STAGGERINGLY PROBLEMATIC: STAGGERED DID ISSUES AND SOLUTIONS

2023 AUDIT MIDYEAR

JOHN M. BARRIOS | WASH U – OLIN & NBER

OVERVIEW

I. Philosophy of Causality and Methodology

II. DiD refresher
   a) Canonical Model
   b) Staggered -Time Variation in Treatment

III. Issues with Staggered DiD

IV. Overview of potential solutions
   a) Bacon Goodman – timing diagnostics and weights
   b) Stacking Event Studies

PHILOSOPHICAL POINT ON CAUSALITY AND METHODOLOGY

While not all accounting research papers may explicitly estimate causal quantities, the implications and takeaways of the research often have a causal interpretation.

Causality lies at the heart of every exercise

The concept of causality has deep philosophical roots

“We do not have knowledge of a thing until we have grasped its why, that is to say, its cause.”

-Aristotle

If Aristotle didn’t settle the question, unlikely will researchers in the 21t century will

PHILOSOPHICAL POINT ON CAUSALITY AND METHODOLOGY

Accounting research is messy, and a careful discussion of causality entails two dimensions:

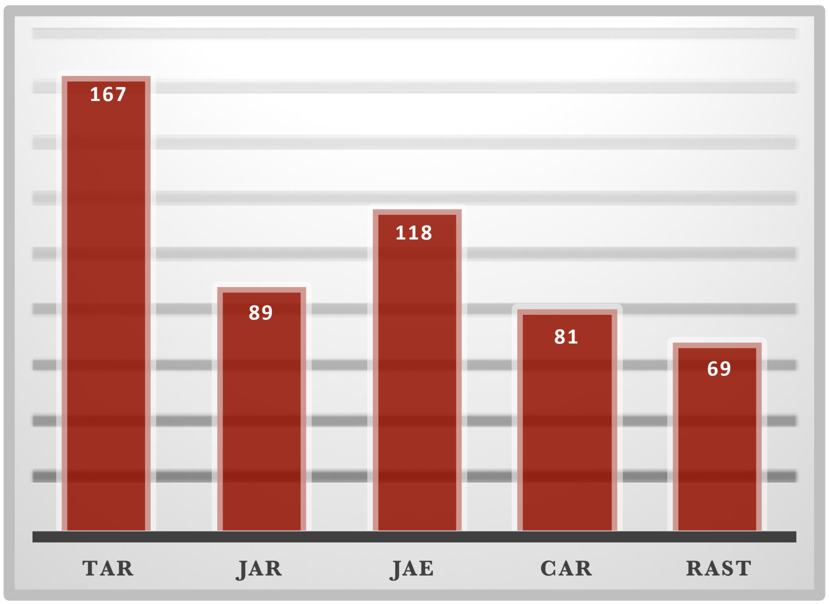
1. Good designs that articulate your assumptions
2. Readers that understand the frameworks

In a nutshell: any research design requires assumptions and we should be ***clear and transparent*** on what we are assuming and be ***honest*** as writers ***as well as readers*** of academic research

Keywords in Published Papers Since 2011

2011/2020 TAR/JAE/JAR/CAR/RAST published 500+ DD papers:

30% had timing variation mentioned



"difference-in-differences" | "differences-in-differences" | "diff-in-diff"

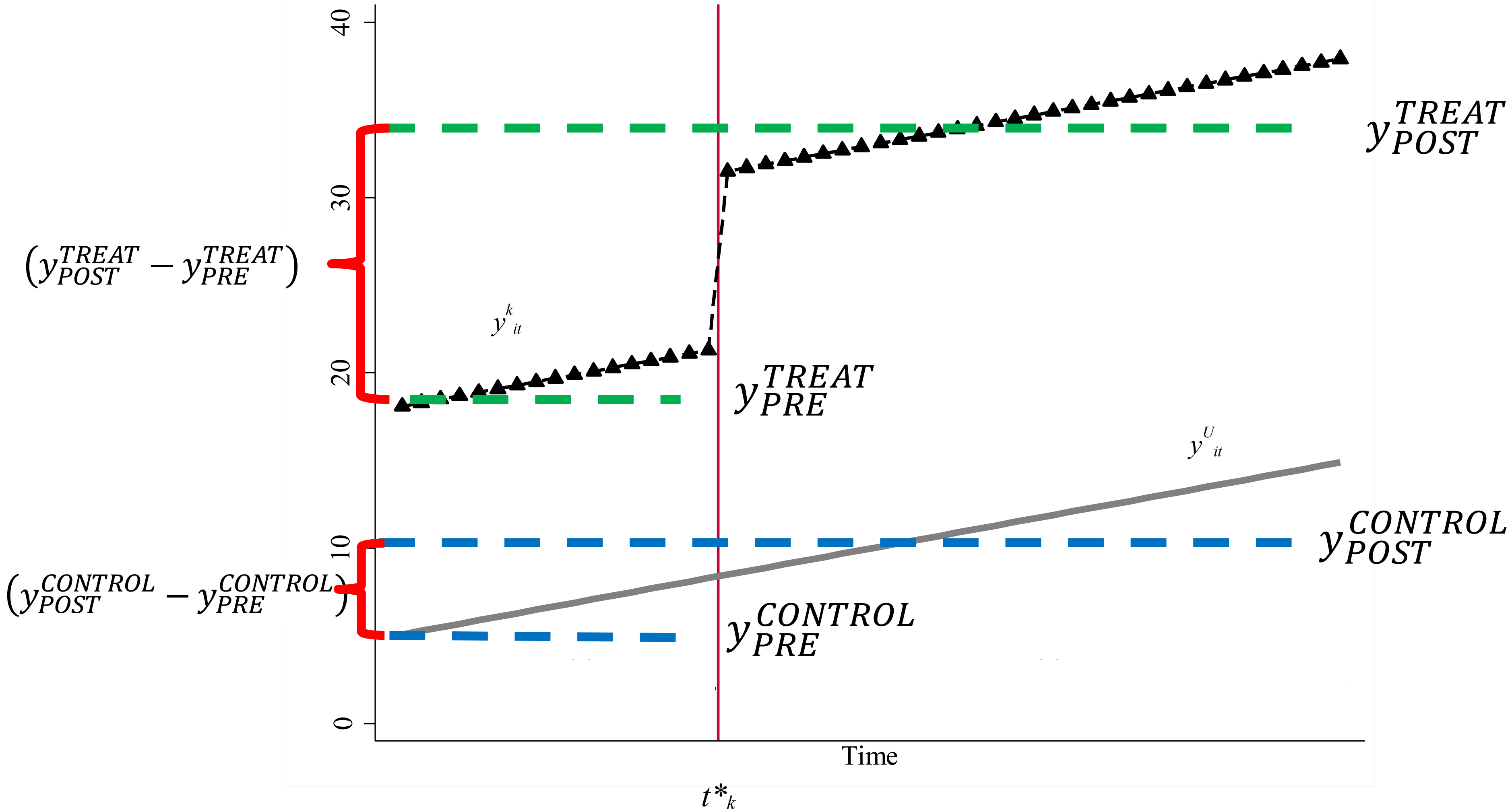
BARRIOS (2021)

