Difference-in-Differences and Roll-out (Staggered) Treatment Designs

Daniel Halvarsson

February 12, 2024

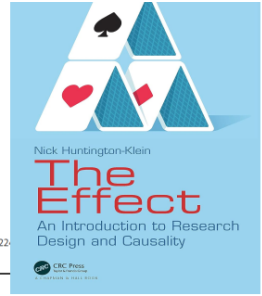
Don’t take my word for it

Difference-in-Differences with Multiple Time Periods

Brantly Callaway
University of Georgia

Designing Difference in Difference Studies With Staggered Treatment Adoption: Key Concepts and Practical Guidelines

Seth M. Freedman, Alex Hollingsworth, Kosali I. Simon, Coady Wing & Madeline Yozwiak



ELSEVIER

Journal of Econometrics
Volume 235, Issue 2, August 2023, Pages 2218-22

What’s trending in difference-in-differences? A synthesis of the recent econometrics literature ☆

ro H.C. Sant'Anna [b] ✉, Alyssa Bilinski [a] ✉,

paulgp / applied-methods-phd

Code | Issues 2 | Pull request

main | 3 branches | 0 tags

paulgp Clean up folder for 2024

exam_questions

lecture_notes

Difference-in-Difference and empirical relevance

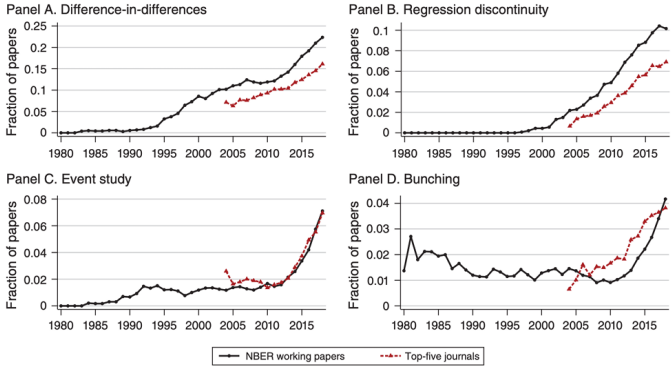


FIGURE 4. QUASI-EXPERIMENTAL METHODS

Notes: This figure shows the fraction of papers referring to each type of quasi-experimental approach. See Table A.I for a list of terms. The series show five-year moving averages.

- Difference-in-Difference (or DiD) is one of the most widely used methods for estimating causal effects in non-experimental settings.

DiD and causal parameter estimation

- What *type of causality* are we after?
- In a DiD setting the likely **causal parameter (estimand) of interest** is the ATT (average treatment effect of the treated).
- It is well known that for basic setups the DiD estimate corresponds to the ATT under the combined assumption of *parallel trends* and *no-anticipation*.
- The work horse model for estimating DiD is the so called two-way fixed effect linear regression model, or for short **TWFE**.

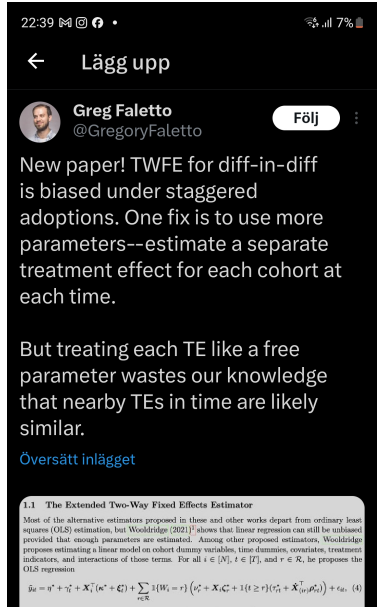
The problem with TWFE and ”the secrete shame of econometrics”

- The TWFE model is well understood and commonly applied to the basic 2×2 setup or to the multiple periods (event-study) setup.
- However, for a roll-out (staggered) treatment design, with different groups getting treated at different times, the TWFE model no longer works.
- **The problem** can be quite serious: with DiD estimates displaying a negative average effect despite the true effect being positive for everyone in the sample.

