Canonical Research Designs I: Difference-in-Differences III:

Paul Goldsmith-Pinkham

March 6, 2025

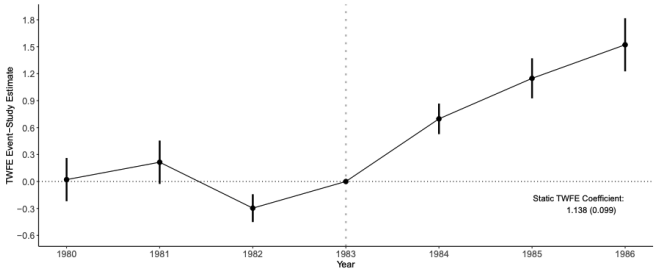
Today's Topics

- Complications that arise in DiD:
  1. continuous treatments
  2. multiple treatments
  3. random timing
  4. Covariates + time-varying covariates

- Checklist of what you need to consider

Fix ideas with example (Callaway, Goodman-Bacon, Sant'anna (2024))

Figure 1: Two-Way Fixed Effects Event-Study Estimates of the Effect of Medicare's Reimbursement Reform on Hospital Input Mix



Notes: The figure plots TWFE event-study coefficients and their 95% confidence intervals from regressions with hospital fixed effects, year fixed effects, and the 1983 Medicare inpatient share (m_i) interacted with either a dummy for years after 1983 or the year dummies. The outcome variable is the depreciation share of total operating expenses, a measure of hospitals' capital/labor ratio. The data cover the years 1980-1986 and come from the American Hospital Association's annual survey (American Hospital Association, 1986). We dropped 860 hospitals (out of 6741) that have missing data for the outcome. We also report the static TWFE coefficient and standard errors associated with (1.1). All standard errors are clustered at the hospital level.

$$y_{it} = \theta_t + \eta_i + \beta D_i \cdot Post_t + \varepsilon_{it}$$

In 1983, Prospective Payment System changed payment for Medicare patients to hospitals

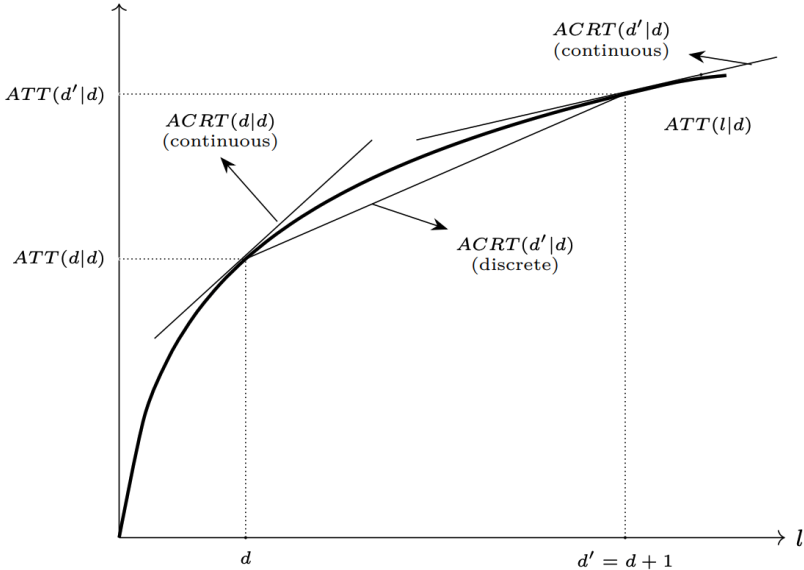
Continuous treatment

- Recall that estimating treatment effects with random assignment and continuous treatments was data hungry, but doable
- Without random assignment, and unconstrained heterogeneity, the challenge is more complicated
- First, consider what the estimand of interest is. Note that with potential outcomes, we now have $Y_i(D_i)$ with D_i multivalued.
- So what is the contrast of interest?

$$ATT(d|d) = E(Y_i(d) - Y_i(0)|D_i = d) \tag{1}$$

$$ACRT(d|d) = \frac{\partial E(Y_i(l)|D_i = d)}{\partial l}\bigg|_{l=d} \tag{2}$$

