Part C: Panel Data Methods

C3: Staggered-Adoption Difference-in-Differences

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2024

C3 outline

1. Staggered adoption: Setting and estimands
2. Traditional estimators
3. What to do instead
4. Extensions

Staggered adoption/rollout setting

	$i = Z$	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$					
$t = 2$					
$t = 3$					
$t = 4$					
$t = 5$					
$t = 6$					

- Assume binary treatment; i gets treated at $t = E_i$ and stays treated forever: $D_{it} = \mathbf{1}[t \geq E_i]$
  - "Cohort" = units with the same E_i
- May or may not have never-treated units ($E_D = \infty$), always-treated units ($E_Z = 1$)
  - Come back to always-treated units later

Why staggered adoption?

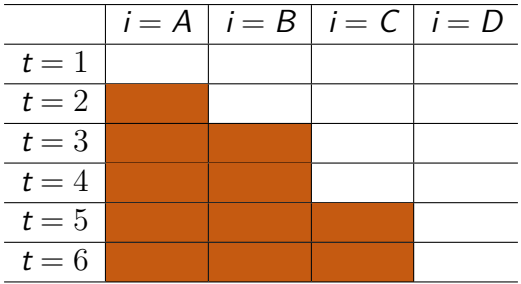
- A fact of life
  - Unilateral divorce laws adopted in different years across states
- Researcher's choice to have more comparable units
  - A panel of mothers, where E_i = year of birth of first child
  - May intentionally drop women without kids, as they are not expected to be on parallel trends

Notation and assumptions

- Fixed sample $\Omega = \{it\}$ with untreated obs Ω_0 and treated obs Ω_1
- Potential outcomes: $Y_{it}(e)$ if first treated in period e
  - No anticipation effects: $Y_{it}(e) = Y_{it}(e') \equiv Y_{it}(\infty)$ for $e, e' > t$
  - Could equivalently define $Y_{it}(\#$ of periods since treatment$)$
- Look for causal effects $Y_{it}(E_i) - Y_{it}(\infty)$ only
  - Simplify potential outcomes to $Y_{it}(d)$; single causal effect: $\tau_{it} = \mathbb{E}\left[Y_{it}(1) - Y_{it}(0)\right]$
  - Interpret heterogeneity of τ_{it} as reflecting, in part, dynamics
- Parallel trends: $\mathbb{E}\left[Y_{it}(0)\right] = \alpha_i + \beta_t$
- Linear target estimand: $\sum_{it \in \Omega_1} w_{it}\tau_{it}$ for w_{it} chosen by researcher...

Some estimands of interest

- ATT: $\frac{1}{|\Omega_1|} \sum_{it \in \Omega_1} \tau_{it}$ where $\Omega_1 = \{it : \ D_{it} = 1\}$



Some estimands of interest

- ATT

- ATT $h \geq 0$ periods since treatment (typically fewer units for longer horizons)

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$	$h = 0$			
$t = 3$	$h = 1$	$h = 0$		
$t = 4$	$h = 2$	$h = 1$		
$t = 5$	$h = 3$	$h = 2$	$h = 0$	
$t = 6$	$h = 4$	$h = 3$	$h = 1$	

Some estimands of interest

- ATT
- ATT $h \geq 0$ periods since treatment
- ATT $h \geq 0$ periods since treatment on a balanced set of units for different h

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$	$h = 0$			
$t = 3$	$h = 1$	$h = 0$		
$t = 4$	$h = 2$	$h = 1$		
$t = 5$	$h = 3$	$h = 2$	$h = 0$	
$t = 6$	$h = 4$	$h = 3$	$h = 1$	

Some estimands of interest

- ATT
- ATT $h \geq 0$ periods since treatment
- ATT $h \geq 0$ periods since treatment on a balanced set of units for different h
- Difference of ATT across subgroups
- Size-weighted ATT; etc.

Outline

1 Staggered adoption: Setting and estimands

2 Traditional estimators

3 What to do instead

4 Extensions

Conventional practice

By analogy with non-staggered DiD, [used to be] common to estimate:

- Static TWFE specification — to get a single summary statistic of treatment effects:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau_{\text{static}} D_{it} + \varepsilon_{it}$$

- Event study (dynamic) specification:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{\substack{h=-a \\ h \neq -1}}^{b-1} \tau_h \mathbf{1}\left[t = E_i + h\right] + \tau_{b+} \mathbf{1}\left[t \geq E_i + b\right] + \varepsilon_{it}$$

  - "Fully-dynamic" if $h = -1$ is the only omitted term
  - Some dummies are often binned or dropped on the left and/or on the right

Static TWFE specification

- Does the static TWFE specification estimate the ATT?

