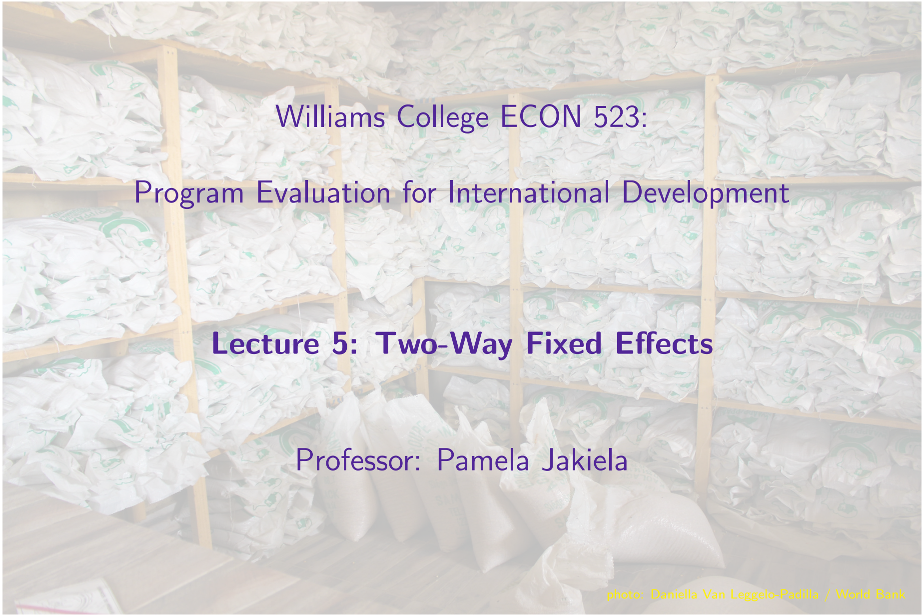


Williams College ECON 523:

Program Evaluation for International Development

Lecture 5: Two-Way Fixed Effects

Professor: Pamela Jakiela

photo: Daniella Van Leggelo-Padilla / World Bank

Variation in Treatment Timing

Example: States Adopted Medicaid at Different Times

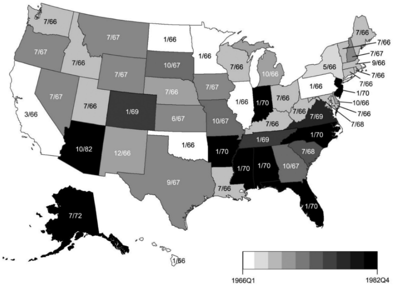


Figure 2.
Medicaid Adoption by Quarter
Notes: Adoption dates come from the Department of Health Education and Welfare (1970) & Social Security Administration (2013). The map is shaded relative to the quarter of adoption and states are labeled with the month and year of adoption.

source: Boudreaux, Golberstein, and McAlpine (Journal of Health Economics, 2016)

Example: Counties Opening Community Health Centers

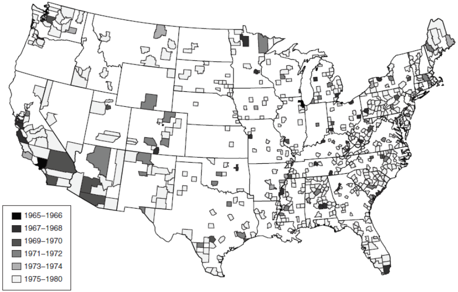


Figure 3. Establishment of Community Health Centers by County of Service Delivery, 1965–1980

Note: Dates are the first year that a CHC was established in the county.
Source: Information on CHCs drawn from NACAP and PHS reports.

source: Bailey and Goodman-Bacon (AER, 2015)

Example: African Countries Introducing Multi-Party Elections

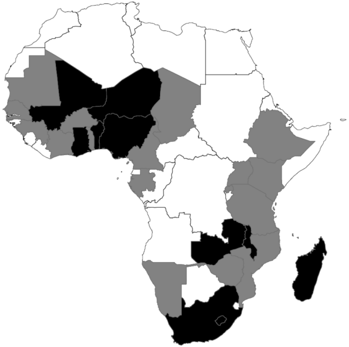


FIGURE A.2. Geographical distribution of democratized countries since 1990. Black-colored countries are democratized since 1990; grey-colored countries the other countries in the sample for infant mortality analysis. Democratized countries include the Comoros, tiny islands to the northwest of Madagascar, which may not be visible as black-colored.

