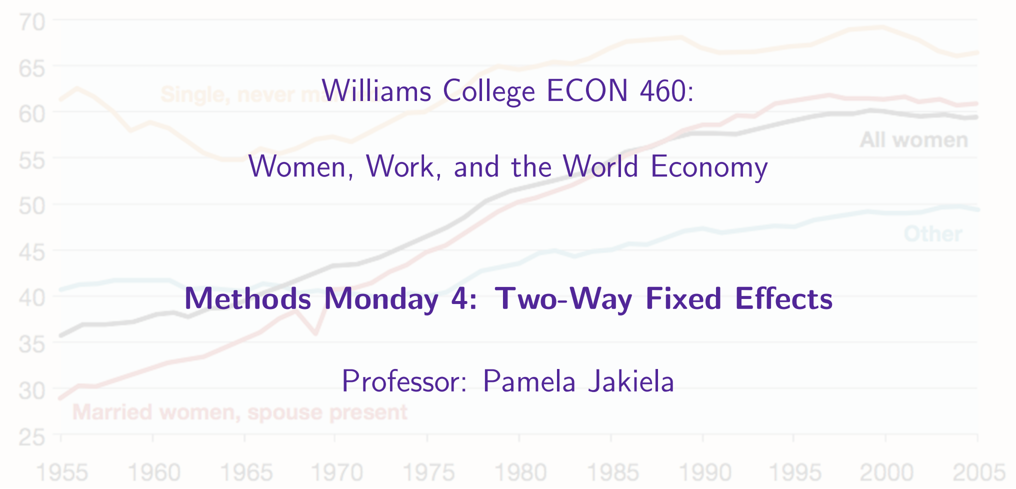
Williams College ECON 460:

Women, Work, and the World Economy

Methods Monday 4: Two-Way Fixed Effects

Professor: Pamela Jakiela



Source: Our World in Data

Outline

- 2×2 difference-in-differences
- Difference-in-differences in a panel data framework (fixed effects)
- Two-way fixed effects with staggered treatment timing

False Counterfactuals

Pre vs. Post Comparisons:

- **Compares:** same units before vs. after program implementation
- **Drawback:** does not control for time trends (in potential outcomes without treatment)

Participant vs. Non-Participant Comparisons:

- **Compares:** participants to those who choose not to participate in a program
- **Drawback:** potential for selection bias (participants differ from non-participants)

Neither approach provides credible estimates of program impacts

Two Wrongs Sometimes Make a Right

Difference-in-differences combines the two (flawed) false counterfactual approaches

- Observe self-selected treatment, comparison groups before and after treatment (i.e. before and after the treatment group participates in the program)
- May overcome problems of both false counterfactual approaches when:
  - Selection bias relates to fixed characteristics of units
  - Time trends are common to treatment and comparison groups

The difference-in-differences (or diff-in-diff, DD, or DiD) estimator is:

$$DD = \bar{Y}_{post}^{treatment} - \bar{Y}_{pre}^{treatment} - \left(\bar{Y}_{post}^{comparison} - \bar{Y}_{pre}^{comparison}\right)$$

Difference-in-Differences Estimation

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

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

Where:

- D_i = treatment dummy
- $Post_i$ = dummy for post-treatment period
- $D_i * Post_i$ = interaction term

	comparison	treatment
pre	$\bar{Y}_{pre}^{comparison}$	$\bar{Y}_{pre}^{treatment}$
post	$\bar{Y}_{post}^{comparison}$	$\bar{Y}_{post}^{treatment}$

When to Use 2×2 Diff-in-Diff

The simple 2×2 regression equation is rarely used in practice

- To assess the common trends assumption, more than two periods are required
  - With multiple pre-treatment periods, the constant in the 2×2 specification would capture the average outcome in the never-treated comparison group in the pre-treatment period
  - There may be considerable temporal variation across pre-treatment time periods
  - Explaining that variation (which is not correlated with treatment) will reduce standard errors