”For decades researchers were basically unaware of this problem and used two-way fixed effects anyway.” - The Effect

The State of the DiD literature

- Things are moving fast
- This literature has had a certain amount of upheaval over the past 5-6 years.
- With the upheaval there is a **tension** for how people currently and historically have used DiD.
- The modern literature has pointed out many issues but has provided solutions to almost all of the them.
- Good tools are now readily available (including Stata), so nothing to prevent you from using a DiD with staggered analysis.



Simple setup 2×2 DiD

- Assume we have n units i and $T = 2$ time periods t
- Consider a **binary** policy D_{it}, and we are interested in estimating its effect on outcomes Y_{it}.
- Consider the **potential outcome** notation for Y_{it}:
  - $Y_{it}(0,0)$ is the outcome in period $t \in \{1,2\}$ if untreated in both periods.
  - $Y_{it}(0,1)$ is the outcome in period $t \in \{1,2\}$ if untreated in first period, treated in second.
  - Can simplify to just $Y_{it}(0)$ and $Y_{it}(1)$, but when we have many time periods, want to account for path of treatments.
- The inherent problem is that D_{it} is **not** necessarily randomly assigned, but we still want to estimate the ATT in period 2:

$$\tau_2^{ATT} = E(Y_{i,2}(1) - Y_{i,2}(0)|D_i = 1)$$

- It's a measure of *the average causal effect on the treated units* at the time they are treated.

How can we identify the ATT in the basic 2x2 DiD setup?

- The challenge lies in the untreated potential outcome $Y_{i,2}(0)$ for the treated group, **which is unobserved**.

- Difference in difference methods overcome this challenge via two assumptions, namely:

1. Parallel Trends: "in the **absence** of the treatment, the average outcomes would have evolved in parallel"

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

2. No-anticipation: policy has no effect prior to treatment

$$Y_{i,1}(0) = Y_{i,1}(1)$$

- Under assumptions 1 and 2 it follows that τ_2^{ATT} is identified by the DiD of population means (i.e. DiD is unbiased estimator of ATT):

$$\Delta_{DiD} = E(Y_{i,2} - Y_{i,1}|D_i = 1) - E(Y_{i,2} - Y_{i,1}|D_i = 0) = \tau_2^{ATT}.$$

Estimation using linear regression (TWFE) in the 2×2 setup

- Provided D_i is not correlated with α_i, a simple linear regression will identify the ATT effect with two time periods:

$$Y_{ist} = \alpha + \beta_1 Treat_s + \beta_2 Post_t + \beta_{DiD} Treat_s \times Post_t + \epsilon_{it} \quad (1)$$

  where β_{DiD} is the Difference-in-Difference estimate corresponding to Δ_{did}.

- If D_i is correlated with α_i a fixed effects regression will identify the ATT:

$$Y_{it} = \alpha_i + \gamma_t + D_{it}\beta_{DiD} + \epsilon_{it} \quad (2)$$

- This latter setup is referred to as the Two-way Fixed Effects estimator (TWFE).

- Necessary condition: two time periods! (within estimation is equivalent to first-differences estimation) What if we have more periods?

Multiple time periods in basic setup event study

- More time periods helps in several ways:

1. If we have multiple periods *before* the policy implementation, we can partially test the underlying assumptions
2. If we have multiple periods *after* the policy implementation, we can examine effect timing
   - If you pool all time periods together into one "post" variable, this estimates the average effect.