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau_{\text{static}} D_{it} + \varepsilon_{it}$$

- de Chaisemartin and D'Haultfœuille (AER 2020), Borusyak, Jaravel, Spiess (2024) (BJS):
  - Yes if the effects are homogeneous across units and periods. Not otherwise!
  - Under PTA, estimand $\tau_{\text{static}} = \sum_{it \in \Omega_1} w_{it}^{\text{static}} \tau_{it}$ for some weights w_{it}^{static} that add up to one
  - But $w_{it}^{\text{static}} \neq \frac{1}{|\Omega_1|}$ and some can be negative due to "forbidden comparisons"...

Forbidden comparisons

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \tau_{\text{static}} D_{it} + \varepsilon_{it}$$

$\mathbb{E}[Y_{it}]$	$i = A$	$i = B$
$t = 1$	α_A	α_B
$t = 2$	$\alpha_A + \beta_2 + \tau_{A2}$	$\alpha_B + \beta_2$
$t = 3$	$\alpha_A + \beta_3 + \tau_{A3}$	$\alpha_B + \beta_3 + \tau_{B3}$

Here $\hat{\tau}_{\text{static}} = \left(Y_{A2} - Y_{B2} \right) - \frac{1}{2} \left(Y_{A1} - Y_{B1} \right) - \frac{1}{2} \left(Y_{A3} - Y_{B3} \right)$

- Treated observations for early adopters are used as controls for treated observations of late adopters
- Long-term effects for early adopters can get a negative weight: $\tau_{\text{static}} = \tau_{A2} + \frac{1}{2}\tau_{B3} - \frac{1}{2}\tau_{A3}$

Mechanics of negative weights

- By Frisch-Waugh-Lovell, $\hat{\tau}_{\text{static}}$ can be obtained from

$$Y_{it} = \tau_{\text{static}} D_{it}^{\perp} + \text{error}$$

  where $D_{it}^{\perp}$ are residuals from regressing $D_{it} = a_i + b_t + \text{error}$.

- Thus, $\hat{\tau}_{\text{static}} = \frac{\sum_{it} D_{it}^{\perp} Y_{it}}{\sum_{js} \left(D_{js}^{\perp}\right)^2}$. Weights $\frac{D_{it}^{\perp}}{\sum_{js} \left(D_{js}^{\perp}\right)^2}$ are easy to compute

- But they can be negative for some treated observations: where $\hat{a}_i$ is high (early adopters) and $\hat{b}_t$ is high (late periods if few never-treated units)

- Angrist (1998) result does not apply!

Characterizing negative weights

- You can compute w_{it}^{static} (by observation or group totals) and total negative weights
- Goodman-Bacon (2021) provides a decomposition of τ_{static} as convex weighted average of several types of comparisons (package *bacondecomp*):
  - Treated vs. never treated *(good)*
  - Early adopters vs. late adopters *(good)*
  - Late adopters vs. early adopters *(forbidden)*
  - Treated during the sample vs. always-treated *(forbidden)*
- *Note*: only useful if you plan to use the static TWFE specification, and you don't

Under-identification of the fully-dynamic specification

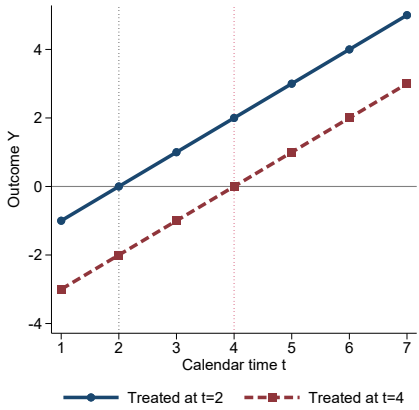
- "Fully-dynamic" specification:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{h \neq -1} \tau_h \mathbf{1}\left[t = E_i + h\right] + \varepsilon_{it}$$

- **Proposition** (BJS): Without never-treated units, the path $\{\tau_h\}_{h \neq -1}$ is not point identified in the fully-dynamic specification. Adding any linear trend to this path, $\{\tau_h + \kappa(h+1)\}$, fits the data equally well

  - This transformation corresponds to adding $\kappa\left(t - (E_i - 1)\right)$, which is collinear with unit and time FEs

Under-identification of the fully-dynamic specification



(from BJS)

- Diff-in-diff doesn't work without some assumption of no anticipation effects!

Spurious identification of very long-run effects

- "Semi-dynamic" specification can be estimated:

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{h \geq 0} \tau_h \mathbf{1}\left[t = E_i + h\right] + \varepsilon_{it}$$

- **Proposition** (BJS): Without never-treated units and with heterogeneous effects, long-run effects ($h \geq \max_i E_i - \min_i E_i$) are not identified by PTA, while the semi-dynamic specification produces some (spurious) estimates

	$i = A$	$i = B$	$i = C$
$t = 1$			
$t = 2$	✓		
$t = 3$	✓	✓	
$t = 4$	✓	✓	
$t = 5$	✗	✗	✗
$t = 6$	✗	✗	✗

Spurious identification of very long-run effects

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{h=0}^{1} \tau_h \mathbf{1}\left[t = E_i + h\right] + \varepsilon_{it}$$

