell, T]$, denoted by $POST(\ell)$. The treatment effect is 10 in timing group k and 15 in timing group ℓ.



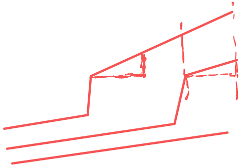


Fig. 2. The four simple (2x2) difference-in-differences estimates in the three group case. Notes: The figure plots outcomes for the subsamples that generate the four simple 2x2 difference-in-difference estimates in the three timing group case from Fig. 1. Each panel plots the data structure for one 2x2 DD. Panel A compares early treated units to untreated units ($\hat{\beta}^{DD}_{kU}$); panel B compares late treated units to untreated units ($\hat{\beta}^{DD}_{\ell U}$); panel C compares early treated units to late treated units during the late timing group's pre-period ($\hat{\beta}^{DD,k}_{k\ell}$); panel D compares late treated units to early treated units during the early timing group's post-period ($\hat{\beta}^{DD,\ell}_{k\ell}$). The treatment times mean that $\overline{D}_k = 0.67$ and $\overline{D}_\ell = 0.16$, so with equal group sizes, the decomposition weights on the 2x2 estimate from each panel are 0.365 for panel A, 0.222 for panel B, 0.278 for panel C, and 0.135 for panel D.

It turns out that

- The OLS estimate, $\hat{\beta}^{\text{DD}}$, in a two-way fixed-effects regression is a weighted average of all possible two-by-two DD estimators

$$\hat{\beta}^{DD} = \sum_{k \neq U} s_{kU} \hat{\beta}^{2\times 2}_{kU} + \sum_{k \neq U} \sum_{\ell > k} \left[s^{k}_{k\ell} \hat{\beta}^{2\times 2,k}_{k\ell} + s^{\ell}_{k\ell} \hat{\beta}^{2\times 2,\ell}_{k\ell} \right].$$

- where the 2×2 DD estimators are:

$$\begin{aligned}
\hat{\beta}^{2\times 2}_{kU} &\equiv \left(\bar{y}_k^{\text{POST}(k)} - \bar{y}_k^{\text{PRE}(k)} \right) - \left(\bar{y}_U^{\text{POST}(k)} - \bar{y}_U^{\text{PRE}(k)} \right), \\
\hat{\beta}^{2\times 2,k}_{k\ell} &\equiv \left(\bar{y}_k^{\text{MID}(k,\ell)} - \bar{y}_k^{\text{PRE}(k)} \right) - \left(\bar{y}_\ell^{\text{MID}(k,\ell)} - \bar{y}_\ell^{\text{PRE}(k)} \right), \\
\hat{\beta}^{2\times 2,\ell}_{k\ell} &\equiv \left(\bar{y}_\ell^{\text{POST}(\ell)} - \bar{y}_\ell^{\text{MID}(k,\ell)} \right) - \left(\bar{y}_k^{\text{POST}(\ell)} - \bar{y}_k^{\text{MID}(k,\ell)} \right).
\end{aligned}$$

The weights

$n_{kU} = \frac{\Sigma_i (t_i \geq k)}{N_k + N_U}$

DDI.

$$s_{kU} = \frac{(n_k + n_U)^2 \overbrace{n_{kU}(1-n_{kU})\overline{D}_k(1-\overline{D}_k)}^{\hat{V}^D_{kU}}}{\hat{V}^D}$$

$n_k, n_U \uparrow, s_{kU} \uparrow$

数据方差. (B分布)

时间方差.

$\overline{D}_k = 0.5$ [illegible]最大.

$$s^k_{k\ell} = \frac{((n_k + n_\ell)(1-\overline{D}_\ell))^2 \overbrace{n_{k\ell}(1-n_{k\ell}) \frac{\overline{D}_k - \overline{D}_\ell}{1-\overline{D}_\ell} \frac{1-\overline{D}_k}{1-\overline{D}_\ell}}^{\hat{V}^{D,k}_{k\ell}}}{\hat{V}^D}$$

$$s^\ell_{k\ell} = \frac{((n_k + n_\ell)\overline{D}_k)^2 n_{k\ell}(1-n_{k\ell}) \overbrace{\frac{\overline{D}_\ell}{\overline{D}_k} \frac{\overline{D}_k - \overline{D}_\ell}{\overline{D}_k}}^{\hat{V}^{D,\ell}}}{\hat{V}^{D,\ell}}$$