- How is an event-study implemented? Let's consider a policy that occurs all at t_0. Event-study TWFE yields:

$$Y_{it} = \sum_{h=1}^{t_0-2} \alpha_h \underbrace{1\left[D_i = 1\right] \times 1\left[t = h\right]}_{\text{Pre-period d.v.}} + \sum_{k=t_0}^{T} \beta_k \underbrace{1\left[D_i = 1\right] \times 1\left[t = k\right]}_{\text{Post-period d.v.}} + \alpha_i + \gamma_t + \varepsilon_{ist}$$

- It works by comparing outcomes with a never treated group in the pre-period time $t = t_0 - 1$.

Pre-testing and structural assumptions

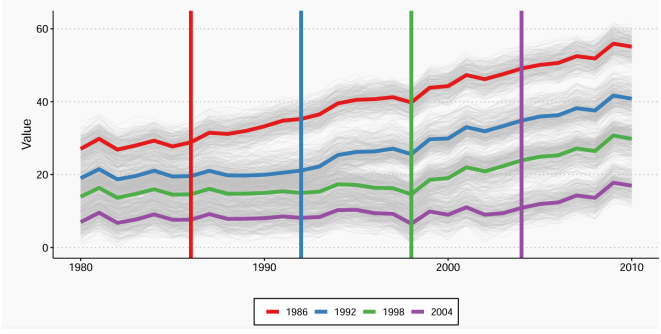
- To partially test the assumption we can impose a stronger assumption about trends
  - If the common trends assumption applies to **all periods**, not just the post-periods in combination with no-anticipation we expect that $ATT(t_0, t_0 - h) = 0$. Thus, rejection the null that $\alpha_h = 0$ implies that the stronger common trend + no anticipation assumption fail.

- This is very powerful and has helped spark the growth in DinD regressions
  - Visual demonstrate of "pre-trends" **helps support the validity of the design**
  - Worth doing!

- Two key issues:
  1. Pre-testing can cause statistical problems as it will cause you to potentially contaminate your design.
     - See suggested solution from Roth (2020) to incorporate robustness to pre-trends into your analysis. Rambachan and Roth (2020) present results on testing sensitivity of DiD results to pre-trends.
  2. What do parallel trends even mean?
     - Roth and Sant'Anna (2021) directly discuss this issue at length.

Roll-out adoption (staggered treatment) design

- Expands the event-study and 2×2 designs to allow for *multiple groups* with *different treatment adoption dates*. Example is a policy with a regional roll-out such that the policy is adopted in different areas at different dates.



- **Staggered treatment adoption** simply means that: $D_{i,t} = 1 \implies D_{i,t+1} = 1$ for $t = 1, 2, ..., T$, such that the treatment is absorbing and does not turn on-off.

”We focus on the setting where units, e.g., individuals, firms, or states, adopt the policy or treatment of interest at a particular point in time, and then remain exposed to this treatment at all times afterwards. **The adoption date at which units are first exposed to the policy may, but need not, vary by unit**. We refer to this as a staggered adoption design” (Athey and Imbens, NBER 2018)

Staggered treatment as many different sub-experiments

- For each treated group s, the causal parameter $ATT(t_s, t_s + k)$ and the DiD comparison is identical to the event-study specification, but where the control group is defined as all groups with adoption dates after the focal post-period.

- Plenty of causal effect estimates: What to do with them?
  - In a small design, it could be be beneficial to analyze each of the sub-experiments separately to gain insight into treatment effect heterogeneity, although it may be statistically inefficient.

  - In a larger study, a separate analysis may not simply be feasible (nor desirable).

- For both cases it's often desirable to average or aggregate the *group* $\times$ *time* estimates into a single causal effect parameter or envent-study parameters (ATT)

- For this purpose we **can't rely on the TWFE model for aggregation**, without imposing strong and quite artificial assumptions. In general the TWFE model don't map to any conceptual causal parameter of interest.

When TWFE is NOT a problem in staggered adoption design

- However, despite the severity of the problem, as emphasized at the beginning, there are situations where TWFE estimates may not be very different from more theoretically favorable approaches, and therefore be appealing because of their simplicity.