DIFF – IN – DIFF

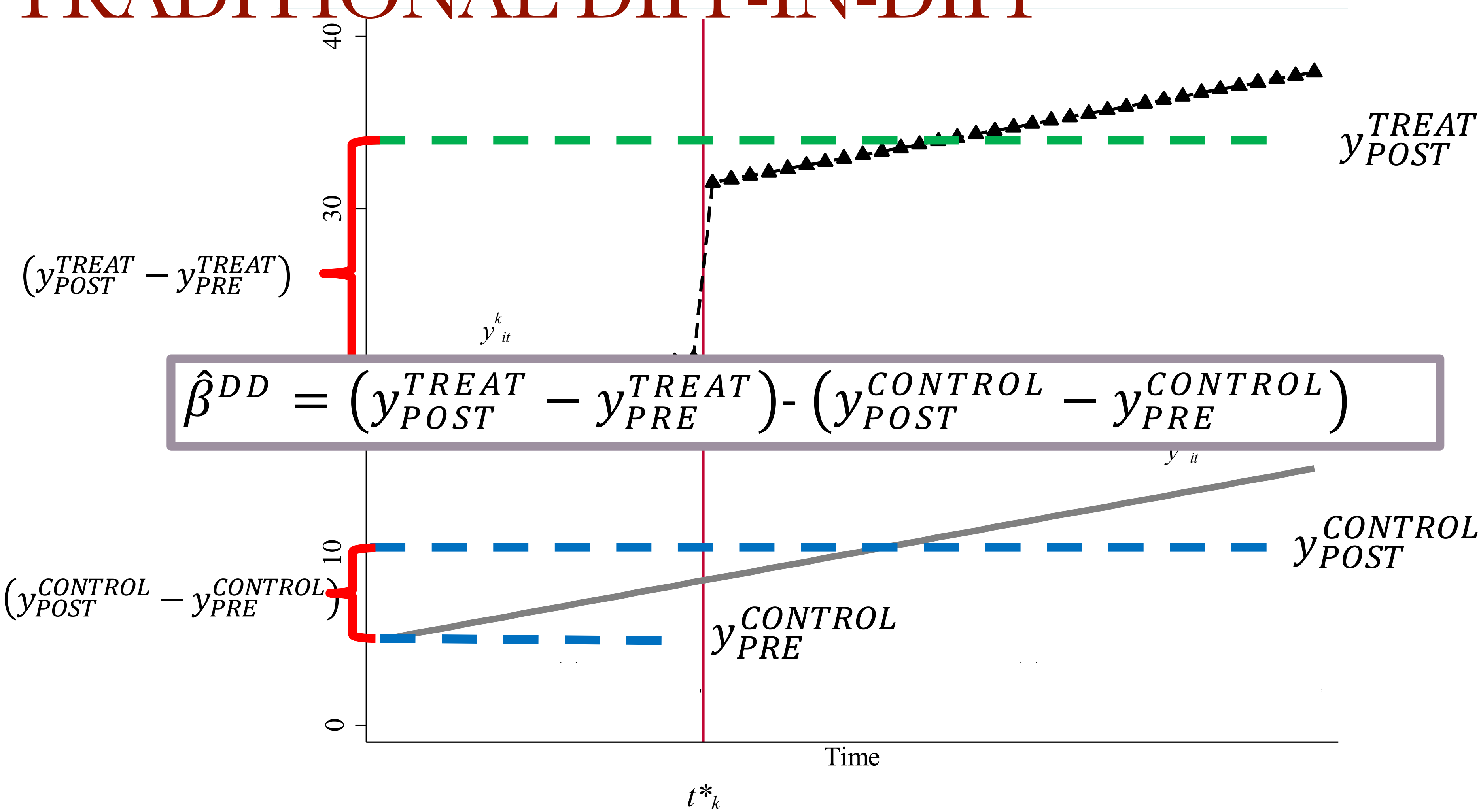
One of the most popular identification strategies in applied accounting work today