- Where $\sum_{k \neq U} s_{kU} + \sum_{k \neq U} \sum_{\ell > k} [s^k_{k\ell} + s^\ell_{k\ell}] = 1$, $\overline{D}_k \equiv \sum_t \mathbb{1}\{t \geq k\}/T$, $n_k \equiv \frac{\sum_i \mathbb{1}\{t_i \geq k\}}{N}$

t≥k的比例

Takeaways from the Bacon decomposition: weights

- First, the Bacon decomposition shows that it's group variation that two-way fixed effects is using to calculate that parameter you're seeking. The more states that adopted a law at the same time, the bigger they influence that final aggregate estimate itself.
- The other thing that matters in these weights is within-group treatment variance.
  - Treatment variance is maximized at D=0.5
  - Being treated in the middle of the panel actually directly influences the numerical value you get when employing TWFE estimation.
  - That means lengthening or shortening the panel can actually change the point estimate purely by changing group treatment variance and nothing more.
  - Isn't that kind of strange though?

New notatios for ATT

- Define any year-specific ATT as

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

- Define it over a time window W (e.g., A post-treatment window)

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

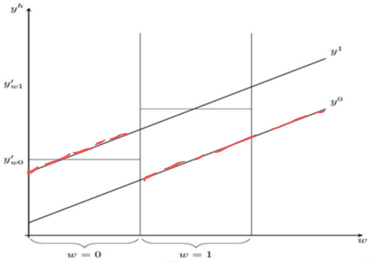
average on W.

- Define differences in average potential outcomes over time as

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

for $h = 0$ (i.e., Y^0) or $h = 1$ (i.e., Y^1)

- With trends, differences in mean potential outcomes is non-zero



$j \in \{\ell, k\}$

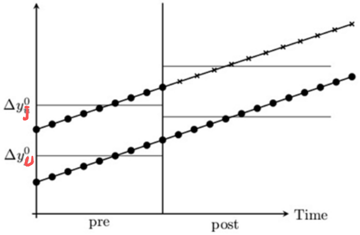
$$\begin{aligned}\widehat{\delta}^{2\times 2}_{kU} &= (E[Y_j \mid \text{Post}] - E[Y_j \mid \text{Pre}]) - (E[Y_U \mid \text{Post}] - E[Y_U \mid \text{Pre}]) \\ &= \underbrace{\Big(E[Y^1_j \mid \text{Post}] - E[Y^0_j \mid \text{Pre}]\Big) - \Big(E[Y^0_U \mid \text{Post}] - E[Y^0_U \mid \text{Pre}]\Big)}_{\text{Switching equation}} \\ &\quad + \underbrace{E[Y^0_j \mid \text{Post}] - E[Y^0_j \mid \text{Post}]}_{\text{Adding zero}} \\ &= \underbrace{E[Y^1_j \mid \text{Post}] - E[Y^0_j \mid \text{Post}]}_{\text{ATT}} \\ &\quad + \underbrace{\Big(E[Y^0_j \mid \text{Post}] - E[Y^0_j \mid \text{Pre}]\Big) - \Big(E[Y^0_U \mid \text{Post}] - E[Y^0_U \mid \text{Pre}]\Big)}_{\text{Non-parallel trends bias in } 2\times 2 \text{ case}}\end{aligned}$$

That is,

[illegible]

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

- The 2 × 2 DD can be expressed as the sum of the ATT itself plus a parallel trends assumption, and without parallel trends, the estimator is biased.