ACRT vs. ATT


$ACRT(d'|d)$
(continuous)
$ATT(d'|d)$
$ACRT(d|d)$
(continuous)
$ATT(l|d)$
$ACRT(d'|d)$
(discrete)
$ATT(d|d)$
l
d
$d' = d + 1$

Continuous treatment – Estimand

- How do these differ? Consider their interpretations:
  - ATT: effect relative to baseline of zero effect
  - ACRT: effect of marginally increasing treatment, evaluted at a given point (can integrate too!)

- Under parallel trends, what can we identify?
  - First, parallel in what counterfactual? Under *no* treatment

$$E(Y_2(0) - Y_1(0)|D = d) = E(Y_2(0) - Y_1(0)|D = 0). \tag{3}$$

  implies

$$ATT(d|d) = E(\Delta Y|D = d) - E(\Delta Y|D = 0). \tag{4}$$

- However, this does not identify anything about the ACRT

Continuous treatment – Estimand

- Can we move between ATT and ACRT?

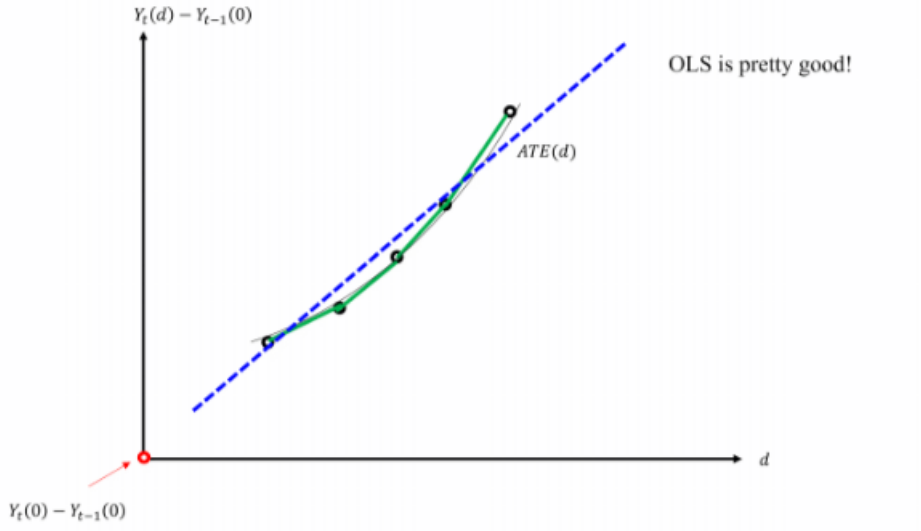
- Under standard parallel trends,

$$\frac{\partial E(\Delta Y|D=d)}{\partial d} = \frac{\partial ATT(d|d)}{\partial d} = ACRT(d|d) + \frac{\partial ATT(d|l)}{\partial l}|_{l=d} \tag{5}$$

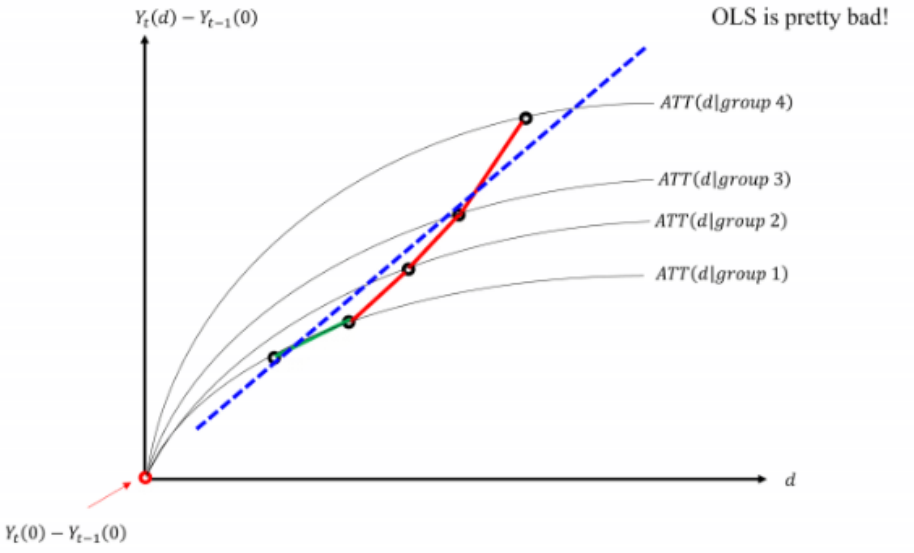
- Consider comparing two ATTs:

$$\begin{aligned} ATT(d|d) - ATT(d'|d') &= (E[\Delta Y|D=d] - E[\Delta Y|D=0]) \\ &\quad - (E[\Delta Y|D=d'] - E[\Delta Y|D=0]) \\ &= E[\Delta Y|D=d] - E[\Delta Y|D=d'] \\ &= E[\Delta Y(d) - \Delta Y(d')|D=d] + ATT(d'|d) - ATT(d'|d') \end{aligned}$$

Continuous treatment – challenge



Continuous treatment – challenge



Continuous treatment – stronger assumption about heterogeneity

- "Strong" parallel trends assumption: $E(\Delta Y_i(d)) = E(\Delta Y_i(d)|D_i = d)$
- No selection bias "on average"
- Effectively assuming some kind of homogeneity in the treatment across groups
- Under this set of assumptions, you can identify the ATE as well as the ATT (and the ACR along with the ACRT)

Continuous treatment – what if there is no untreated group?

- Can consider parallel trends relative to some d_l "baseline" group
- However, this mixes the selection effect based on treatment:

$$E(\Delta Y|D_i = d) - E(\Delta Y|D_i = d_l) = ATT(d|d) - ATT(d_l|d_l) \tag{6}$$

and would need that $ATT(d_l|d) = ATT(d_l|d_l)$ to identify the ATT

So what does TWFE do?

Table 1: TWFE Decomposition Weights

Decomposition	$D > 0$ Weights	$D = 0$ Weights
Causal response	$w_1^{\text{acr}}(l) = \frac{(\mathbb{E}[D\|D \geq l] - \mathbb{E}[D])\mathbb{P}(D \geq l)}{\text{Var}(D)}$	$w_0^{\text{acr}} = \frac{(\mathbb{E}[D\|D > 0] - \mathbb{E}[D])\mathbb{P}(D > 0)d_L}{\text{Var}(D)}$
Levels	$w_1^{\text{lev}}(l) = \frac{(l - \mathbb{E}[D])}{\text{Var}(D)} f_D(l)$	$w_0^{lev} = -\frac{\mathbb{E}[D]\mathbb{P}(D = 0)}{\text{Var}(D)}$
Scaled levels	$w^{\text{s}}(l) = l\frac{(l - \mathbb{E}[D])}{\text{Var}(D)} f_D(l)$	
Scaled 2×2	$w_1^{2\times 2}(l, h) = \frac{(h - l)^2 f_D(h) f_D(l)}{\text{Var}(D)}$	$w_0^{2\times 2}(h) = \frac{h^2 f_D(h)\mathbb{P}(D = 0)}{\text{Var}(D)}$

Notes: The table provides the formulas for the weights used in the decompositions of β^{twfe} provided in this section.

Theorem 3.4. *Under Assumptions 1, 2(a), 3, and 4, β^{twfe} can be decomposed in the following ways:*

(a) Causal Response Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1^{acr}(l) \left(ACRT(l|l) + \underbrace{\frac{\partial ATT(l|h)}{\partial h}\Big|_{h=l}}_{\text{selection bias}} \right) dl + w_0^{acr} \frac{ATT(d_L|d_L)}{d_L}$$

where the weights are always positive and integrate to 1.

So what does TWFE do?