- More than 500 published papers use DiD
  - Top 5 Accounting journals 2011-2020

TRADITIONAL DIFF-IN-DIFF



TRADITIONAL DIFF-IN-DIFF

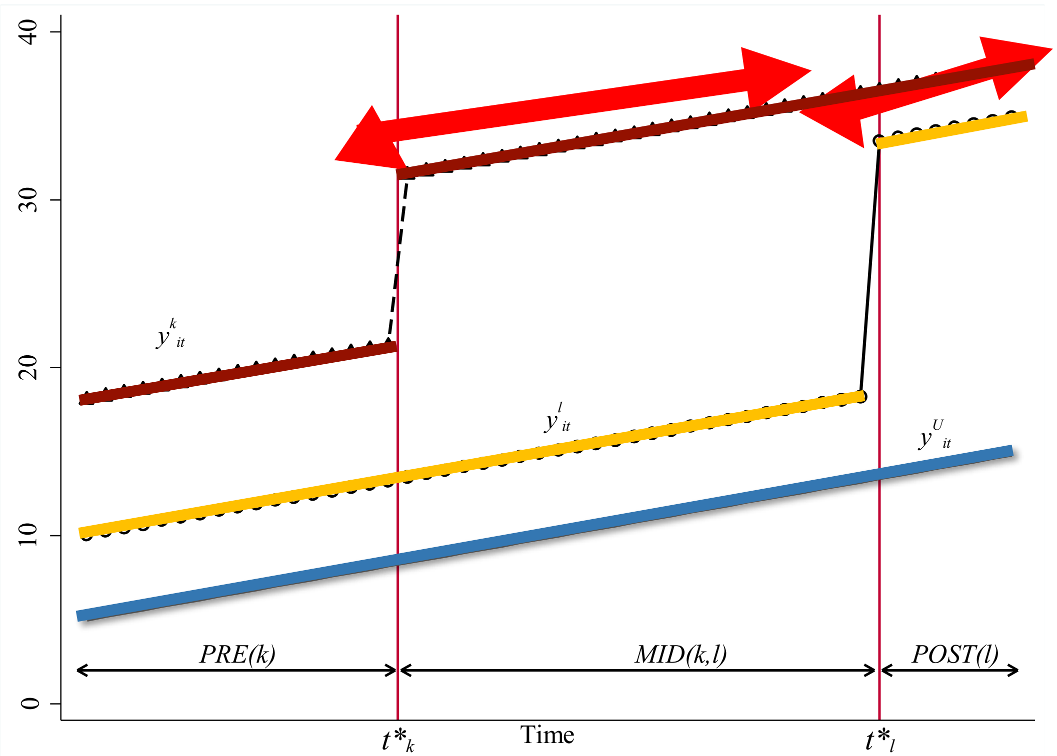


$$\hat{\beta}^{DD} = \left(y_{POST}^{TREAT} - y_{PRE}^{TREAT}\right) - \left(y_{POST}^{CONTROL} - y_{PRE}^{CONTROL}\right)$$

FROM TRADITIONAL 2X2 TO STAGGERED DIFF-IN-DIFF

However, most DiD applications exploit variation in the *timing* of *treatments across groups* and *units*

D_{it} turns on at different times t^*



BARRIOS (2021)

VARIATION IN TIMING

Cases where the treatment turns on at a different time for different units

- IPOS
- Big-4 Audits
- Law Changes
- Natural disasters
- Mass layoffs, …

FROM TRADITIONAL 2X2 TO STAGGERED DIFF-IN-DIFF

To test, most paper run a two-way fixed effects (TWFE) estimator

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

The coefficient that comes from this specification is ***not as easily interpretable as the 2x2 coefficient*** discussed before

Numerous papers have now begun to examine the interpretation of this coefficient

RECENT RESEARCH ON DiD WITH TIMING

Abraham, Sarah, and Liyang Sun. 2018. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." Working Paper.

Athey, Susan, and Guido W. Imbens. 2018. "Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption." Working Paper.

Bitler, Marianne P., Jonah B. Gelbach, and Hilary W. Hoynes. 2003. "Some Evidence on Race, Welfare Reform, and Household Income." The American Economic Review 93 (2):293-298. doi: 10.2307/3132242.

Borusyak, Kirill, and Xavier Jaravel. 2017. "Revisiting Event Study Designs." Harvard University Working Paper.

Callaway, Brantly, and Pedro Sant'Anna. 2018. "Difference-in-Differences With Multiple Time Periods." Journal of Econometrics, 2020

de Chaisemartin, C., and X. D'HaultfŒuille. 2018a. "Fuzzy Differences-in-Differences." The Review of Economic Studies 85 (2):999-1028. doi: 10.1093/restud/rdx049.

de Chaisemartin, C.;, and X. D'HaultfŒuille. 2018b. "Two-way fixed effects estimators with heterogeneous treatment effects." Working Paper.

Freyaldenhoven, Simon, Christian Hansen, and Jesse M. Shapiro. 2018. "Pre-event Trends in the Panel Event-study Design." National Bureau of Economic Research Working Paper Series No. 24565. doi: 10.3386/w24565.

Gibbons, Charles, E., Juan Carlos Suárez Serrato, and Michael Urbancic, B. 2018. Broken or Fixed Effects? In Journal of Econometric Methods.

Goodman-Bacon, Andrew. 2018. "Difference-in-Differences with Variation in Treatment Timing." *National Bureau of Economic Research Working Paper Series*

Imai, Kosuke, In Song Kim, and Erik Wang. 2018. "Matching Methods for Causal Inference with Time-Series Cross-Section Data." Working Paper.

Krolikowski, Pawel. 2017. "Choosing a Control Group for Displaced Workers." ILR Review:0019793917743707. doi: 10.1177/0019793917743707.

Sant'Anna, Pedro H. C. and Jun Zhao, "Doubly robust difference-in-differences estimators," Journal of Econometrics, November 2020, 219 (1), 101–122.

Sloczynski, Tymon. 2017. "A General Weighted Average Representation of the Ordinary and Two-Stage Least Squares Estimands." Working Paper.

Strezhnev, Anton. 2018. "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs." *Working Paper*.

Wooldridge, Jeffrey M. 2005. "Fixed-Effects and Related Estimators for Correlated Random-Coefficient and Treatment-Effect Panel Data Models." The Review of Economics and Statistics 87 (2):385-390.

Two-Way Fixed Effects Estimator

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

Unit fixed effects

Time fixed effects

Treatment dummy

What is $\widehat{\beta}^{DD}$?

This has been unclear, but we do have good intuition for subsamples

EXAMPLE OF THE TWFE ESTIMATOR IN THE CASE OF THREE GROUPS:

1. Never Treated
2. Treated Early
3. Treated Later

FROM TRADITIONAL 2X2 TO STAGGERED DIFF-IN-DIFF

Most DiD applications exploit variation in the timing of treatments across groups and units

Yet most paper run a two-way fixed effects (TWFE) estimator when there are more than two units and periods

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

The coefficient that comes from this not easily interpretable parameter as the 2x2 coefficient discussed before