$\mathbb{E}\left[Y_{it}\right]$	$i = A$	$i = B$
$t = 1$	α_A	α_B
$t = 2$	$\alpha_A + \beta_2 + \tau_0$	$\alpha_B + \beta_2$
$t = 3$	$\alpha_A + \beta_3 + \tau_1$	$\alpha_B + \beta_3 + \tau_0$

- Here $\hat{\tau}_1 = \left(Y_{A3} - Y_{B3} \right) + \left(Y_{A2} - Y_{B2} \right) - 2\left(Y_{A1} - Y_{B1}\right)$

Spurious identification of very long-run effects

$$Y_{it} = \tilde{\alpha}_i + \tilde{\beta}_t + \sum_{h=0}^{1} \tau_h \mathbf{1}\left[t = E_i + h\right] + \varepsilon_{it}$$

$\mathbb{E}\left[Y_{it}\right]$	$i = A$	$i = B$
$t = 1$	α_A	α_B
$t = 2$	$\alpha_A + \beta_2 + \tau_{A2}$	$\alpha_B + \beta_2$
$t = 3$	$\alpha_A + \beta_3 + \tau_{A3}$	$\alpha_B + \beta_3 + \tau_{B3}$

- Here $\hat{\tau}_1 = \left(Y_{A3} - Y_{B3}\right) + \left(Y_{A2} - Y_{B2}\right) - 2\left(Y_{A1} - Y_{B1}\right)$
- Estimand $\tau_1 = \tau_{A3} + \tau_{A2} - \tau_{B3}$ inevitably involves extrapolation that is invalid with heterogeneous effects

Cross-horizon contamination

Sun and Abraham (2021):

- Similar problems occur even for short-run effects in dynamic specification
  - Estimand τ_h is not an average of horizon-h effects: contaminated by heterogeneity of effects at other horizons
- And pre-trend coefficients are contaminated by treatment effect heterogeneity
  - Can be significant even if PTA holds!
- *Note*: these problems tend to be small in practice

Pre-testing problems

Roth (AER:I 2022):

- Estimators of causal effects and pre-trends are correlated, e.g. by using the same reference period
  - In the non-staggered case, $(\bar{Y}_{\text{treated},E+h} - \bar{Y}_{\text{control},E+h}) - (\bar{Y}_{\text{treated},E-1} - \bar{Y}_{\text{control},E-1})$ and $(\bar{Y}_{\text{treated},E-\ell} - \bar{Y}_{\text{control},E-\ell}) - (\bar{Y}_{\text{treated},E-1} - \bar{Y}_{\text{control},E-1})$
- Suppose you only report the results if some pre-trend test doesn't reject ("pre-testing")
- If PTA holds, you are distorting inference for causal effects
- If PTA doesn't hold, you are changing the bias; under some conditions increase it

Outline

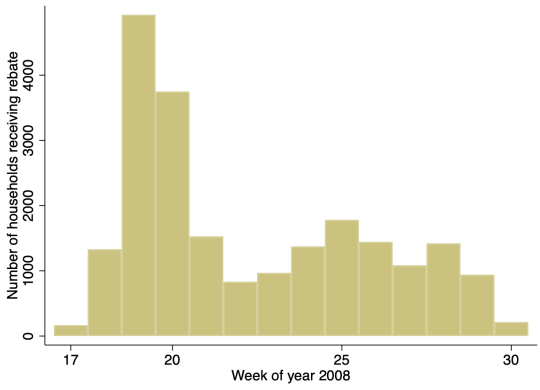
1 Staggered adoption: Setting and estimands

2 Traditional estimators

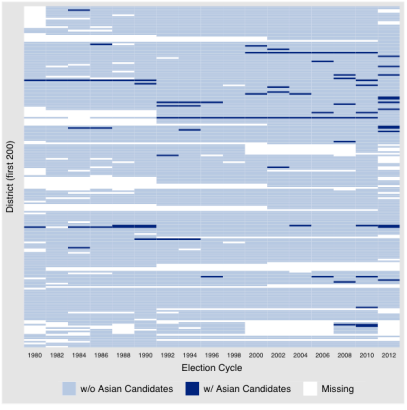
3 What to do instead

4 Extensions

Plot treatment timing

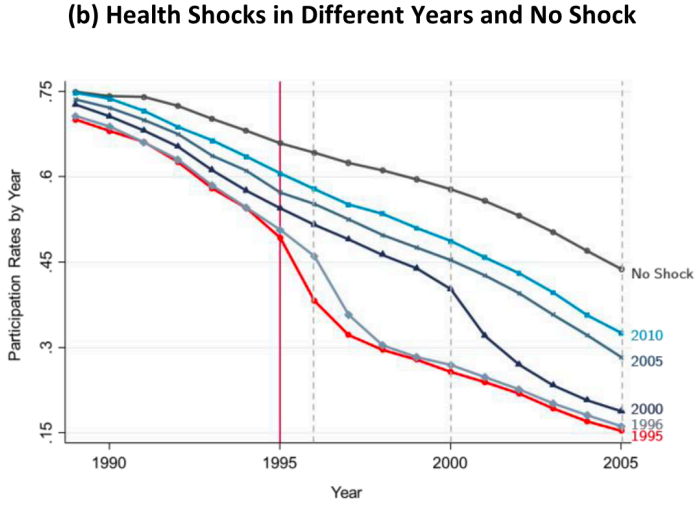


(From BJS)