Table 1: TWFE Decomposition Weights

Decomposition	$D > 0$ Weights	$D = 0$ Weights
Causal response	$w_1^{\text{acr}}(l) = \frac{(\mathbb{E}[D\|D \geq l] - \mathbb{E}[D])\mathbb{P}(D \geq l)}{\text{Var}(D)}$	$w_0^{\text{acr}} = \frac{(\mathbb{E}[D\|D > 0] - \mathbb{E}[D])\mathbb{P}(D > 0)d_L}{\text{Var}(D)}$
Levels	$w_1^{\text{lev}}(l) = \frac{(l - \mathbb{E}[D])}{\text{Var}(D)} f_D(l)$	$w_0^{lev} = -\frac{\mathbb{E}[D]\mathbb{P}(D = 0)}{\text{Var}(D)}$
Scaled levels	$w^{\text{s}}(l) = l\frac{(l - \mathbb{E}[D])}{\text{Var}(D)} f_D(l)$	
Scaled 2×2	$w_1^{2\times 2}(l, h) = \frac{(h - l)^2 f_D(h) f_D(l)}{\text{Var}(D)}$	$w_0^{2\times 2}(h) = \frac{h^2 f_D(h)\mathbb{P}(D = 0)}{\text{Var}(D)}$

Notes: The table provides the formulas for the weights used in the decompositions of β^{twfe} provided in this section.

(b) Levels Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1^{lev}(l) ATT(l|l)\, dl,$$

where $w_1^{lev}(l) \lesseqgtr 0$ *for* $l \lesseqgtr \mathbb{E}[D]$, *and* $\int_{d_L}^{d_U} w_1^{lev}(l)\, dl + w_0^{lev} = 0$.

(c) Scaled Levels Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w^{s}(l) \frac{ATT(l|l)}{l}\, dl,$$

where $w^{s}(l) \lesseqgtr 0$ *for* $l \lesseqgtr \mathbb{E}[D]$, *and* $\int_{d_L}^{d_U} w^{s}(l)\, dl = 1$.

Under strong parallel trends, TWFE gets convex ACR

Theorem 3.4. *Under Assumptions 1, 2(a), 3, and 4, β^{twfe} can be decomposed in the following ways:*

(a) Causal Response Decomposition:

$$\beta^{twfe} = \int_{d_L}^{d_U} w_1^{acr}(l) \left(ACRT(l|l) + \underbrace{\frac{\partial ATT(l|h)}{\partial h}\Big|_{h=l}}_{\text{selection bias}} \right) dl + w_0^{acr} \frac{ATT(d_L|d_L)}{d_L}$$

where the weights are always positive and integrate to 1.

- Under strong parallel trends, the selection bias terms disappear (and with $d_l = 0$, last term disappears in (a))
- Weights are also convex and weakly causal! However, don't necessary map to an estimand of interest. Related to OLS weights from previous discussions
- ATT composition does not have convexity property

So what can you do?

- Easiest thing is to use untreated units as baseline category
- When treatments are discrete:

$$\Delta Y_i = \beta_0 + \sum_j 1(D_i = d_j)\beta_j + \epsilon_i \tag{7}$$

Under parallel trends, each β is $ATT(d|d)$, and with strong parallel trends, $\beta_j - \beta_{j-1}$ is $ACR(d_j)$.

- Life is more complicated when d is continuous of the groups are quite small at a given treatment:

$$\Delta Y_i = \sum_k \Psi_{Kk}(D)\beta_{Kk} + \epsilon_i, \tag{8}$$

where $\Psi_{Kk}(D)$ is a basis function for the treatment (implementation can be found in paper)

Multiple Treatments in Diff-in-diff

- Hull (2018) and deChaisemartin and D'hautefeuille (2022) are relevant cites

- One "easy" context that this shows up (Hull 2018) is "mover" designs
  - I move from city i to city j – if my move was random, can we use it to understand the effect of city j?
  - Without strong additional assumptions, challenging to interpret coefficients from standard TWFE

$$Y_{it} = \alpha_i + \tau_t + \sum_{j \neq 0} \beta_j D_{ijt} + \epsilon_{it} \tag{9}$$

- Consider 2 period case, and use first differences:

$$\Delta Y_i = \tau + \sum_{j \neq 0} \beta_j \Delta D_{ij} + \Delta \epsilon_i \tag{10}$$

- Recall Goldsmith-Pinkham, Hull and Kolesar (2022) – multiple treatments (ΔD_{ij}) is potentially contaminated and negative weighted.