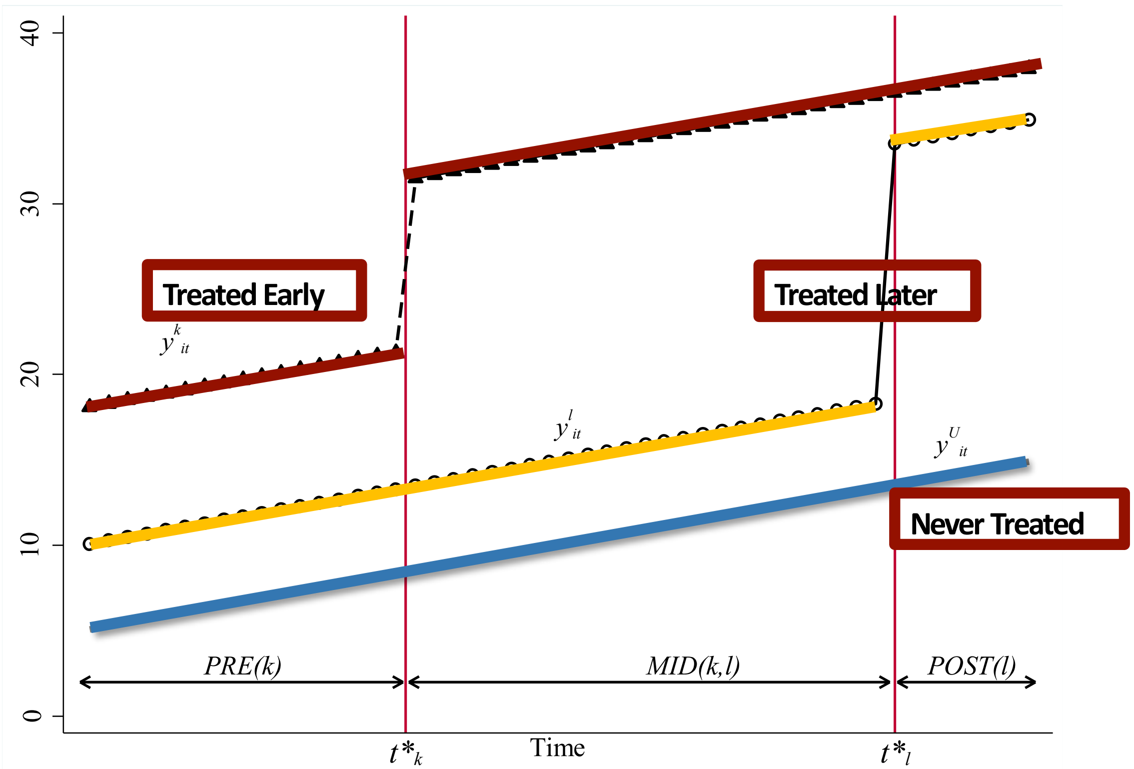
Consensus that the coefficient is a weighted average of many different 2x2 treatment effects

These weights are often negative and non-intuitive requiring accounting researchers to further examine what the $\hat{\beta}^{DD}$ is reflecting

Follow Goodman Bacon 2019 to explain the interpretation of $\hat{\beta}^{DD}$ as

a weighted average of all possible 2x2 DiD estimators in the dataset

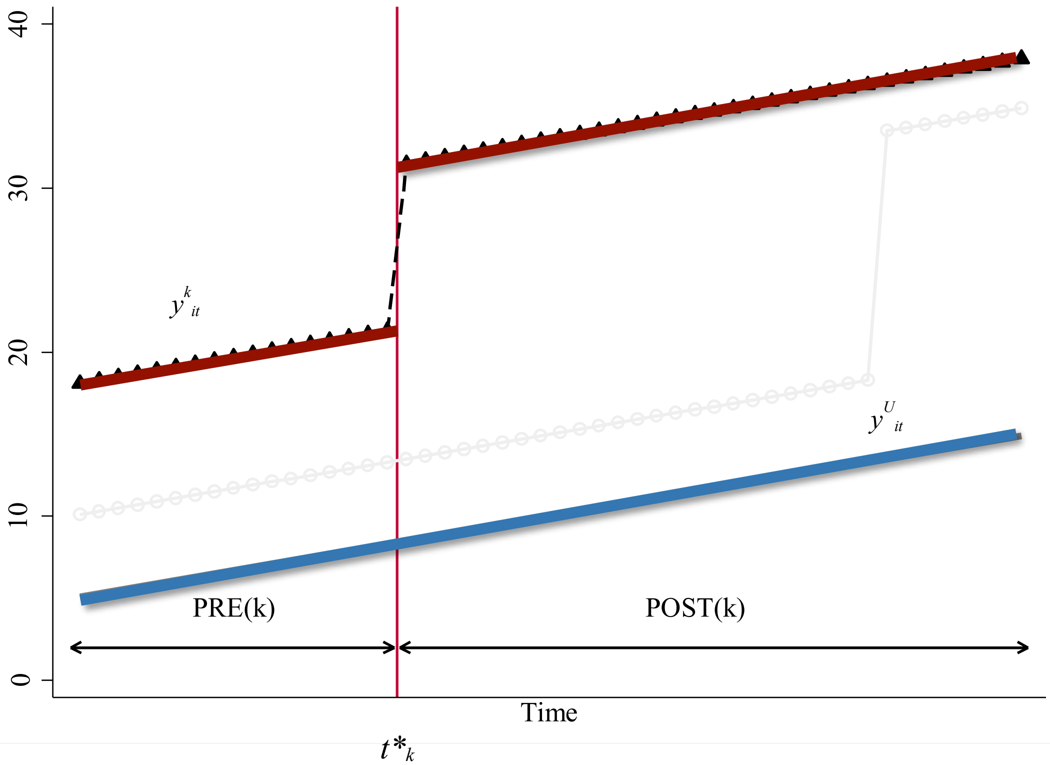
2x2 DDs: subsamples with two groups (treat/control) and two periods (pre/post)



BARRIOS (2021)

EXAMPLE OF THE TWFE ESTIMATOR IN THE CASE OF THREE GROUPS:

1. Never Treated
2. Treated Early
3. Treated Later



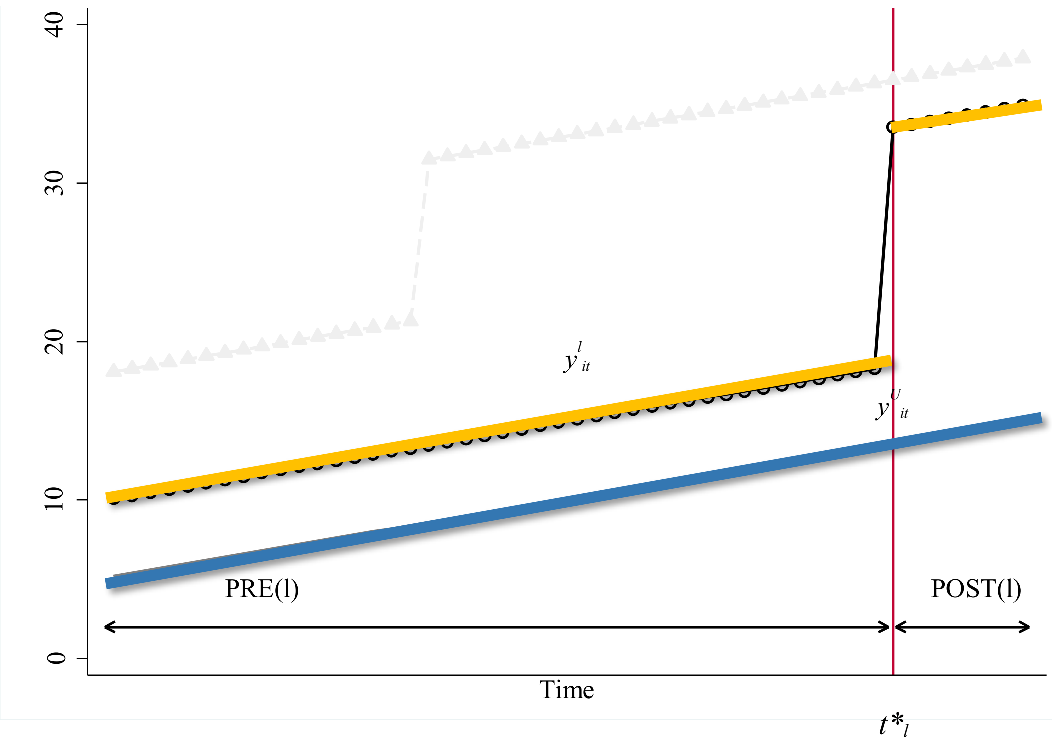
BARRIOS (2021)

1ST - 2X2 COMPARISON

TREATED EARLY VS NEVER TREATED

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

$$\widehat{\beta}^{2x2}_{kU} = \left(y_k^{POST(k)} - y_k^{PRE(k)}\right) - \left(y_U^{POST(k)} - y_U^{PRE(k)}\right)$$



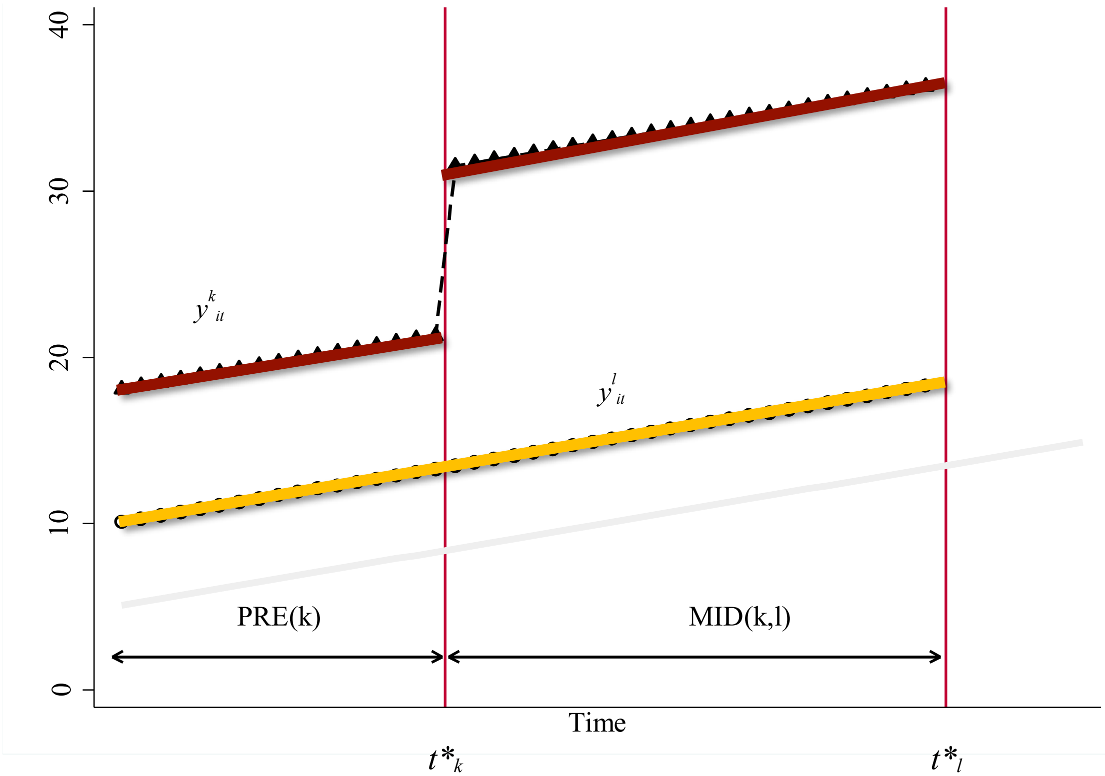
BARRIOS (2021)

2ND - 2X2 COMPARISON

TREATED LATE VS NEVER TREATED

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

$$\widehat{\beta}^{2x2}_{\ell U} = \left(y_{\ell}^{POST(\ell)} - y_{\ell}^{PRE(\ell)}\right) - \left(y_{U}^{POST(\ell)} - y_{U}^{PRE(\ell)}\right)$$



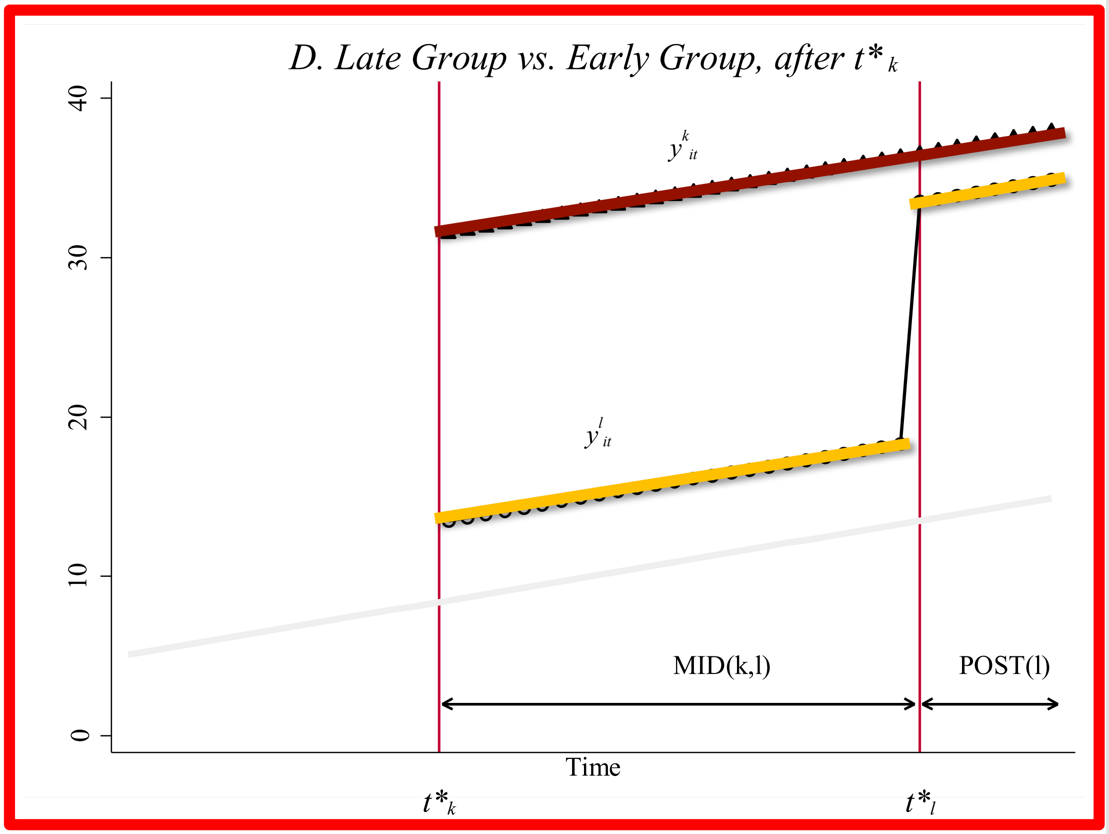
BARRIOS (2021)