(from Chiu, Lan, Liu, Xu 2023;
package *panelView* in Stata & R)

Plot raw outcome data by cohort



(from Fadlon and Nielsen, 2015 version)

Estimation robust to heterogeneous effects

- The problems arise from conventional specifications being too restrictive
- They are not fundamental to staggered adoption DiD
  - Under PTA, there are many valid 2x2 contrasts
- How to combine them?
  - Manual averaging approaches
    - ⋆ Yield simpler estimators, more closely parallel conventional event studies
  - Imputation approaches
    - ⋆ Transparently link assumptions to estimators, more versatile, often more efficient
  - Regression implementations of both approaches

Manual averaging estimators

de Chaisemartin and D'Haultfœuille (AER 2020) for $h = 0$:

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$	$h = 0$			
$t = 3$		$h = 0$		
$t = 4$				
$t = 5$			$h = 0$	
$t = 6$				

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$	$h = 0$			
$t = 3$		$h = 0$		
$t = 4$				
$t = 5$			$h = 0$	
$t = 6$				

- For each cohort $E_i = e$: form the clean control group; compute **cohort-average treatment effect** ($CATT_{e,e+0}$) by comparing $Y_{ie} - Y_{i,e-1}$
- Average across cohorts weighting by cohort size
- *Note:* PTA is not fully exploited by comparing to $e - 1$ only, without earlier periods

Manual averaging estimators (2)

de Chaisemartin and D'Haultfœuille (2023 WP) for $h \geq 0$:

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$				
$t = 3$	$h = 1$			
$t = 4$		$h = 1$		
$t = 5$				
$t = 6$			$h = 1$	

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$				
$t = 3$	$h = 1$			
$t = 4$		$h = 1$		
$t = 5$				
$t = 6$			$h = 1$	

- For each cohort $E_i = e$: form the control group as cohorts not treated by $e + h$; compute $CATT_{e,e+h}$ by comparing $Y_{i,e+h} - Y_{i,e-1}$
- Get SE by bootstrap or analytical formula
- Sun and Abraham (2021): same but use never-treated controls only
  - If no never-treated, use latest-treated cohort instead

Manual averaging: Pre-trend tests

de Chaisemartin and D'Haultfœuille (2023 WP) pre-trend tests:

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$			$\ell = 3$	
$t = 3$				
$t = 4$				
$t = 5$				
$t = 6$				

	$i = A$	$i = B$	$i = C$	$i = D$
$t = 1$				
$t = 2$			$\ell = 3$	
$t = 3$				
$t = 4$				
$t = 5$				
$t = 6$				

- For cohort e and lead $\ell > 1$, measure $Y_{i,e-\ell} - Y_{i,e-1}$ **or** $Y_{i,e-\ell} - Y_{i,e-(\ell-1)}$
  - Do not confuse the interpretations (Roth 2024)
- Compare to the control group of units not treated by e
- Average across cohorts; test it's zero

Imputation estimators

- PTA requires $\mathbb{E}[Y_{it}(0)] = \alpha_i + \beta_t$
- No anticipation effects means $Y_{it} = Y_{it}(0)$ for untreated observations ($it \in \Omega_0$)
- Imputation approach:
  1. Estimate α_i and β_t from untreated observations
     - ⋆ Need pre-treatment obs for every unit to get $\hat{\alpha}_i$
     - ⋆ Need untreated obs in every period to get $\hat{\beta}_t$
  2. For each treated observation $it \in \Omega_1$, compute $\hat{\tau}_{it} = Y_{it} - \hat{\alpha}_i - \hat{\beta}_t$
     - ⋆ Each $\hat{\tau}_{it}$ is very noisy!
  3. Estimate $\tau_w = \sum_{it \in \Omega_1} w_{it} \tau_{it}$ by $\hat{\tau}_w = \sum_{it \in \Omega_1} w_{it} \hat{\tau}_{it}$
     - ⋆ Averaging across many units makes $\hat{\tau}_w$ consistent

Efficient imputation

How to estimate α_i and β_t?