- Given the TWFE specification

$$Y_{i,t} = \alpha_i + \gamma_t + D_{i,t}\beta + \varepsilon_{i,t},$$

  with $D_{i,t}$ being an indicator for whether unit i is treated in time t.

- If there is no treatment effect heterogeneity across both units or time the population regression coefficient β equals the ATT.

- There is a common alternative ”dynamic specification” to the ”static” TWFE above with time relative to treatment. This version of the TWFE results in sensible estimates when there is only treatment effect heterogeneity in the time size treatment.

The problem

- The problem is somewhat complex.
- TWFE relies on **within variation** for comparing treated and controls. It means that units that *remains untreated* during the period end up as controls, but the same goes for units that *remains treated* from earlier roll-outs.
- **IF** treatment effect varies with treatment time e.g. effects grow larger over time, earlier treated units in the control group will have an increasing trend that is distinct from just-now-treated units,
- **THEN** parallel trends assumption breaks and identification fails.

Goodman-Bacon Decomposition

- An influential paper, Goodman-Bacon (2021) showed that TWFE estimators in a staggered setting can be expressed as a weighted average of the of all underlying DiDs.

- **Their contribution** was the insight that some of the DiDs are actually confounded despite them individually satisfying the common trend + no anticipation assumptions!

- There are 4 types of DiD comparisons between treated and control implicitly made in the TWFE model
  1. Treated (early and late) vs. Never treated DiDs
  2. Early vs. Late DiDs
  3. Late vs. Early DiDs ☠

  The potentially problematic comparison being the 'Late vs. Early' where a treatment group is compared to a group that is already treated! .

- How large is the problem: Use **bacondecomp** in Stata for Goodman-Bacon decomposition.

What to do with staggered timing in DiD?

"What to do then, when we have a nice roll-out design? Don't use two-way fixed effects, but also don't despair. **You're not out of luck, you're just moving into the realm of what the pros do**." (The effect)

- There's really no reason to use the baseline TWFE in staggered timings
  - A perfect example wherein the estimator does not generate an estimate that maps to a meaningful estimand
- There are different approaches proposed in the literature that are just as good!
  - Here, I look at Callaway and Sant'anna (2020) and Sun and Abraham (2020)
- These all are robust to this issue, but circumstances may vary slightly, e.g. are the **panel balanced** or the **treatment absorbing**?

Modified event-study design: Sun and Abraham, 2020

- Starting from the basic event-study design, it can be modified to include many groups with a staggered roll-out.

- First, each of the sub-experiments are **centered** relative to their own treatment start.
  - It keeps track on the already treated groups so that they don't get included in comparisons.

- Second, Sun and Abraham then propose that the relative time dummies are **interacted** with group dummies such that each treated $group \times relative - time - dummies$ get their own effect.

- It's then up to you to avoid making bad comparisons when averaging coefficients to get e.g. the time-varying treatment effect (see example below).

- Each effect is either compared to the group of (i) never treated observation or (ii) not yet treated observations (from the last treated group(s)).

Modified event-study design: Sun and Abraham, 2020

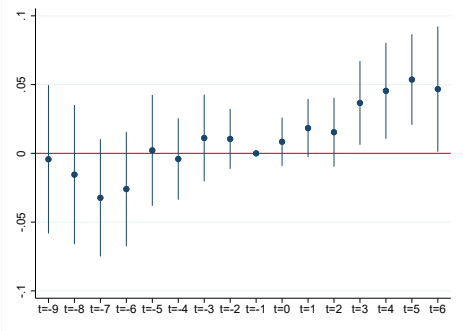
- Sun and Abraham (2020) can be accessed in Stata using the **eventstudyintereact** package.
- In my experience:
  - it's easy to implement, but not very robust when tinkering with choice parameter.
  - Ideally you have a balanced panel (which I did not).
  - Beware of the potentially huge number of parameter to be estimated (the new Twitter estimator).
  - It's also fairly straight forward to compare effects across groups, e.g. how males and females are affected differently by foreign acquisitions (Halvarsson et. al., WP 2023)
  - Lastly, it's very difficult to get customized output in the form of aggregate result tables.