3RD - 2X2 COMPARISON

TREATED EARLY VS TREATED LATE PRE t*

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

$$\widehat{\beta}^{2x2,k}_{k\ell} = \left(y_k^{MID(k,\ell)} - y_k^{PRE(k)}\right) - \left(y_\ell^{MID(k,\ell)} - y_\ell^{PRE(k)}\right)$$



BARRIOS (2021)

4^{TH} - 2X2 COMPARISON

TREATED LATE
VS
TREATED EARLY AFTER t^*

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

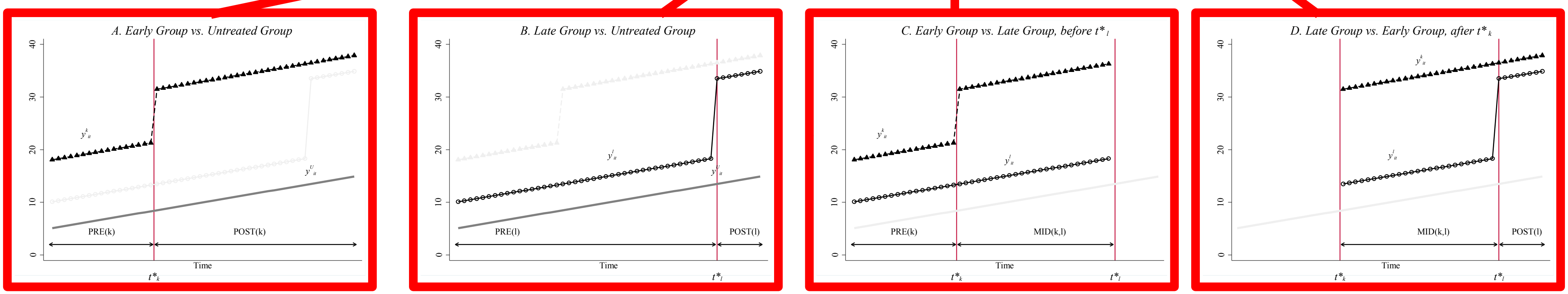
$$\hat{\beta}_{k\ell}^{2x2,\ell} = \left(y_\ell^{Post(l)} - y_\ell^{MID(k,\ell)}\right) - \left(y_k^{Post(\ell)} - y_k^{MID(k,\ell)}\right)$$

DIFFERENCE-IN-DIFFERENCES DECOMPOSITION (3 GROUP CASE)

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

For three groups:

$$\hat{\beta}^{DD} = s_{kU} \boxed{\hat{\beta}^{2x2}_{kU}} + s_{\ell U} \boxed{\hat{\beta}^{2x2}_{\ell U}} + s^{k}_{k\ell} \boxed{\hat{\beta}^{2x2,k}_{k\ell}} + s^{k}_{k\ell} \boxed{\hat{\beta}^{2x2,\ell}_{k\ell}}$$



WHAT DO WE LEARN FROM THE DECOMPOSITION
COMPARISON OF THE 2X2S

1. Helps clarify what comparisons are being made when we run the regression:

   "switchers vs untreated"?

   "early switchers vs late switchers"?

   "late vs early" (this is less obvious)?

   It's all of these

2. "What is the control group?"
   Every group acts as a control (sometimes)

3. Help clarifies theory
   We understand the estimand (ATET)
   Identification assumption (common trends) for each 2x2
   late vs. early comparisons are biased if treatment effects vary over time

DIFFERENCE-IN-DIFFERENCES DECOMPOSITION (3 GROUP CASE)

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

For three groups:

$$\hat{\beta}^{DD} = \boxed{s_{kU}} \hat{\beta}^{2x2}_{kU} + \boxed{s_{\ell U}} \hat{\beta}^{2x2}_{\ell U} + \boxed{s^{k}_{k\ell}} \hat{\beta}^{2x2,k}_{k\ell} + \boxed{s^{\ell}_{k\ell}} \hat{\beta}^{2x2,\ell}_{k\ell}$$

A. Early Group vs. Untreated Group

Size:

$$(n_k + n_U)^2$$
$$\times$$

Variance:

$$n_{kU}(1 - n_{kU}) D_k (1 - D_k)$$

B. Late Group vs. Untreated Group

Size:

$$(n_\ell + n_U)^2$$
$$\times$$

Variance:

$$n_{\ell U}(1 - n_{\ell U}) D_\ell (1 - D_\ell)$$

*C. Early Group vs. Late Group, before t**

Size:

$$\left((n_k + n_\ell)(1 - D_\ell)\right)^2$$
$$\times$$

Variance:

$$n_{k\ell}(1 - n_{k\ell}) \frac{D_k - D_\ell}{1 - D_\ell} \frac{1 - D_k}{1 - D_\ell}$$

*D. Late Group vs. Early Group, after t**

Size:

$$\left((n_k + n_\ell) D_k\right)^2$$
$$\times$$

Variance:

$$n_{k\ell}(1 - n_{k\ell}) \frac{D_\ell}{D_k} \frac{D_k - D_\ell}{D_k}$$

Weights: $\dfrac{(\textit{subsample share})^2 (\textit{subsample variance of FE-adjusted } D)}{\textit{total variance of FE-adjusted } D}$

WHAT DO WE LEARN FROM THE DECOMPOSITION WEIGHTS

1. Relative importance of each kind of comparison.

More important if bigger sample size or treated closer to middle of the panel (i.e. bigger variance).