- $\hat{\tau}_w$ is unbiased for any unbiased $\hat{\alpha}_i, \hat{\beta}_t$

Proposition (BJS) If $Y_{it}(0) = \alpha_i + \beta_t + \varepsilon_{it}$ for *spherical* ε_{it} (i.e., homoskedastic and serially uncorrelated), estimating $\hat{\alpha}_i, \hat{\beta}_t$ by OLS in the untreated sample yields most efficient $\hat{\tau}_w$ for any τ_w

- This imputation estimator can be obtained by OLS from a very flexible regression

$$Y_{it} = \alpha_i + \beta_t + \tau_{it} D_{it} + \varepsilon_{it}$$

  where each treated observation gets its own coef

- By Gauss-Markov, OLS is efficient for the vector of τ_{it} and for any linear combination τ_w

Comparison to manual averaging

- **Proposition** (BJS): Any unbiased estimator for τ_w under arbitrary heterogeneity of treatment effects can be represented as an imputation estimator for some unbiased $\hat{\alpha}_i, \hat{\beta}_t$

- de Chaisemartin and D'Haultfœuille (2023) and Sun and Abraham (2021) are also imputation estimators that use less information to estimate $\hat{\alpha}_i, \hat{\beta}_t$

- Exception (Harmon 2022): If ε_{it} is a random walk, de Chaisemartin and D'Haultfœuille's estimator is efficient for $h = 0$

  - Outcomes at $E_i - 1$ contain all useful information; previous periods only add noise

BJS: Asymptotic standard errors

- Represent $\hat{\tau}_w = \sum_{it} v_{it} Y_{it}$: $v_{it} = w_{it}$ for $it \in \Omega_1$; v_{it} can be computed for Ω_0
- True variance: $\mathrm{Var}\left[\hat{\tau}_w\right] = \mathrm{Var}\left[\sum_{it} v_{it}\varepsilon_{it}\right] = \mathbb{E}\left[\sum_i \left(\sum_t v_{it}\varepsilon_{it}\right)^2\right]$
- Plug-in estimator: $\hat{\sigma}_w^2 = \sum_i \left(\sum_t v_{it}\hat{\varepsilon}_{it}\right)^2$ where $\hat{\varepsilon}_{it} = Y_{it} - \hat{\alpha}_i - \hat{\beta}_t$ for untreated obs.
- Key challenge: $Y_{it} - \hat{\alpha}_i - \hat{\beta}_t - \hat{\tau}_{it} = 0$ for treated obs. by construction
  - Impossible to separate variation in τ_{it} from ε_{it}:
    $\hat{\tau}_{it} = \tau_{it} + \varepsilon_{it} +$ noise from estimating FEs
- Obtain conservative SE by attributing some τ_{it} variation to ε_{it}

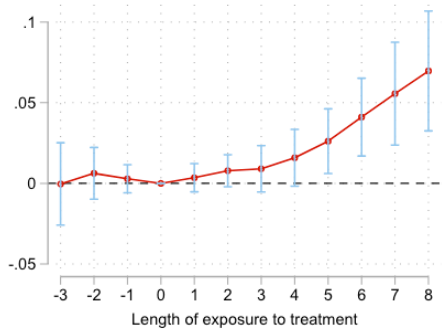
BJS: Asymptotic standard errors (2)

- E.g. $\tilde{\varepsilon}_{it} = \hat{\tau}_{it} - \hat{\tau}_{E_i t}$ where $\hat{\tau}_{et} =$ Avg of $\hat{\tau}_{jt}$ in the cohort $E_j = e$
- If cohorts are large, $\tilde{\varepsilon}_{it} \approx \varepsilon_{it} + (\tau_{it} - \bar{\tau}_{E_i t})$
  - SE are conservative when there is variation in τ_{it} within cohorts; otherwise asymptotically exact
- If cohorts are small, can replace $\hat{\tau}_{E_i t}$ with averages that pool multiple cohorts or use leave-out estimation; see BJS
  - SE of other estimators assume random sample of units $\Longrightarrow$ require large cohorts

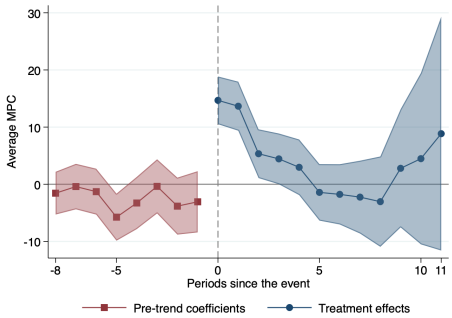
BJS: Pre-trend testing

- To test for pre-trends, always use only untreated observations
- Null hypothesis: $Y_{it} = \alpha_i + \beta_t + \varepsilon_{it}$
- Choose a richer alternative model: $Y_{it} = \alpha_i + \beta_t + \eta' W_{it} + \varepsilon_{it}$
  - E.g. anticipation effects: W_{it} are $\mathbf{1}[t = E_i - 1], \ldots, \mathbf{1}[t = E_i - L]$
  - Non-parallel linear trends: W_{it} are cohort dummies $\times t$
  - Structural break: W_{it} are cohort dummies $\times$ post financial crisis
- Use F-test for $\eta = 0$. (Don't include too many covariates to avoid low power)
- *Note:* not all violations affect causal estimates much
  - But Rambachan-Roth approach is not available for imputation yet
- *Bonus:* pre-trend coefs are not correlated w/ causal effects under spherical errors

dCDH and BJS event study graphs



(from de Chaisemartin, D'Haultfœuille, 2023 WP)