Generalized Diff-in-Diff with Fixed Effects

Widely used panel data diff-in-diff specification:

$$Y_{i,t} = \alpha + \gamma D_{i,t} + \delta \left(D_{i,t} \times Post_t \right) + \nu_t + \varepsilon_{i,t}$$

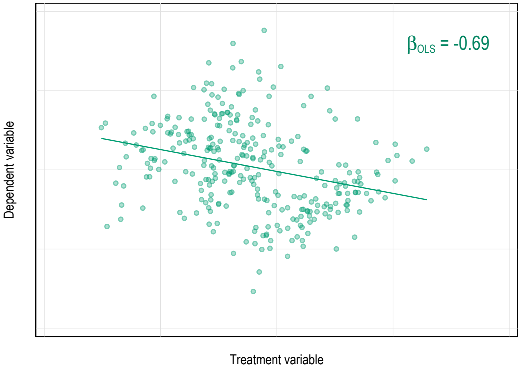
where:

- $D_{i,t}$ = dummy for ever-treated group/unit
- δ = diff-in-diff estimate of treatment effect
- ν_t = time-period fixed effects

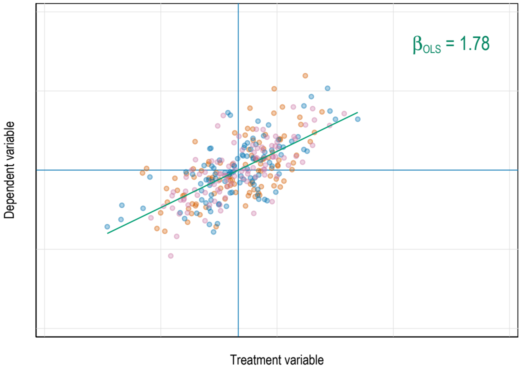
What Are Fixed Effects?

- Individual dummy variables for mutually exclusive groups in your data
  - Dummy for male or female
  - Age (or age group) fixed effects
  - Continent/country/state/district fixed effects
  - Year fixed effects
- Why use fixed effects?
  - Estimation using **within** rather than **between** variation
- We often use multiple sets of fixed effects in empirical work

Simpson's Paradox



What Do Fixed Effects Do?



What Do Fixed Effects Do?

- Fixed effects are equivalent to:
  - Transforming both independent and dependent variables by subtracting off the mean in each group and adding back the mean in the omitted category (the blue group in the figure)
    - Equivalently: subtracting off the difference in means between group and omitted group
  - Runing OLS in your transformed (i.e. re-centered) data
- Fixed effects mattered because treatment varied across groups
  - When treatment doesn't vary, FEs can improve precision but won't change slope estimate
- If you regress X and Y on FEs, de-meaned variables are the residuals

The Frisch-Waugh-Lovell Theorem

$$Y = \alpha + \beta X + \gamma Z$$

is equivalent to

$$\tilde{Y} = \alpha + \beta \tilde{X}$$

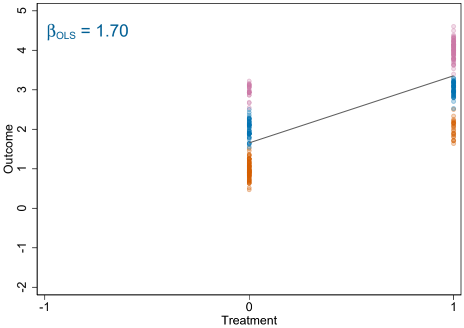
where

$\tilde{Y}$ = residuals from regressing Y on Z

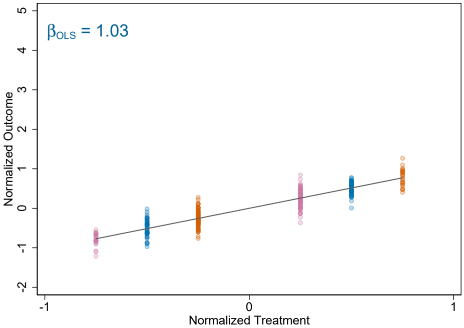
$\tilde{X}$ = residuals from regressing X on Z