"How much" comes from timing vs comparisons to untreated.

2. Importance of specific 2x2 DDs.

Sometimes a few of the 2x2 DD terms dominate.

3. Clarifies theory

The estimand and ID assumption are "variance weighted";

Can compare estimand to "parameters of interest" and conduct a proper balance test

POTENTIAL ISSUES WITH TWFE ESTIMATION AND STAGGERED DID

1st Two-way fixed effects estimates of DiD relying on staggered treatments only recover the average treatment effect ***when treatment effects are homogeneous***.

- When treatment effects are ***heterogeneous*** across firms, OLS ***over-weights units with more variance in treatment status*** in order to achieve a more precise estimate of the treatment effect
- Thus, if you want the ***average treatment effect on the treated*** - some sort of ***re-weighting is required***

2nd DiD estimates are biased when treatment effects change over time within unit.

- Occurs because already treated units serve as controls in some of the 2X2 DiDs underlying the weighted average
- Thus, when treatment effects are not constant over time **using already treated units as controls biases estimates of the treatment effect**
  - Basically, you are *differencing a change in the treatment effect* by using the already treated units

WHY MIGHT TREATMENT EFFECTS VARY OVER TIME?

Phase in effects: it might take some time for an intervention to produce measurable changes in behavior.

- Knowledge of a new tax might spread gradually. People might not change consumption or production behavior until they figure it out.
- Marijuana might be legalized in one year, but it will take a few years for dispensaries to open.

Calendar Period x Treatment Effect Interactions: the same policy may have a different effect on behavior in some calendar periods than others.

- Expanded Unemployment Insurance benefits might have negligible effects on labor supply when the economy is depressed. But they could have larger effects on labor supply when the economy is booming.

PROPOSED SOLUTIONS

Callaway & Sant'Anna

Develop an **inverse propensity weighted long-difference in cohort-specific average treatment effects** between treated and untreated units for a given treatment cohort

Abraham & Sun

Extension of the standard event-study TWFE model where we **saturate the event time indicators (i.e. t = -2, -1, …) with indicators for the treatment initiation year group and sum to overall aggregate event time indicators by cohort size**

Cengiz et al. – Stacking Procedure

Akin to matching treated to control firms that do not experience treatment within the pre- and post-window as defined by the researcher

Stack each comparison

Run the same TWFE estimates as in standard DiD

But include **interactions for the cohort-specific dataset with all of the fixed effects, controls, and clusters**.

BARRIOS (2021

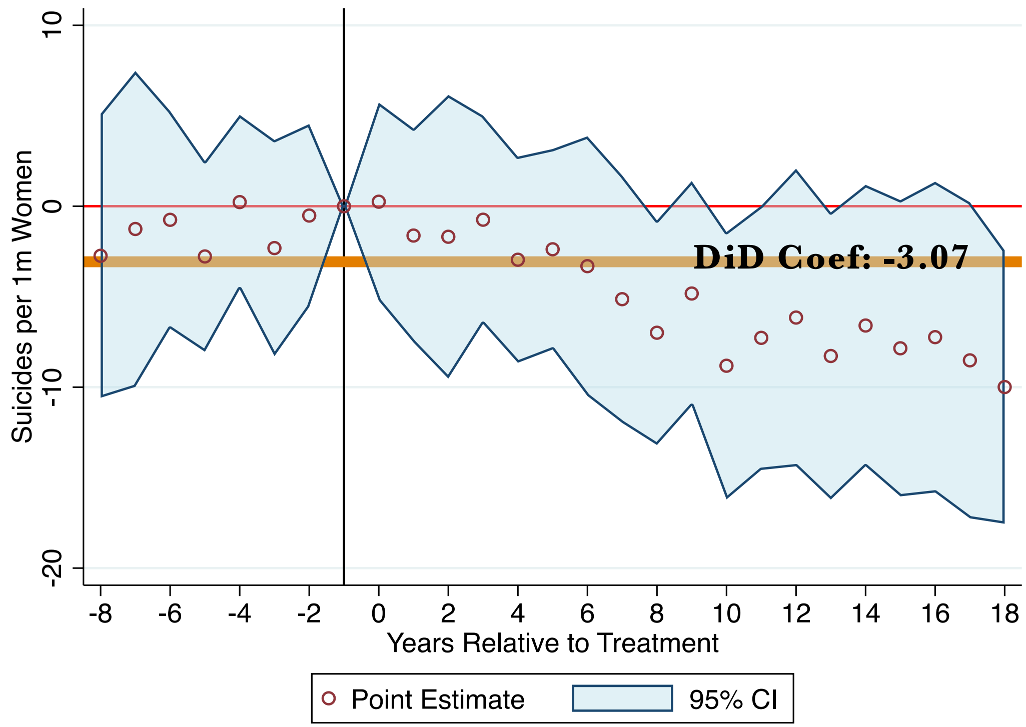
REAL WORLD EXAMPLES

1) THE EFFECT OF UNILATERAL DIVORCE ON SUICIDE

2) ACCOUNTING – 150 HR RULE IN THE PAPER

BARRIOS (2021)