- ×: Counterfactual; ●: factual

$$\widehat{\delta}^{2\times 2}_{kU} = ATT_{k,\text{Post}} + \underline{\Delta Y^0_k(\text{Post}(k),\text{Pre}(k)) - \Delta Y^0_U(\text{Post}(k),\text{Pre})} \quad \checkmark$$

$$\widehat{\delta}^{2\times 2}_{k\ell} = ATT_k(\text{Mid}) + \underline{\Delta Y^0_k(\text{Mid},\text{Pre}) - \Delta Y^0_\ell(\text{Mid},\text{Pre})} \quad \checkmark$$

$$\widehat{\delta}^{2\times 2}_{\ell k} = ATT_{\ell,\text{Post}(\ell)}$$

$$+ \underbrace{\Delta Y^0_\ell(\text{Post}(\ell),\text{Mid}) - \Delta Y^0_k(\text{Post}(\ell),\text{Mid})}_{\text{Parallel-trends bias}}$$

mid

Post

$$- \underbrace{(ATT_k(\text{Post}) - ATT_k(\text{Mid}))}_{\text{Heterogeneity in time bias!}}$$

- The first line is the ATT that we desperately hope to identify.
- The selection bias zeroes out insofar as Y^0 for k and ℓ has the same parallel trends from *mid* to *post* period.
- And the treatment effects bias in the third line zeroes out so long as there are constant treatment effects for a group over time. But if there is heterogeneity in time for a group, then the two ATT terms will not be the same, and therefore will not zero out.

- In sum, the decomposition formula for DD is:

$$
\begin{aligned}
\widehat{\delta}^{DD} &= \sum_{k \neq U} s_{kU}\widehat{\delta}^{2\times 2}_{kU} + \sum_{k \neq U}\sum_{\ell > k} s_{k\ell}\left[\mu_{k\ell}\widehat{\delta}^{2\times 2,k}_{k\ell} + (1-\mu_{k\ell})\widehat{\delta}^{2\times 2,\ell}_{k\ell}\right] \\
\widehat{\delta}^{2\times 2}_{kU} &= ATT_k(\text{Post}) + \Delta Y^0_k(\text{Post,Pre}) - \Delta Y^0_U(\text{Post,Pre}) \\
\widehat{\delta}^{2\times 2,k}_{k\ell} &= ATT_k(\text{Mid}) + \Delta Y^0_k(\text{Mid,Pre}) - \Delta Y^0_\ell(\text{Mid,Pre}) \\
\widehat{\delta}^{2\times 2,\ell}_{\ell k} &= ATT_\ell(\text{Post}(\ell)) + \Delta Y^0_\ell(\text{Post}(\ell),\text{Mid}) - \Delta Y^0_k(\text{Post}(\ell),\text{Mid}) \\
&\quad \boxed{-\left(ATT_k(\text{Post}) - ATT_k(\text{Mid})\right)}
\end{aligned}
$$

bias

- That is

$$
\overset{\underset{n\to\infty}{\text{plim}}}{\underset{n\to\infty}{\cancel{p\lim}}}\ \widehat{\delta}^{DD}_n = \boxed{VWATT} + \boxed{VWCT} - \boxed{\Delta ATT}
$$

Varience weighted

Common Trend

加权为0即可

放松了假设

Variance weighted ATT

- $$\text{VWATT} = \sum_{k \neq U} s_{kU} ATT_k(\text{Post}(k)) + \sum_{k \neq U} \sum_{\ell > k} s_{k\ell} \left[\mu_{k\ell} ATT_k(\text{MID}) + (1 - \mu_{k\ell}) ATT_\ell(\text{POST}(\ell)) \right]$$
- Notice that the VWATT simply contains the three ATTs identified above, each of which was weighted by the weights contained in the decomposition formula. While these weights sum to one, that weighting is irrelevant if the ATT are identical.

Variance weighted common trends

- VWCT stands for variance weighted common trends. This is just the collection of non-parallel-trends biases

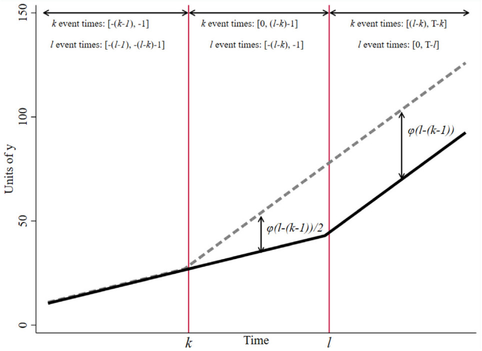
$$\begin{aligned} VWCT = & \sum_{k \neq U} \sigma_{kU} \left[\Delta Y_k^0(\text{Post}(k), \text{Pre}) - \Delta Y_U^0(\text{Post}(k), \text{Pre}) \right] \\ & + \sum_{k \neq U} \sum_{\ell > k} \sigma_{k\ell} \Big[\mu_{k\ell} \left\{ \Delta Y_k^0(\text{Mid}, \text{Pre}(k)) - \Delta Y_\ell^0(\text{Mid},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