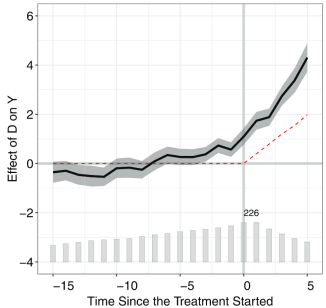


(BJS data, using `event_plot` Stata command)

- Note different reference groups and different behavior of SE

Liu, Wang, Xu (AJPS 2022) pre-trend test

- Liu et al. also derive an imputation approach: "FE counterfactual estimator"
- Both for $h \geq 0$ and $h < 0$, plot averages of $Y_{it} - \hat{\alpha}_i - \hat{\beta}_t$



- Like with dCDH graphs, same interpretation on the left and right
- But no single reference period because imputation-based
- Unlike BJS, no explicit alternative hypothesis
- See Gardner (2021) and Thakral, To (2021) for a similar approach

Flexible regression estimator

- Can the convenience of OLS be preserved without negative weights and other problems?
- Wooldridge (2021): the problem with old-school estimators is restrictive specifications $\Longrightarrow$ let's keep running regressions but more flexibly

$$Y_{it} = \alpha_i + \beta_t + \sum_e \sum_{s \geq e} \tau_{es} \mathbf{1}\left[E_i = e\right] \times \mathbf{1}\left[t = s\right] + \text{error}$$

- τ_{et} estimates CATT for cohort e in period t; then aggregate estimates as required
- In complete panels, equivalent to BJS for CATT-based estimands
- *Note:* Sun and Abraham's estimator also has a regression representation

Local projection DiD

- Dube, Girardi, Jorda, Taylor (2023) propose to estimate, for each h,

$$Y_{i,t+h} - Y_{i,t-1} = \beta_{ht} + \tau_h \mathbf{1}[t = E_i] + \text{error}$$

  on the subsample where $E_i = t$ or $E_i > t + h$

  - Combines the local projections estimator for time series of Jorda (2005) with "stacking" approach of Cengiz, Dube, Lindner, Zipperer (2019, Appendix D)

- τ_h estimates a convex weighted average of treatment effects

  - Can be reweighted to get dCDH
  - Instead of subtracting $Y_{i,t-1}$ can subtract average of several pre-period outcomes $\implies$ closer to imputation

- Can stack the data for different h to get joint confidence intervals

Does heterogeneity-robust estimation matter?

BJS application:

- Yes, relative to specifications that restrict dynamics
- Yes, for the spurious long-run coefficients
- For the short-run, the semi-dynamic specification seems fine

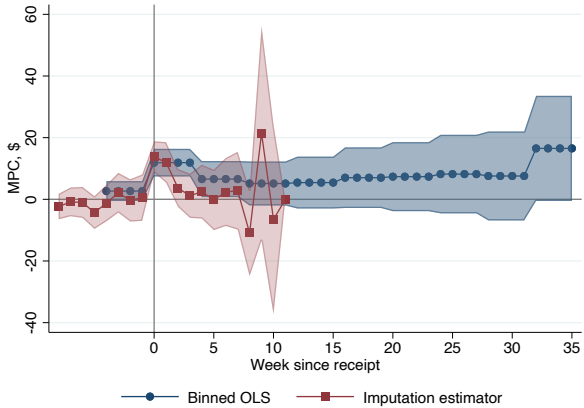
Setting: marginal propensity to spend (MPC) from the 2008 Economic Stimulus Payments (tax rebates)

- Staggered disbursement of rebates. Weekly spending data from Nielsen

Broda and Parker (2014) find a very large MPC from a monthly-binned specification:

$$Y_{it} = \alpha_i + \beta_t + \sum_{m=-1}^{\infty} \tau_m \mathbf{1}\left[t - E_i \in \{4m-3, \ldots, 4m\}\right] + \text{error}_{it}$$

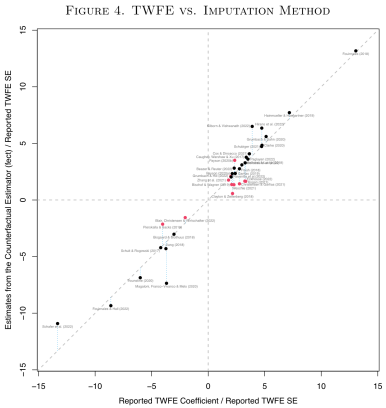
BJS results



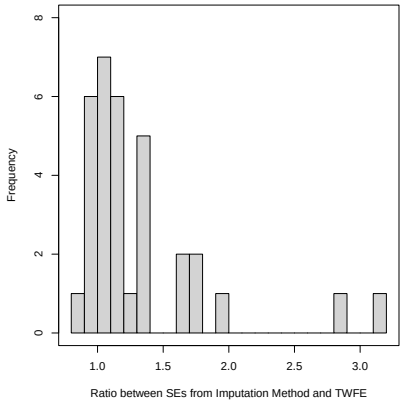
- Because the effects are fast decaying, binned specification overstates them in the first month
- Binned specification also extrapolates them to all future months

Chiu, Lan, Liu, Xu (2023) are less convinced

Chiu et al. reanalyze main significant TWFE coefficients in 37 political science papers and don't find big differences — but also no cost of being robust



(From Chiu et al.)