source: Kudamatsu (JEEA, 2012)

Two-Way Fixed Effects Estimates of β_{twfe}

What exactly is β_{twfe}?

$$Y_{it} = \alpha_i + \gamma_t + \beta_{twfe} D_{it} + \varepsilon_{it}$$

unit fixed effects

time fixed effects

treatment dummy

that turns "on" at different times

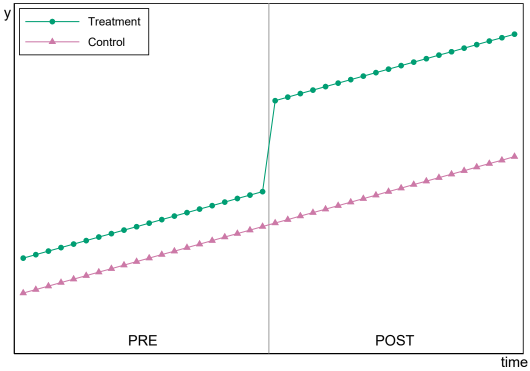
What exactly is β_{twfe}?

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	
ID 1	0	0	1	1	1	treatment
ID 2	0	0	1	1	1	
ID 3	0	0	0	0	0	comparison
ID 4	0	0	0	0	0	

Multiple Treatment and Comparison Groups

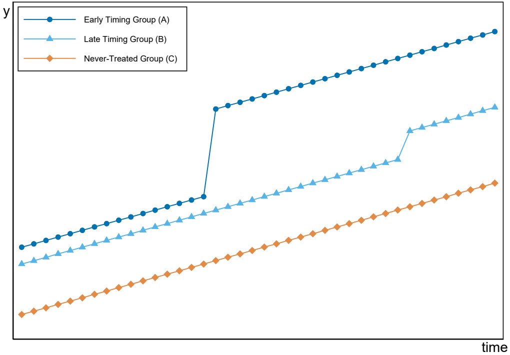
	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	
ID 1	0	0	1	1	1	1	1	1	timing group 1 (early)
ID 2	0	0	1	1	1	1	1	1	
ID 3	0	0	0	0	0	1	1	1	timing group 2 (late)
ID 4	0	0	0	0	0	1	1	1	
ID 5	0	0	0	0	0	0	0	0	never treated
ID 6	0	0	0	0	0	0	0	0	

2 × 2 Difference-in-Differences



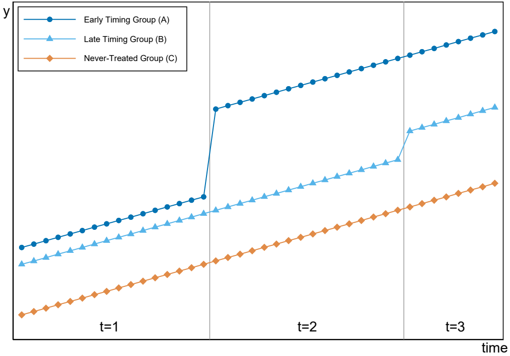
The 2 × 2 DiD estimate of the treatment effect: $\hat{\beta}_{DiD} = \left(\bar{Y}_T^{POST} - \bar{Y}_C^{POST}\right) - \left(\bar{Y}_T^{PRE} - \bar{Y}_C^{PRE}\right)$

Decomposition into Timing Groups



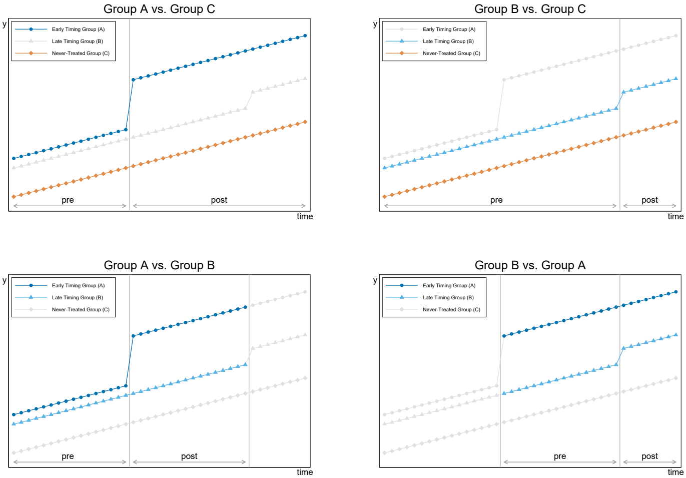
Panel with variation in treatment timing can be decomposed into distinct **timing groups** reflecting observed start of treatment