Fixed Effects with Binary Treatment: Example

Without Fixed Effects



With Fixed Effects



Diff-in-Diff with Time Fixed Effects

	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$
Unit 1	0	0	0	0	0
Unit 2	0	0	0	0	0
Unit 3	0	0	0	1	1
Unit 4	0	0	0	1	1
Unit 5	0	0	0	1	1
$\bar{D}_t$	0	0	0	0.6	0.6

Time fixed effects:

$\Rightarrow$ Subract off mean $D_{i,t}$

Diff-in-Diff with Time Fixed Effects

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$
Unit 1	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$
Unit 2	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$
Unit 3	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$
Unit 4	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$
Unit 5	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$	$\tilde{Y}_{i,t}$
$\bar{Y}_t$	$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$	$\bar{Y}_4$	$\bar{Y}_5$

Time fixed effects:

$\Rightarrow$ Subract off mean $D_{i,t}$

Equivalent to regression on:

$\tilde{D}_{i,t} = D_{i,t} - \bar{D}_t$

With dependent variable:

$\tilde{Y}_{i,t} = Y_{i,t} - \bar{Y}_t$

Diff-in-Diff with Time Fixed Effects

Why used time fixed effects (instead of dummy for post-treatment)?

- Fixed effects "soak up" period-specific shocks better
  - Smaller residuals $\Rightarrow$ smaller standard errors $\Rightarrow$ statistical power
- Inclusion of time fixed effects yield should not lead to substantial changes in coefficients

Two-way fixed effects specification:

$$Y_{i,t} = \alpha + \eta_i + \nu_t + \delta D_{i,t} + \varepsilon_{i,t}$$

where η_i is an individual FE, ν_t is a time FE, and δ is DD estimator

Use two-way fixed effects with caution when treatment starts at different times in different units, treatment is continuous, or variance of treatment differs across treated units for other reasons, as we discuss further in the next module.

Example: States Adopted Medicaid at Different Times

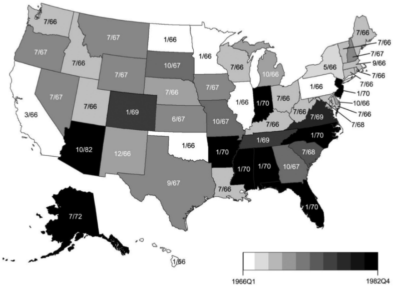


Figure 2.
Medicaid Adoption by Quarter
Notes: Adoption dates come from the Department of Health Education and Welfare (1970) & Social Security Administration (2013). The map is shaded relative to the quarter of adoption and states are labeled with the month and year of adoption.

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

Example: Counties Opening Community Health Centers

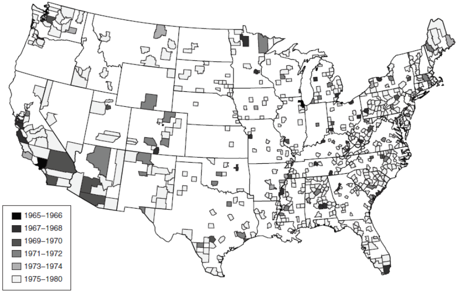


Figure 3. Establishment of Community Health Centers by County of Service Delivery, 1965–1980

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

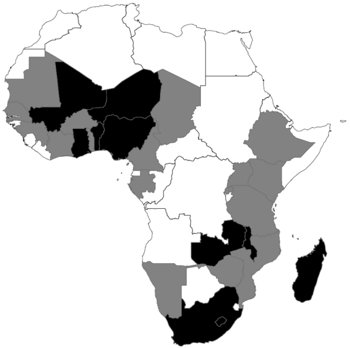


FIGURE A.2. Geographical distribution of democratized countries since 1990. Black-colored countries are democratized since 1990; grey-colored countries the other countries in the sample for infant mortality analysis. Democratized countries include the Comoros, tiny islands to the northwest of Madagascar,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

source: Kudamatsu (JEEA, 2012)

Two-Way Fixed Effects Estimates of β^{DD}

What exactly is β^{DD}?