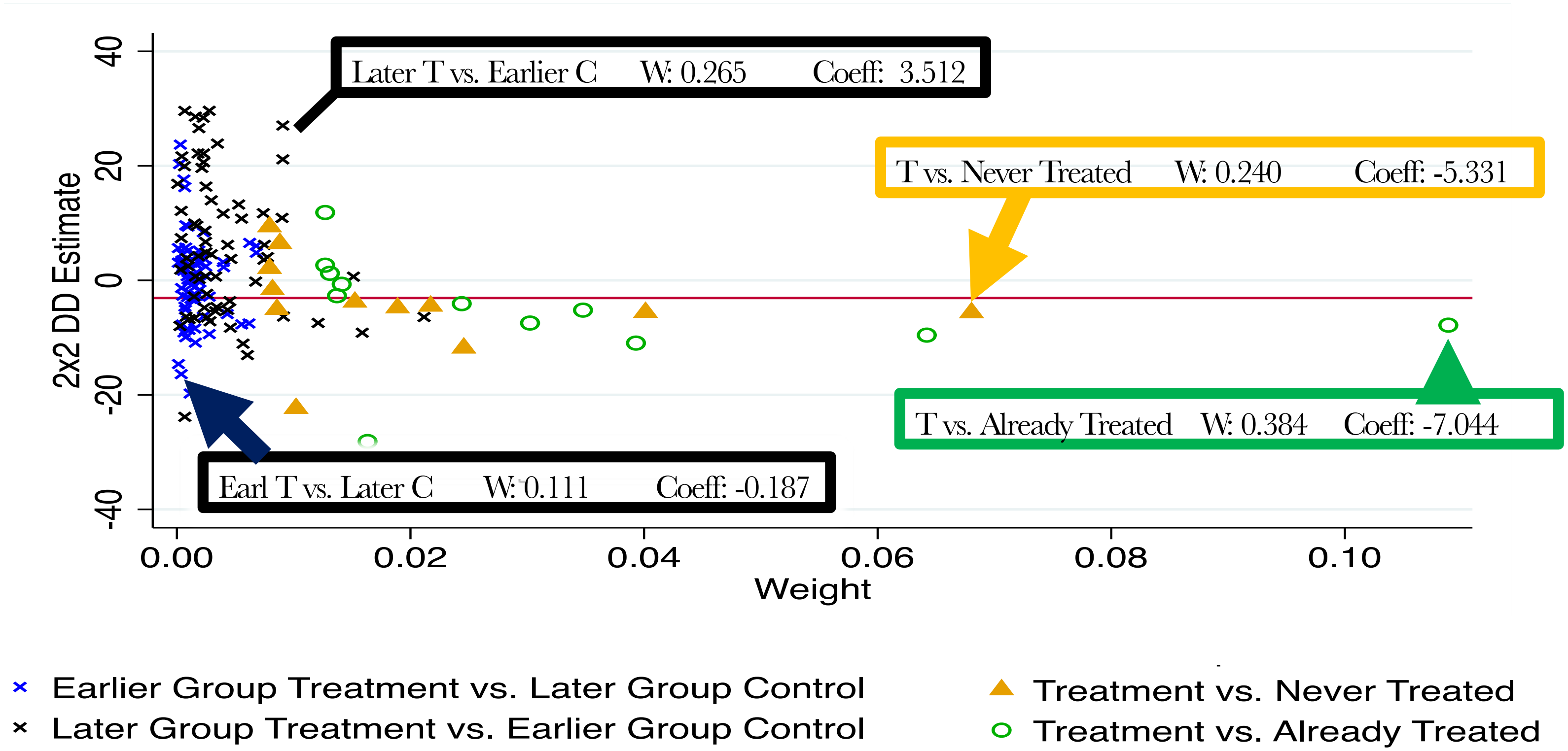
No-Fault Divorce Year	Number of States	Share of States	Treatment Share
Non-Reform States	5	0.10	.
Pre-64 Reform States	8	0.16	.
1969	2	0.04	0.85
1970	2	0.04	0.82
1971	7	0.14	0.79
1972	3	0.06	0.76
1973	10	0.20	0.73
1974	3	0.06	0.70
1975	2	0.04	0.67
1976	1	0.02	0.64
1977	3	0.06	0.61
1980	1	0.02	0.52
1984	1	0.02	0.39
1985	1	0.02	0.36



BARRIOS (2021

PANEL EVENT STUDY

GRAPHING THE DECOMPOSITION



BARRIOS (2021)

HELPFUL TIPS WHEN IMPLEMENTING STAGGERED DESIGNS

It is important to check whether common trends are satisfied in any panel data set being used for DiD analysis. ***Event–Study!!***

- Crucial to use event study design in panel tests
- You can usually ***trust a flat pre period!***

When treatment turns on at different times:

- Group the event study coefficients by cohort or fit a trend break instead of a constant DiD estimator
- Another way is to re-center and stack all the possible two-by-two DD comparisons with clean pre and post period for control units

Look at the weights and 2X2 DiD from the decomposition:

- Can describe in your test how many terms account for "most" of the weight.
- Check if the comparisons to "already treated" groups are very different than the other comparisons – can help give measure of bias
- You can take out the bias from time-varying effects by hand - Average the 2x2 DDs ***except*** *f*or those that use already treated units as controls weighting by the decomposition weights.
  - The estimate will not be biased by time-varying treatment effects (*although it will generally be skewed toward short-run effects)*

BARRIOS (2021)

HELPFUL TIPS WHEN IMPLEMENTING STAGGERED DESIGNS

STACKED DID DESIGN

How do you stack your DID or Event Study?

Checklist: Start with a staggered adoption design and some data.

1. Define an Event Window
2. Enumerate Sub-Experiments 3. Define Inclusion Criteria
4. Stack The Data
5. Specify an Estimating Equation

More info in my paper and code in my website: Johnmbarrios.com

BARRIOS (2021)

TAKE HOME MESSAGE

Staggered Adoption Designs create conditions for bias when treatment effects are time varying, and pose and aggregation/averaging puzzle.

Important Things

- Use an estimation strategy based only on "clean controls", such as stacked DID.
- Cost: trim treated units that don't have "clean controls" and time periods that fall outside the event window.
- Use event study approaches to allow effects to vary over time.
- In the stacked estimator, cluster standard errors at the Unit level to account for duplication.
- Consider further methods to control aggregation. See work by Callaway and Sant'Anna (2020)

TAKE HOME MESSAGE

The new approaches allow us to solve some issues but be careful not to fall in the habit of just running all alternatives estimators.

Most important is to think about the assumptions in the design and be as transparent and honest.

Causal inference is about only exploiting the good variation, i.e., those that respect our assumptions.

THANK YOU

Can find paper and STATA code on my website:

Johnmbarrios.com

HELPFUL STATA PACKAGES

Packages used in the presentation and paper to examine staggered DiD issues.

Also can make your life easier on your papers!

Bacondecomp - implements a decomposition of the difference-in-differences (DD) estimator with variation in treatment timing

Flexpaneldid - causal analysis of treatments with varying start dates and treatment durations within panel data

Eventdd - panel event study corresponding to a difference-in-difference style model. It generates plots of the leads and lags of the treatment