Decomposition into Timing Groups

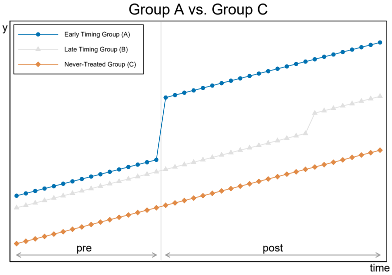


Example: with three timing groups (one of which is never treated), can construct three timing windows (pre, middle, post or $t = 1, 2, 3$)

Decomposition into Standard 2×2 DDs



Decomposition into Standard 2 × 2 DDs



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

$$\hat{\beta}^{DiD}_{AC} = \left(\bar{Y}^{POST}_{A} - \bar{Y}^{POST}_{C}\right) - \left(\bar{Y}^{PRE}_{A} - \bar{Y}^{PRE}_{C}\right)$$
$$= \left(\bar{Y}^{t=2,3}_{A} - \bar{Y}^{t=2,3}_{C}\right) - \left(\bar{Y}^{t=1}_{A} - \bar{Y}y^{t=1}_{C}\right)$$

Bacon Decomposition

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:

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

The two-way fixed effects estimator β_{twfe} is a weighted average of 2×2 diff-in-diff estimators across all possible pairwise combinations of timing groups (Goodman-Bacon 2021)

Bacon Decomposition: Calculating the Weights

Weights depend on sample size, variance of treatment w/in each 2×2 DiD:

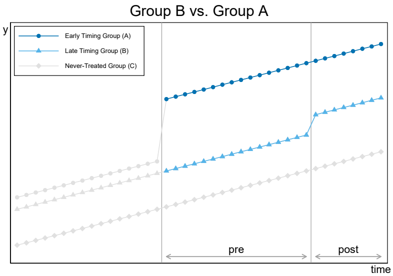
$$s_{kU} = \left[\frac{(n_k + n_U)^2}{\hat{V}^{\tilde{D}}}\right] \underbrace{n_{kU}\,(1 - n_{kU})\,\bar{D}_k(1 - \bar{D}_k)}_{\hat{Var}^{\tilde{D}}_{kU}}$$

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

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

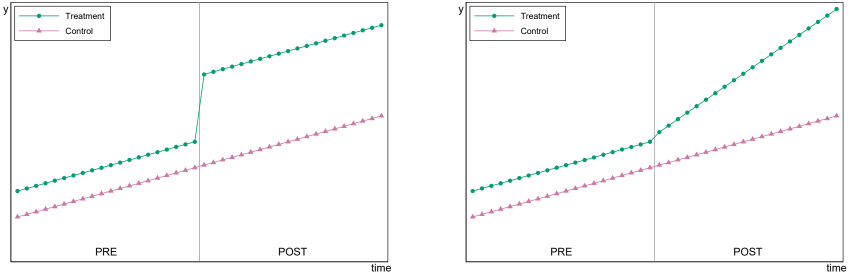
where n_k is the proportion of the sample in group timing group k for all k timing groups, $n_{kj} = n_k/(n_k + n_j)$, and $\bar{D}_k$ is the fraction of sample periods in which k is treated

Forbidden Comparisons



Some 2×2 DiDs (implicitly) use already-treated groups as the comparison

Why Are Forbidden Comparisons Forbidden?