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

unit fixed effects

time fixed effects

treatment dummy

that turns "on" at different times

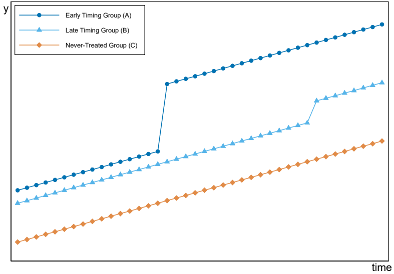
What exactly is β^{DD} in TWFE?

	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	
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

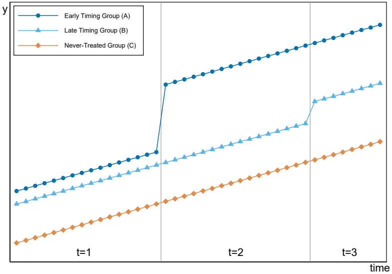
	$t=1$	$t=2$	$t=3$	$t=4$	$t=5$	$t=6$	$t=7$	$t=8$
ID 1	0	0	1	1	1	1	1	1
ID 2	0	0	1	1	1	1	1	1
ID 3	0	0	0	0	0	1	1	1
ID 4	0	0	0	0	0	1	1	1
ID 5	0	0	0	0	0	0	0	0
ID 6	0	0	0	0	0	0	0	0

Decomposition into Timing Groups



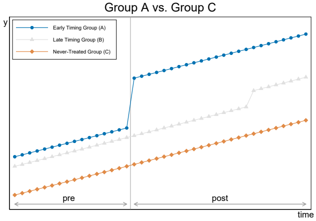
Panel with variation in treatment timing can be decomposed into distinct **timing groups** reflecting observed start of treatment

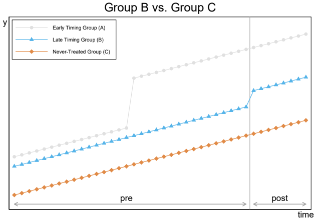
Decomposition into Timing Groups

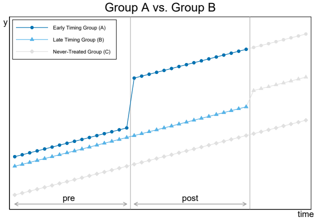


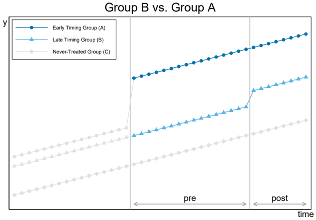
Example: with three timing groups (one of which is never treated), can construct three timing windows (pre, middle, post or $t = 1, 2, 3$)

Decomposition into Standard 2 × 2 DDs

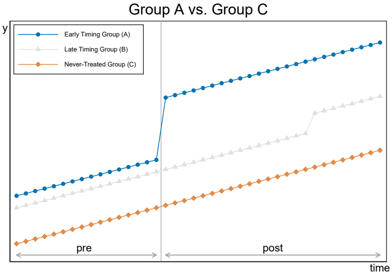








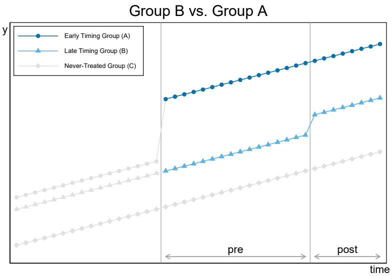
Decomposition into Standard 2 × 2 DDs



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

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

Forbidden Comparisons



Some 2×2 diff-in-diffs (implicitly) use already-treated groups as the comparison

Two-Way Fixed Effects β^{DD} as a Weighted Sum

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

- Some use an already-treated group as comparison
    - Creates problems if treatment effect grows/changes over time
    - TWFE imposes a model of homogeneous treatment effects
    - When treatment effects evolve over time, model is mis-specified