Stata: Sun and Abraham, 2020

```
* Relative time to treatment
gen t = year-treatment_year

* Pretreatment
forvalues k = 9(-1)2 {
    cap drop g_`k'
    gen g_`k' = t == -`k'
}

* Treatment and posttreatment
forvalues k = 0/7{
    cap drop g`k'
    gen g`k' = t == `k'
}
****  Dummy for control group (2015)
gen last_treated_cohort_2015 = 1 if treatment_year >= 2015 & treatment_year !=.
replace last_treated_cohort_2015 = 0 if last_treated_cohort_2015 != 1 & treated == 1
```

Stata: Sun and Abraham, 2020

```
* Event study
eventstudyinteract gap g_* g0-g6 if ${cond1},  cohort(treatment_year) ///
control_cohort(last_treated_cohort_2015) absorb(OrgLopNr year) vce(cluster OrgLopNr)

* Visualize
matrix C = e(b_iw)
mata st_matrix("A",sqrt(diagonal(st_matrix("e(V_iw)"))))
matrix A = A'
matrix C = (C[1,1..8],J(1,1,0),C[1,9..15]) \ (A[1,1..8],J(1,1,0),A[1,9..15])
mat coln C = "t=-9" "t=-8" "t=-7" "t=-6" "t=-5" "t=-4" "t=-3" "t=-2" "t=-1" "t=0" "t=1" "t=2" "t=3" "t=4" "t=5" "t=6"
mat list C

coefplot matrix(C[1]), se(C[2]) vertical yline(0) plotregion(color(white))  graphregion(color(white))
```



Callaway and Sant'anna (2020)

- Callaway and Sant'anna (2020) propose the following building block estimand:

$$\tau_{ATT}(g,t) = E(Y_{it}(g) - Y_{it}(\infty)|D_i = g),$$

the ATT in period t for those units first treated in period g.

- For example, $\tau_{ATT}(2015, 2020)$, gives the average treatment effect in 2020 for the units first treated in 2015
  - In the 2x2 case, this was exactly our effect!

- The ATT can be identified by comparing the outcome at t with that in $g-1$ for the treated group g, with the outcome in t compared to the outcome in $g-1$ for a comparison group that is not yet treated. Formally

$$ATT(g,t) = E(Y_{i,t} - Y_{i,g-1}|G_i = g) - E(Y_{i,t} - Y_{i,g-1}|G_i = g') \text{ for any } g' > t.$$

Callaway and Sant'anna (2020)

- Using these estimands, C&S provide a very natural set of potential ways to aggregate these estimands. Take e.g. the "event-study" of the average treatment effect k times after first treatment across the different groups g (cohorts)

$$ATT_k^w = \sum_g w_g ATT(g, g+k) \tag{3}$$

- The weight w_g can be chosen to weight groups equally, or relative their share of the treated sample.
- C&S highlights two advantages:
  1. It works! And gives full control over the weighting scheme compared to TWFE
  2. Makes transparent which units are used as a control group.
- But they also contributes by allowing for time invariant covariates to be incorporated in the model (Not going to talk about that here).

Callaway and Sant'anna (2020)