If treatment changes the level of Y **and the rate of change in** Y, already-treated units cannot be used as a comparison group (common trends does not hold)

$\longrightarrow$ This problem does not arise (in the same way) in 2×2 DiD

Two-Way Fixed Effects β_{twfe} as a Weighted Sum

The two-way fixed effects estimator β_{twfe} is a weighted average of 2×2 DiD estimators across all possible pairwise combinations of timing groups + the never-treated (Goodman-Bacon 2021)

- Some use an already-treated group as comparison
  - Creates problems if treatment effects grow (or change in other ways) over time
  - TWFE imposes a model of homogeneous treatment effects
  - When treatment effects evolve over time or vary across units, the model is mis-specified

We can use Frisch-Waugh-Lovell to construct the TWFE/OLS weights used to generate β_{twfe}

- Weights on treated units are not always positive (they are also used as comparison)

Two-Way Fixed Effects as Univariate Regression

Two-way fixed effects is equivalent to univariate regression:

$$\tilde{Y}_{it} = \alpha + \tilde{D}_{it} + \epsilon_{it}$$

where $\tilde{Y}_{it} = Y_{it} - \bar{Y}_t - (\bar{Y}_i - \bar{\bar{Y}})$ and $\tilde{D}_{it}$ defined analogously

↑

"grand mean"

(just the mean across i and t)

Two-Way Fixed Effects as Univariate Regression

Two-way fixed effects is equivalent to univariate regression:

$$\tilde{Y}_{it} = \alpha + \tilde{D}_{it} + \epsilon_{it}$$

where $\tilde{Y}_{it} = Y_{it} - \bar{Y}_t - \left(\bar{Y}_i - \bar{\bar{Y}} \right)$ and $\tilde{D}_{it}$ defined analogously

$\Rightarrow$ Treatment dummy now effectively continuous measure $\tilde{D}_{it}$

$$\hat{\beta}_{twfe} = \sum_{it} \tilde{Y}_{it} \underbrace{\left(\frac{\tilde{D}_{it} - \bar{\tilde{D}}_{it}}{\sum_i \left(\tilde{D}_{it} - \bar{\tilde{D}}_{it} \right)^2} \right)}_{\text{OLS weight}}$$

Two-Way Fixed Effects as Univariate Regression

Two-way fixed effects is equivalent to univariate regression:

$$\tilde{Y}_{it} = \alpha + \tilde{D}_{it} + \epsilon_{it}$$

where $\tilde{Y}_{it} = Y_{it} - \bar{Y}_t - (\bar{Y}_i - \bar{\bar{Y}})$ and $\tilde{D}_{it}$ defined analogously

$\Rightarrow$ Treatment dummy now effectively continuous measure $\tilde{D}_{it}$

$$\hat{\beta}_{twfe} = \sum_{it} \tilde{Y}_{it} \underbrace{\left(\frac{\color{red}{\tilde{D}_{it}} - \bar{\tilde{D}}_{it}}{\sum_i \left(\tilde{D}_{it} - \bar{\tilde{D}}_{it} \right)^2} \right)}_{\text{OLS weight}} \quad \text{where } \bar{\tilde{D}}_{it} = 0$$

Negative Weights $\Leftrightarrow \hat{D}_{it} > 1$

- The weight on each (treated) observation depends on $\tilde{D}_{it}$:

$$\hat{\beta}_{twfe} = \sum_{it} \tilde{Y}_{it} \left(\frac{\tilde{D}_{it}}{\sum_i \tilde{D}_{it}^2} \right)$$

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

$$\tilde{D}_{it} = D_{it} - \hat{D}_{it}$$

where $\hat{D}_{it}$ is the predicted value of D_{it} from the regression of D_{it} on (all) the fixed effects

$\Rightarrow$ For treated units, $\tilde{D}_{it} < 0 \Leftrightarrow \bar{D}_t + \left(\bar{D}_i - \bar{\bar{D}} \right) > 1 \Leftrightarrow \hat{D}_{it} > 1$

DiD without Staggered Treatment Timing: Review

	$t = 1$	$t = 2$	$t = 3$	$t = 4$
ID 1	0	1	1	1
ID 2	0	0	0	0

DiD without Staggered Treatment Timing: Review

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

	$t = 1$	$t = 2$	$t = 3$	$t = 4$
ID 1	-0.375	0.125	0.125	0.125
ID 2	0.375	-0.125	-0.125	-0.125

← Equal weight (ID 1, $t = 2$ to $t = 4$)