We can use Frisch-Waugh-Lovell to construct the TWFE/OLS weights used to generate β^{DD}

- 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 $\tilde{D}_{it}$

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

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

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

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

Diff-in-Diff 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

Diff-in-Diff 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

Diff-in-Diff 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

Diff-in-Diff 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)

Diff-in-Diff with Staggered Treatment Timing: Example

Y_{it} ⟶ $\hat{\beta}_{OLS} = 10$

	$t = 1$	$t = 2$	$t = 3$	$t = 4$
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)

Diff-in-Diff 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)

Diff-in-Diff with Staggered Treatment Timing: Example

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

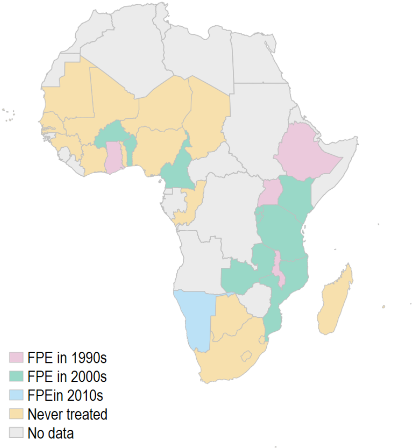
	$t = 1$	$t = 2$	$t = 3$	$t = 4$
ID 1	0	2	2	10
ID 2	0	0	0	2

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
Ethiopia	1995
Ghana	1996
Kenya	2003
Lesotho	2006
Malawi	1994
Mozambique	2005
Namibia	2013
Rwanda	2003
Tanzania	2001
Uganda	1997
Zambia	2002

TWFE Specification: Countries that Implemented Free Primary

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

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

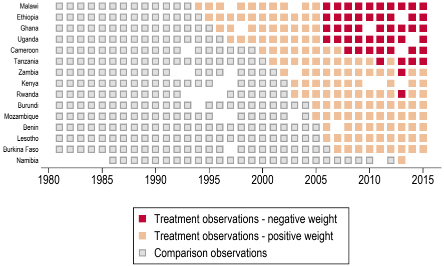
	(1)
	Enrollment
Free primary education	20.428
	(9.120)
	[0.042]
Country fixed effects	Yes
Year fixed effects	Yes
Never treated	No

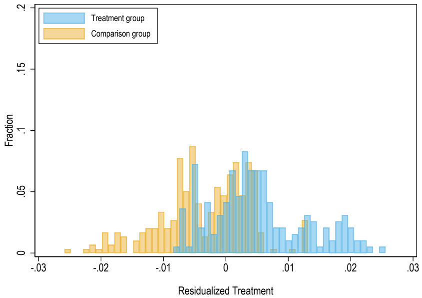
Dependent variable: gross enrollment ratio. Data on gross enrollment ratio in 15 countries comes from the World Development Indicators, years 1981 through 2015. Standard errors (clustered at the country level) in parentheses; p-values in square brackets.

TWFE Diagnostics

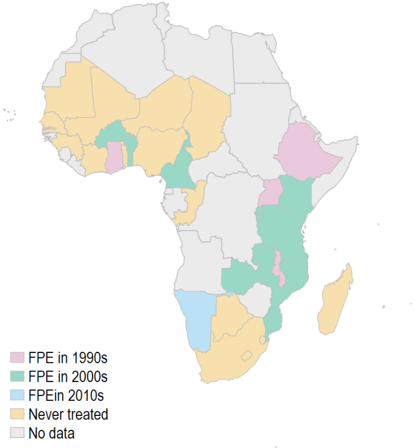
1. Are treated observations getting negative weight in my TWFE estimation?
   - Are treated observations (i.e. country-years) being weighted in a sensible way?
2. Are treatment effects (likely to be) heterogeneous? If yes, how?
   - Conceptually: do you expect the treatment effects to vary over time, across units, or both?
   - Do you see evidence contradicting the assumption of homogeneous treatment effects?
     - Event study specifications
     - Scatter plots of residuals
   - Are your estimated treatment effects robust across specifications?