- Goodman-Bacon (2019) shows us is that technically, you don't need identical trends because the weights can make it hold even if we don't have exact parallel trends.

ATT heterogeneity within time bias

$$\Delta ATT = \sum_{k \neq U} \sum_{\ell > k} (1 - \mu_{k\ell}) \left[ATT_k(\text{Post}(\ell)) - ATT_k(\text{Mid}) \right]$$

- Heterogeneity in the ATT has two interpretations: you can have heterogeneous treatment effects across groups, and you can have heterogeneous treatment effects within groups over time. The ATT is concerned with the latter only.
- Time-varying treatment effects, even if they are identical across units, generate cross-group heterogeneity because of the differing post-treatment windows and the fact that earlier-treated groups are serving as controls for later-treated groups.



- **Notes:** The figure plots a stylized example of a timing-only DD set up with a treatment effect that is a trend-break rather than a level shift (see Meer and West, 2016). The trend-break effect equals $\phi \cdot (t - t_i + 1)$. The top of the figure notes which event-times lie in the PRE(k), MID(k, ℓ), and POST(ℓ) periods for each unit. The figure also notes the average difference between timing groups in each of these periods.
- In the MID(k, ℓ) period, outcomes differ by $\frac{\phi}{2}(\ell - (k-1))$ on average.
- In the POST(ℓ) period, however, outcomes had already been growing in the early group for $\ell - k$ periods, and so they differ by $\phi(\ell - (k-1))$ on average. The 2x2 DD that compares the later-treated group to the earlier-treated group is biased and, in the linear trend-break case, weakly negative despite a positive and growing treatment effect.

Time varying treatment effect

- **When is this a problem?**
- It's a problem if there are a lot of those 2×2s or if their weights are large. If they are negligible portions of the estimate, then even if it exists, then given their weights are small (as group shares are also an important piece of the weighting not just the variance in treatment) the bias may be small.
- Time-varying treatment effects cause bias when you use a TWFE regression to analyze data from a staggered adoption design.
- Stata package
  ddtiming

Stevenson and Wolfers'(2006)

- Stevenson and Wolfers' (2006) examine the no-fault divorce reforms and female suicide.
- Unilateral (or no-fault) divorce allowed either spouse to end a marriage, redistributing property rights and bargaining power relative to fault-based divorce regimes.
- Stevenson and Wolfers exploit "the natural variation resulting from the different timing of the adoption of unilateral divorce laws" in 37 states from 1969–1985 (see Table 1) using the "remaining fourteen states as controls" to evaluate the effect of these reforms on female suicide rates.