DiD without Staggered Treatment Timing: Review

$$w_{it} = \tilde{D}_{it} / \left(\sum_{it} \tilde{D}_{it}^2 \right)$$

	$t=1$	$t=2$	$t=3$	$t=4$
ID 1	-1	$0.\bar{3}$	$0.\bar{3}$	$0.\bar{3}$
ID 2	1	$-0.\bar{3}$	$-0.\bar{3}$	$-0.\bar{3}$

← Equal weight

$$\hat{\beta}_{ols} = \sum_i Y_i w_i = \sum_i Y_i \frac{\tilde{D}_{it}}{\sum_{it} \tilde{D}_{it}^2}$$

DiD without Staggered Treatment Timing: Review

$w_{it} = \tilde{D}_{it} / \left(\sum_{it} \tilde{D}_{it}^2 \right)$

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	
ID 1	-1	$0.\bar{3}$	$0.\bar{3}$	$0.\bar{3}$	⟵ Equal weight
ID 2	1	$-0.\bar{3}$	$-0.\bar{3}$	$-0.\bar{3}$	⟵ Equal weight

$$\hat{\beta}_{ols} = \sum_i Y_i w_i$$

$$= \sum_{ET,pre} Y_i w_i + \sum_{ET,post} Y_i w_i + \sum_{NT,pre} Y_i w_i + \sum_{NT,post} Y_i w_i$$

DiD without Staggered Treatment Timing: Review

$$w_{it} = \tilde{D}_{it} / \left(\sum_{it} \tilde{D}_{it}^2 \right)$$

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	
ID 1	-1	$0.\bar{3}$	$0.\bar{3}$	$0.\bar{3}$	⟵ Equal weight
ID 2	1	$-0.\bar{3}$	$-0.\bar{3}$	$-0.\bar{3}$	⟵ Equal weight

$$\begin{aligned}
\hat{\beta}_{ols} &= \sum_i Y_i w_i \\
&= \sum_{ET,pre} Y_i w_i + \sum_{ET,post} Y_i w_i + \sum_{NT,pre} Y_i w_i + \sum_{NT,post} Y_i w_i \\
&= \ldots \\
&= \bar{Y}_{ET,pre}(-1) + 3\bar{Y}_{ET,post}(0.\bar{3}) + \bar{Y}_{NT,pre}(1) + 3\bar{Y}_{NT,post}(-0.\bar{3})
\end{aligned}$$

DiD with Staggered Treatment Timing: Example

	$t=1$	$t=2$	$t=3$	$t=4$
ID 1	0	1	1	1
ID 2	0	0	0	1
$\bar{D}_t$	0	0.5	0.5	1

mean treatment in period t

DiD with Staggered Treatment Timing: Example

	Y_{it}			
	$t = 1$	$t = 2$	$t = 3$	$t = 4$
ID 1	0	10	10	10
ID 2	0	0	0	10

Let $Y_{it} = \gamma_i + \lambda_t + \delta_{it}$

Treated cells

Positive weights (in treatment group)

DiD with Staggered Treatment Timing: Example

$Y_{it} \longrightarrow \hat{\beta}_{OLS} = 10$

	$t = 1$	$t = 2$	$t = 3$	$t = 4$
ID 1	0	10	10	10
ID 2	0	0	0	10

homogeneous impacts:
$E[\hat{\beta}_{OLS}] = \text{ATE}$

Treated cells

Positive weights (in treatment group)

DiD with Staggered Treatment Timing: Example

$Y_{it} \longrightarrow \hat{\beta}_{OLS} = 6$

?

	$t=1$	$t=2$	$t=3$	$t=4$
ID 1	0	2	2	2
ID 2	0	0	0	10

Treated cells

Positive weights (in treatment group)

DiD with Staggered Treatment Timing: Example

$Y_{it} \longrightarrow \hat{\beta}_{OLS} = -2$

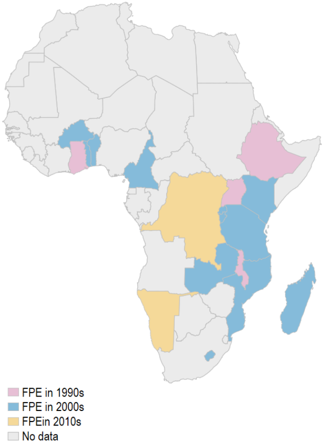
	$t = 1$	$t = 2$	$t = 3$	$t = 4$
ID 1	0	2	2	10
ID 2	0	0	0	2