(Computation by Yixing Xu)

Stata/R implementations

Method	Stata	R
de Chaisemartin, D'Haultfoeuille 2020, 2023	`did_multiplegt_dyn`	
Sun and Abraham 2021	`eventstudyinteract`	—
Callaway and Sant'Anna 2021	`csdid`	`did`
Borusyak et al. 2022	`did_imputation`	`didimputation` *(limited)*
Plotting any estimates	`event_plot`	—

Outline

1. Staggered adoption: Setting and estimands
2. Traditional estimators
3. What to do instead
4. Extensions

Extensions

Feature	Baseline	Extensions
Reason for PTA:	None	Linking to selection models
Model of $Y(0)$:	TWFE (PTA)	Covariates Multiplicative and duration models
Model of $Y(1)$:	Arbitrary heterogeneity	*Ex ante* restrictions
Treatment:	Binary, absorbing	Continuous treatments Treatment reversals
Data structure:	Panel	Two-dimensional cross-sections Triple-differences

Link to selection models

New literature tries to reconcile PTA with self-selection into treatment

- de Chaisemartin, D'Haultfœuille (2022), Ghanem, Sant'Anna, Wuthrich (2023), Marx, Tamer, Tang (2023)
- Themes:
  - Is self-selection based on time-invariant or time-varying unobservables?
  - Is self-selection based on gains from treatment or on untreated potential outcomes ("Ashenfelter's dip")
  - Are agents forward-looking?
  - Can they learn about future potential outcomes from observing earlier outcomes?
  - Is $Y_{it}(0)$ a random walk, up to period FEs?

Covariates

Generalize the two approaches from canonical DiD:

- Imputation approach extends directly:
  - Estimate $Y_{it}(0) = \alpha_i + \beta_t + \gamma' X_{it}$ from untreated observations in the first step
  - Use $\hat{\tau}_{it} = Y_{it} - \hat{Y}_{it}(0)$ for treated observations

Covariates (2)

Generalize the two approaches from canonical DiD:

- Callaway and Sant'Anna (2021) apply the doubly-robust approach to canonical DiD (Sant'Anna and Zhao 2020) to dCDH
  - Impose $\mathbb{E}\left[Y_{it}(0) - Y_{i,t-1}(0) \mid X_i, E_i = e\right] = \mathbb{E}\left[Y_{it}(0) - Y_{i,t-1}(0) \mid X_i, E_i > s\right]$ for all $s \geq t \geq e$
  - For $t \geq e$ define $S_{ite} = \mathbf{1}\left[E_i = e \text{ or } E_i > t\right]$ and

$$p_{e,t}(X_i) = Pr(E_i = e \mid X_i; S_{ite} = 1); \quad m_{e,t}(X_i) = \mathbb{E}\left[Y_{it} - Y_{i,e-1} \mid X_i; E_i > t\right]$$

  - Estimate $p_{e,t}(X_i)$ and $m_{e,t}(X_i)$ to get $CATT_{et}$ by AIPW (or just regression adjustment or IPW), then average across cohorts
  - *Optional:* use Sun and Abraham (2021) instead of dCDCH
  - *Note:* good software packages even for the case without covariates

Restrictions on treatment effects

So far we have considered estimators robust to arbitrary effect heterogeneity

- This is a strong demand on the estimator $\implies$ can't get all estimands
- Model is very asymmetric: strong assumptions on $Y_{it}(0)$ and none on $Y_{it}(1)$
- With so much heterogeneity possible, are ATTs informative about future policy?
  "Anyone who makes a living out of data analysis probably believes that heterogeneity is limited enough that the well-understood past can be informative about the future" (Angrist and Pischke 2010)

Restrictions on treatment effects

Can we get identify more estimands (e.g. long-run effects) or get more power for the same estimands using extra restrictions on τ_{it} and $Y_{it}(1)$?