Random Timing in Diff-in-Diff

- Athey and Imbens (2022) take design-based approach to diff-in-diff
  - What does this mean? Can consider a well-defined propensity scores for the treatment
  - Focus on staggered adoption with binary treatment
  - Since treamtent is absorbing, this simplifies problem into "when did I get the treatment (if ever)?"

- E.g. define a propensity score: $Pr(A_i = a)$ over when the assignment occurs.

- Key question: who is the relevant counterfactual group?

Three key assumptions (not testable!)

1. Random assignment (can make this happen by design)
2. No anticipation (we assume this already)
3. Invariance to history ($1_{a \leq t}(Y_{it}(a) - Y_{it}(1)) = 0$) (no causal effect of an early adotpion vs. a later adoption on outcome as long as adoption occured before or on period t)

Efficiency in estimating staggered random roll-out

- Roth and Sant'anna (2023) show that it is much more efficient to condition on lagged outcomes over using standard diff-in-diff in staggered roll-outs

- E.g. Consider the class of estimators:

$$\hat{\theta}_\beta = (\overline{Y}_{22} - \overline{Y}_{2\infty}) - \beta(\overline{Y}_{12} - \overline{Y}_{1\infty}) \tag{11}$$

  Did is the special case of $\beta = 1$.

- We want to put more weight on this setting when the lagged outcomes are more predictive, and less when it is not!
  - Intuition is similar to synthetic controls

Covariates

- With time-invariant covariates that treatment does not affect, controlling for covariates is conditional parallel trends assumption

$$E(Y_{i,2}(0) - Y_{i,1}(0)|D_i = 1, X_i) = E(Y_{i,2}(0) - Y_{i,1}(0)|D_i = 0, X_i) \tag{12}$$

- However, this isn't enough to satisfy this TWFE regression approach in two periods:

$$Y_{it} = \alpha_i + \phi_t + 1(t = 2)D_i\beta + X_i 1(t = 2)\gamma + \epsilon_{it} \tag{13}$$

Why? (consider age)

- Even worse, if covariates affected by treatment, this is a form of collider bias!
  - Should be very careful about time-varying covariates

Checklist

- First, consider the treatment features of your setup
  - What is your treatment? Binary or continuous?
  - Single or multiple?
  - Absorbing or not?
  - How many treated units? Single or many?
  - How many interventions? Single or staggered timing?
- Second, consider the relevant identifying assumptions for your condition
- Third, what tests are feasible in your setting?
- Fourth, what estimates do you want?
- Fifth, what relaxations of assumptions can you make?

Treatment features

- If treatment is binary or only a few values, easy to consider the simple case
  - Particularly helps if it's absorbing!
- If continuous and/or not absorbing, need to worry about many more homogeneity assumptions
- First order of business – *define your estimand*.
  - What treatment do you care about?
- Can you simplify the problem in some way if it's more complex?

Assumptions

- Write out carefully in math and then words what assumptions you need
  - Parallel trends (with or without covariates)?
  - No anticipation?
  - Others?

- If you know your estimand, much easier to know what you need

- Do you need to construct your estimator using lagged outcomes, e.g. synthetic control?

Testable assumptions

- What is testable? Pre-trends can be useful, but underpowered
- If using random timing, not much is testable
  - But could look for balance across timing groups in exogeneous features

What estimates do you want?

- Do you need a long run effect? Short-run?
  - Long-run effect relies *heavily* on parametric extrapolation. Is it plausible?
- Are you studying ATTs? CATTs? ACRTs?

Relaxation of assumptions?

- What group comparaisons are you making for parallel trends?
- Could you assume random timing instead of parallel trends?
- Could you condition on lagged outcomes?