Negative Weights





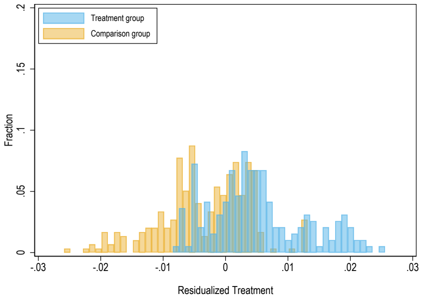
Including Never Treated Countries to Eliminate Negative Weights

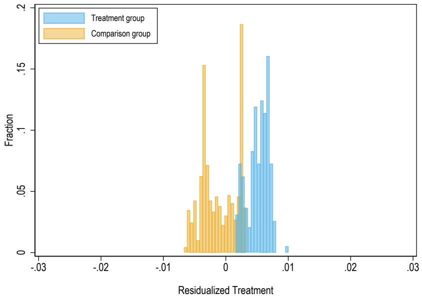


	(1) Enrollment	(2) Enrollment
Free primary education	20.428	18.951
	(9.120)	(6.733)
	[0.042]	[0.008]
Country fixed effects	Yes	Yes
Year fixed effects	Yes	Yes
Never treated	No	Yes

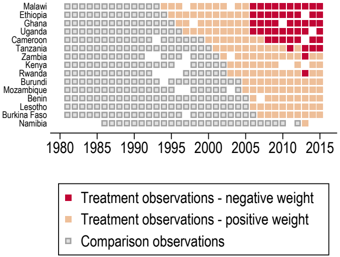
Dependent variable: gross enrollment ratio. Data on gross enrollment ratio in 15 countries comes from the World Development Indicators, years 1981 through 2015. Standard errors (clustered at the country level) in parentheses; p-values in square brackets.

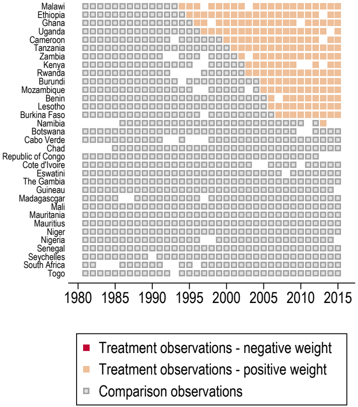
Including Never Treated Countries to Eliminate Negative Weights





Including Never Treated Countries to Eliminate Negative Weights





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 timet)
- 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

Event Study Specifications

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

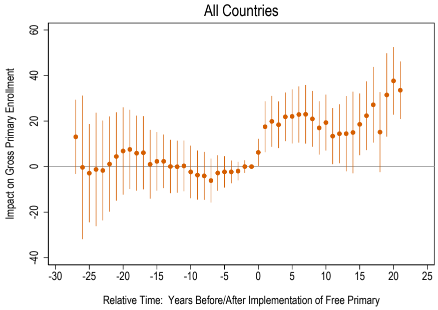
$\Rightarrow R_{it} = t - G_i$ is relative time, and treatment starts when $R_{it} = 0$

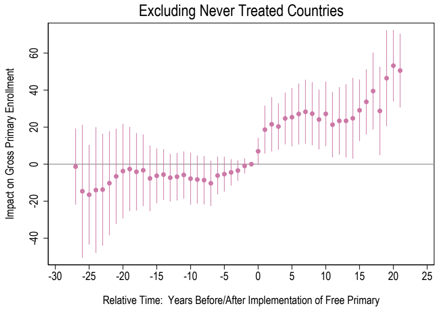
TWFE event study specification:

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

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

The Impacts of Free Primary: Event Study Specifications





The End!