Treated cells

Positive weights (in treatment group)

TWFE in Practice

Policy Context: Free Primary Education in Sub-Saharan Africa



Country	FPE Year
Benin	2006
Burkina Faso	2007
Burundi	2005
Cameroon	2000
Democratic Republic of Congo	2019
Ethiopia	1995
Ghana	1996
Kenya	2003
Lesotho	2006
Madagascar	2003
Malawi	1994
Mozambique	2005
Namibia	2013
Rwanda	2003
Tanzania	2001
Togo	2008
Uganda	1997
Zambia	2002

Outcome Variables

1. FPE is likely to increase **gross enrollment in primary school** immediately
   - Evidence from single-country studies suggests this (e.g. Lucas and Mbiti 2012)
2. Impacts on **primary school completion** might emerge more slowly over time
   - In first year of FPE, only pre-FPE Grade 8 students will complete primary
   - If FPE reduces the drop-out rate, impacts on completion may increase over time

     $\Rightarrow$ When (positive) impacts grow larger over time, $\hat{\beta}_{twfe}$ may be biased down

TWFE Estimates of Treatment Effects

TWFE specification: $PrimaryEnrollment_{it} = \alpha_i + \gamma_t + \beta FPE_{it} + \varepsilon_{it}$

Stata code: `reg primary fpe i.cid i.year, cluster(cid)`

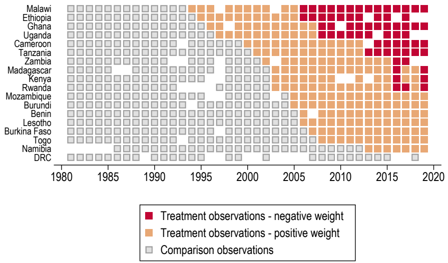
	Dep. Var.: Primary School...	
	ENROLLMENT	COMPLETION
	(1)	(2)
Free primary education	19.85	7.06
	(7.06)	(4.41)
	[0.01]	[0.13]
Country fixed effects	Yes	Yes
Year fixed effects	Yes	Yes

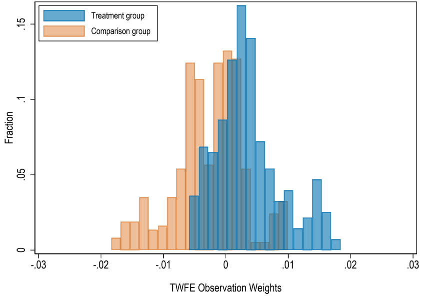
Dependent variable: gross enrollment ratio. Data on gross primary enrollment and primary school completion in 18 countries comes from the World Development Indicators, years 1981 through 2019. Standard errors (clustered at the country level) in parentheses; p-values in brackets.

TWFE Diagnostics

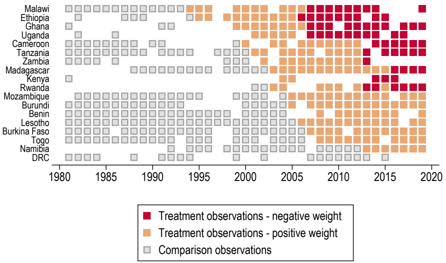
1. Are treated observations getting negative weight in the standard TWFE estimation?
    - Are treated observations (i.e. country-years) being weighted in a sensible way?
2. Are treatment effects (likely to be) heterogeneous? If yes, how?
    - Conceptually: do you expect the treatment effects to vary over time, across units, or both?
    - Do you see evidence contradicting the assumption of homogeneous treatment effects?
        - Imputation-based estimates of treatment effects
        - Event study specifications
3. How do alternative (more robust) estimates of the treatment effect compare to TWFE?