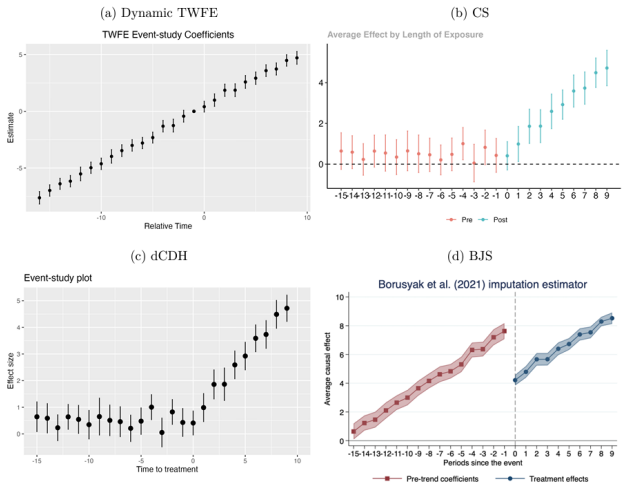
- Have not yet tried C&S myself, but it's sounds intuitive.
- It's one of the most popular staggered DiD estimators alongside: de Chaisemartin and d'Haultfoeuille (2020), which allows for non-absorbing treatments, the "stacked" estimator of Baker, Larcker, and Wang (2021); Wooldridge (2021) two-way Mundlak regrssion; and Dube, Girardi, Jordà & Taylor (2023) local projections approach to staggered DiD, which we ended up using in Halvarsson et al. (2023)
- C&S is now part of the standard DiD package in Stata under **csdid**.
- A further note on covariates in general:
  - While many modern estimators allow for weaker identifying assumptions of parallel trends conditional on a set of time invariant covariates time variant covariants requires strong assumptions:
  - That covariates are not affect by the treatment and don't impact the effect of the treatment.

Most recent contributions

- Roth (2024) ”Interpreting Event-Studies from Recent Difference-in-Difference Methods” identifies an inherent flaw relying on *visual heuristics* for analyzing estimates in e.g. C&S, which can display a kink at the time of treatment.

- Post treatment estimates are still valid, if parallel trends hold. Can’t rely on visuals to judge whether post treatment estimates result from a violation of parallel trends.

- Remedy: Use ”long-differences” for both pre-treatment estimates and post treatment-estimates in event-study plots (the same baseline)



Conclusion

No reason to feel **ashamed** any longer: Software exists. This is doable!

Heterogeneity Robust Estimators for Staggered Treatment Timing		
Package	Software	Description
did, csdid	R, Stata	Implements Callaway and Sant'Anna (2021)
did2s	R, Stata	Implements Gardner (2021), Borusyak et al. (2021), Sun and Abraham (2021), Callaway and Sant'Anna (2021), Roth and Sant'Anna (2021)
didimputation, did_imputation	R, Stata	Implements Borusyak et al. (2021)
DIDmultiplegt, did_multiplegt	R, Stata	Implements de Chaisemartin and D'Haultfœuille (2020)
eventstudyinteract	Stata	Implements Sun and Abraham (2021)
flexpaneldid	Stata	Implements Dettmann (2020), based on Heckman et al. (1998)
fixest	R	Implements Sun and Abraham (2021)
stackedev	Stata	Implements stacking approach in Cengiz et al. (2019)
staggered	R	Implements Roth and Sant'Anna (2021), Callaway and Sant'Anna (2021), and Sun and Abraham (2021)
xtevent	Stata	Implements Freyaldenhoven et al. (2019)
DiD with Covariates		
Package	Software	Description
DRDID, drdid	R, Stata	Implements Sant'Anna and Zhao (2020)
Diagnostics for TWFE with Staggered Timing		
Package	Software	Description
bacondecomp, ddtiming	R, Stata	Diagnostics from Goodman-Bacon (2021)
TwoWayFEWeights	R, Stata	Diagnostics from de Chaisemartin and D'Haultfœuille (2020)
Diagnostic / Sensitivity for Violations of Parallel Trends		
Package	Software	Description
honestDiD	R, Stata	Implements Rambachan and Roth (2022b)
pretrends	R	Diagnostics from Roth (2022)

Note: This table lists R and Stata packages for recent DiD methods, and is based on Asjad Naqvi's repository at https://asjadnaqvi.github.io/DiD/. Several of the packages listed under "Heterogeneity Robust Estimators" also accommodate covariates.