- Impose some simple model of treatment effects, $\tau_{it} = \Gamma_{it}'\theta$: e.g.
  - $\tau_{it} = \bar{\tau}$ is homogeneous across i and t $\implies$ *static TWFE specification*
  - $\tau_{it} \equiv \tau_{t-E_i}$ is homogeneous across i for any given horizon $\implies$ *semi-dynamic*
  - $\tau_{it} = 0$ when $t > E_i + K$ for some K
  - $\tau_{it+1} = \tau_{it}$ when $t > E_i + K$
- **Proposition** (BJS): Under PTA, spherical errors, and model of τ_{it}, the efficient unbiased estimator of τ_w is:
  1. Run OLS on the full sample for the "true" model: $Y_{it} = \alpha_i + \beta_t + D_{it}\Gamma_{it}'\theta + \varepsilon_{it}$
  2. Compute $\hat{\tau}_{it} = \Gamma_{it}'\hat{\theta}$ and $\hat{\tau}_w = \sum_{it \in \Omega_1} w_{it}\hat{\tau}_{it}$

Continuous treatment intensity

Minor changes to the setting if we observe untreated periods for every unit and untreated units for every period

- In BJS, the tax rebate amount is a continuous treatment but that's no problem

- Can identify e.g. the average MPC as % of household-specific rebate:

$$\tau_w = \sum_{t-E_i=h} \frac{\tau_{it}}{\text{Rebate}_i} \quad \text{or} \quad \tau_w = \frac{\sum_{t-E_i=h} \tau_{it}}{\sum_{t-E_i=h} \text{Rebate}_i}$$

Continuous treatment intensity (2)

Otherwise PTA for $Y_{it}(0)$ is not helpful, have to restrict effect heterogeneity in some way

- Explicit model of heterogeneity $\Longrightarrow$ imputation with restrictions
- Explicit contrasts (justified by stronger PTA): de Chaisemartin, D'Haultfœuille, Pasquier, Vazquez-Bare (2023)
  - Consider D_{it} continuously distributed in every period (e.g. tariffs)
  - Decide to compare stayers and switchers with the same initial treatment $\Longrightarrow$ impose $\mathbb{E}\left[\Delta Y_{i2}(d) \mid D_{i1} = d, D_{i2}\right] = \mathbb{E}\left[\Delta Y_{i2}(d) \mid D_{i1} = d\right]$
  - Can't match on D_{i1} exactly $\Longrightarrow$ estimate $\mathbb{E}\left[\Delta Y_{i2}(d) \mid D_{i1} = D_{i2} = d\right]$ nonparametrically for stayers and impute the counterfactual for switchers

Treatment reversals

What if treatment is not an absorbing state?

- E.g. Adda (2016): Economic effects of epidemics (e.g., flu); D_{it} = dummy for school holidays
- Key assumption: **no carryover effects**
  - If lagged treatment can have effects far into the future, there is no control group
- Imputation approach extends immediately
  - Estimate α_i, β_t from untreated observations both before and after treatment
- dCDH (2020) assume PTA on $Y_{it}(1)$ and use groups with $D_{i,t-1} = D_{it} = 1$ as controls for "leavers" who switch from $D_{i,t-1} = 1$, $D_{it} = 0$
  - Parameter is no longer ATT
- The no-carryover effects assumption should be tested

Two-dimensional cross-sections

- Suppose we have data by region i and age group g in a single period
- $D_{ig} = \mathbf{1}[g < E_i]$, e.g. a policy applies to 18y.o. but was rolled out in a staggered way
  - Impose $Y_{ig}(0) = \alpha_i + \beta_g$; use older groups as controls
- dCDH and imputation extend by redefining variables
  - E.g. $D_{ib} = \mathbf{1}[b \geq B_i]$ where b is birth year and B_i is the earliest birth year eligible

Triple-differences

- Manual averaging can be manually extended

- Imputation extends immediately: estimate

$$Y_{igt}(0) = \alpha_{it} + \beta_{gt} + \gamma_{ig} + \text{error}$$

  on all untreated observations

Suggested DiD checklist: *Ex ante*

1. Carefully define treated vs. untreated observations
   - If you study a national shock, is there a control group at all?
   - Think spillovers: if some units could be affected indirectly, can view them as a separate treatment arm
   - Think anticipation effects: if the shock was announced before implementation, can view post-announcement periods as a separate treatment arm
   - Think carryover effects: are some post-periods untreated again?
2. Justify *ex ante* whether (conditional) parallel trends may be expected to hold
   - Which other important shocks happen in your period, observed or unobserved?
   - Which pre- and post-periods may violate PTA?
   - Which control units may violate PTA? E.g., keep or drop never-treated units?
   - Could some covariates help?
3. Introduce potential outcomes, identifying assumptions, and estimand
   - Don't go straight to an empirical specification

Suggested DiD checklist: *Ex post*

4. Visualize treatment timing. Plot raw outcomes by cohort
5. Test the assumptions
   - Make sure your tests use untreated data only, explicitly or implicitly
   - Avoid uninformative tests: pre-trends for robot adoption when robots barely exist
   - Don't oversell low-powered tests. Check robustness to PTA violations instead
   - Consider tests beyond standard pre-trend tests: Balance in levels? Heterogeneous trends across cohorts? Outcomes other than your main one?
6. Estimate the effects
   - Avoid restrictive "convenience" regressions. If you want an average, estimate flexibly then average
   - Use heterogeneity-robust methods except when (1) the estimand of interest is not identified otherwise and (2) you impose restrictions on heterogeneity consciously