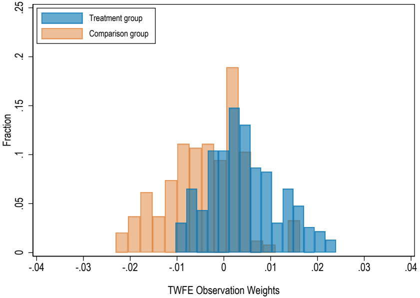
Negative Weights: Gross Primary Enrollment





Negative Weights: Primary School Completion





Negative Weights: Comments

Negative weights are most likely in early-adopter units and later time periods

- Adding never-treated units can eliminate negative weights (common trends?)

Even when all treated observations get positive weight in the calculation of $\hat{\beta}_{twfe}$, staggered treatment timing means that treated observations are not all weighted equally

- Which ATE is desired? Should all observations/units/periods be weighted equally?

Imputing Y_0

Under common trends, $E[Y_{0,it}] = \alpha_i + \gamma_t$

- Gardner et al. 2024 and Borusyak et al. 2024 propose an imputation-based estimator of the average impact of treatment, calculating $Y_{0,it}$ using untreated observations
- Stata: `did2s`

$Y_{it} - E[Y_{0,it}]$ provides an estimate of the (observation-specific) treatment effect

- Is it heterogeneous? (probably)
- Does it vary across units? Over (relative or calendar) time?

An Imputation-Based Estimator

```
did2s primary, first_stage(i.year i.id) second_stage(treatment) treatment(treatment)
cluster(id)
```

- First stage: regress outcomes on fixed effects using only untreated country-years (observations), predict country-specific and year-specific means and subtract them from Y (equivalent to FEs)
- Second stage: regress implicitly residualized Y on treatment dummy

Three simulated examples:

1. Treatment effect is 0
2. Treatment effect is 10
3. Treatment effect is equal to 2 times the number of years since treatment
    - Average treatment effect across treated observations is 17.7

An Imputation-Based Estimator: Example

	Example 1	Example 2	Example 3
TWFE	0.160	10.160	1.155
	(0.919)	(0.919)	(2.126)
	[$t = 0.17$]	[$t = 15.68$]	[$t = 0.54$]
`did2s`	-0.857	9.143	16.517
	(0.771)	(0.701)	(1.083)
	[$t = -1.11$]	[$t = 11.86$]	[$t = 15.26$]

Comparison of TWFE and the Imputation-Based Estimator

	Dep. Var.: Primary School...	
	ENROLLMENT	COMPLETION
	(1)	(2)
Panel A. TWFE		
Free primary education	19.85	7.06
	(7.06)	(4.41)
	[0.01]	[0.13]
Panel B. Gardner et al. (2024) Imputation-Based Estimator		
Free primary education	24.72	18.28
	(5.630)	(2.85)
	[$p < 0.001$]	[$p < 0.001$]
Country fixed effects	Yes	Yes
Year fixed effects	Yes	Yes

Dependent variable: gross enrollment ratio. Data on gross primary enrollment and primary school completion in 18 countries comes from the World Development Indicators, years 1981 through 2019. Standard errors (clustered at the country level) in parentheses; p-values in brackets.

Event Study Specifications

Negative weights are a major issue if treatment effects change over (relative) time

- Relative time is the number of years since treatment was implemented (in country t)
- We can also think of negative relative time as years until treatment starts (in country t)

An **event study** specification allows us to estimate treatment effects for every (relative) time

- Provides direct evidence on the stability of the treatment effect (over time)
- Also allows us to check for violations of common (pre)trends
- Because we are estimating many parameters instead of one, statistical power is an issue
  - A (relatively large) never-treated group is (relatively) important

Event Study Specifications

Let G_i indicate the time t when treatment starts in country i

⇒ $R_{it} = t - G_i$ is relative time, and treatment starts when $R_{it} = 0$

TWFE event study specification:

$$Primary_{it} = \alpha_i + \gamma_t + \sum_{r \leq 2} \beta_r \mathbf{1}\left[R_{it} = r\right] + \sum_{r \geq 0} \delta_r \mathbf{1}\left[R_{it} = r\right] + \varepsilon_{it}$$

Impacts are defined relative to $R_{it} = -1$, the last period before treatment

⇒ Relies on never-treated group for identification (graph shows data to 2012)

⇒ Can still be biased when treatment effect heterogeneity is **not** over relative time

The Impact of Free Primary on Completion: Event Study