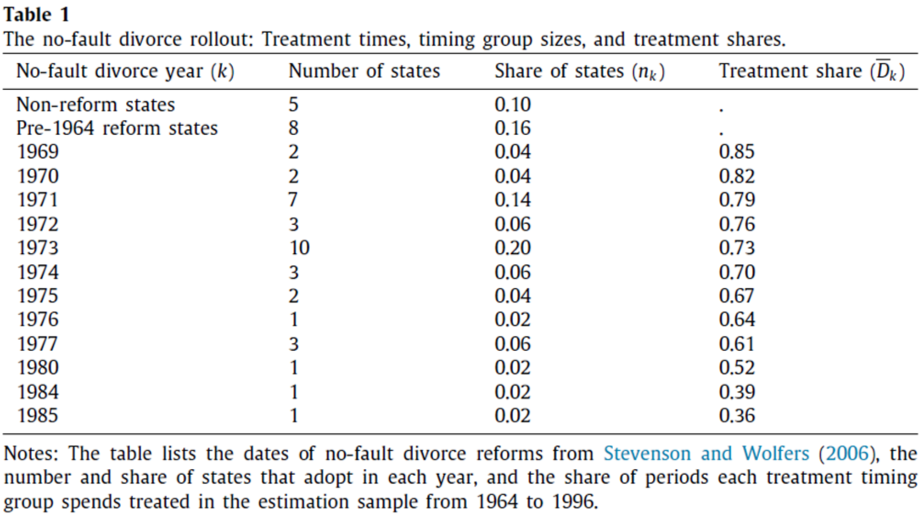
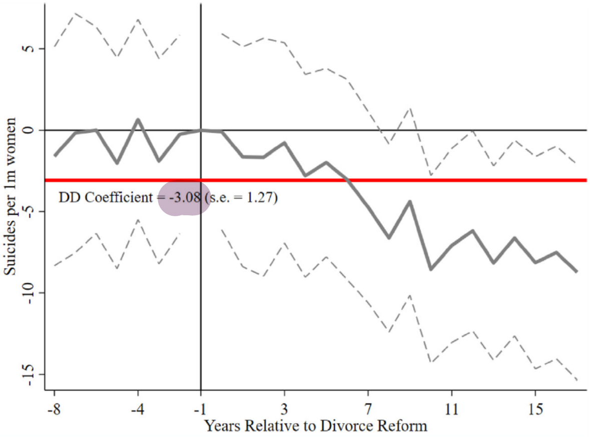


Table 1
The no-fault divorce rollout: Treatment times, timing group sizes, and treatment shares.

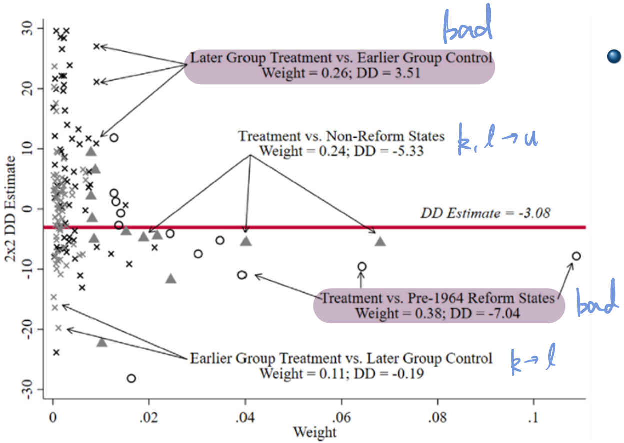
No-fault divorce year (k)	Number of states	Share of states (n_k)	Treatment share ($\overline{D}_k$)
Non-reform states	5	0.10	.
Pre-1964 reform states	8	0.16	.
1969	2	0.04	0.85
1970	2	0.04	0.82
1971	7	0.14	0.79
1972	3	0.06	0.76
1973	10	0.20	0.73
1974	3	0.06	0.70
1975	2	0.04	0.67
1976	1	0.02	0.64
1977	3	0.06	0.61
1980	1	0.02	0.52
1984	1	0.02	0.39
1985	1	0.02	0.36

Notes: The table lists the dates of no-fault divorce reforms from Stevenson and Wolfers (2006), the number and share of states that adopt in each year, and the share of periods each treatment timing group spends treated in the estimation sample from 1964 to 1996.



- Fig. 5. Event-study and difference-in-differences estimates of the effect of no-fault divorce on female suicide: Replication of Stevenson and Wolfers (2006).

- Notes: The figure plots event-study estimates from the two-way fixed effects regression equation, along with the DD coefficient. The specification does not include other controls and does not weigh by population.



- Difference-in-differences decomposition for unilateral divorce and female suicide.
- The figure plots each 2x2 DD component from the decomposition theorem against their weight for the unilateral divorce analysis.
    - The open circles, one timing group acts as the treatment group and the pre-1964 reform states act as the control group.
    - The closed triangles, treatment group vs. the non-reform states
    - The x's are the timing-only terms.
    - The two-way fixed effects estimate, -3.08, equals the average of the y-axis values weighted by their x-axis value.

Thanks for your Attention!

Any questions or comments please write to:

chenglingguo@nju.edu.cn