

# Advanced DID

*MULTIPLE PERIODS + STAGGERED TREATMENT TIMING*

---



# Staggered Timing

- Remember that in the canonical DiD model we had:
  - Two periods and a common treatment date
  - Identification from parallel trends and no anticipation
  - A large number of clusters for inference
- A very active recent literature has focused on relaxing the first assumption: **what if there are multiple periods and units adopt treatment at different times?**
- This literature typically maintains the remaining ingredients: parallel trends and many clusters

# Overview of Staggered Timing Literature

1. Negative results: TWFE OLS doesn't give us what we want with treatment effect heterogeneity
2. New estimators: perform better under treatment effect heterogeneity

# Staggered timing set-up

- Panel of observations for periods  $t = 1, \dots, T$
- Suppose units adopt a binary treatment at different dates  $G_i \in \{1, \dots, T\} \cup \infty$  (where  $G_i = \infty$  means “never-treated”)
  - Literature is now starting to consider cases with continuous treatment & treatments that turn on/off – that lit is still developing (see Section 3.4 of review paper)
- Potential outcomes  $Y_{it}(g)$  – depend on time and time you were first-treated

# Extending the Identifying Assumptions

- The key identifying assumptions from the canonical model are extended in the natural way
- **Parallel trends:** Intuitively, says that if treatment hadn't happened, all "adoption cohorts" would have parallel average outcomes in all periods

$$E[Y_{it}(\infty) - Y_{i,t-1}(\infty)|G_i = g] = E[Y_{it}(\infty) - Y_{i,t-1}(\infty)|G_i = g'] \text{ for all } g, g', t$$

Note: can impose slightly weaker versions (e.g. only require PT post-treatment)

- **No anticipation:** Intuitively, says that treatment has no impact before it is implemented

$$Y_{it}(g) = Y_{it}(\infty) \text{ for all } t < g$$

# Negative results

- Suppose we again run the regression

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it},$$

where  $D_{it} = 1[t \geq G_i]$  is a treatment indicator.

- Suppose we're willing to assume no anticipation and parallel trends across all adoption cohorts as described above

# Negative results

- Suppose we again run the regression

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it},$$

where  $D_{it} = 1[t \geq G_i]$  is a treatment indicator.

- Suppose we're willing to assume no anticipation and parallel trends across all adoption cohorts as described above
- Good news: if treatment effects are constant across time and units,  $Y_{it}(g) - Y_{it}(\infty) \equiv \tau$ , then  $\beta = \tau$

# Negative results

- Suppose we again run the regression

$$Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it},$$

where  $D_{it} = 1[t \geq G_i]$  is a treatment indicator.

- Suppose we're willing to assume no anticipation and parallel trends across all adoption cohorts as described above
- Good news: if treatment effects are constant across time and units,  $Y_{it}(g) - Y_{it}(\infty) \equiv \tau$ , then  $\beta = \tau$
- Bad news: if treatment effects are heterogeneous, then  $\beta$  may put negative weights on treatment effects for some units and time periods
  - E.g., if treatment effect depends on time since treatment,  $Y_{it}(t - r) - Y_{it}(\infty) = \tau_r$ , then some  $\tau_r$ s may get negative weight



# Where do these negative results come from?

- The intuition for these negative results is that the TWFE OLS specification combines two sources of comparisons:
  1. **Clean comparisons:** DiD's between treated and not-yet-treated units
  2. **Forbidden comparisons:** DiD's between two sets of already-treated units (who began treatment at different times)
- These forbidden comparisons can lead to negative weights: the “control group” is already treated, so we run into problems if their treatment effects change over time

## Some intuition for forbidden comparisons

- Consider the two period model, except suppose now that our two groups are **always-treated** units (treated in both periods) and **switchers** (treated only in period 2)
- With two periods, the coefficient  $\beta$  from  $Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}$  is the same as from the first-differenced regression  $\Delta Y_i = \alpha + \Delta D_i \beta + u_i$
- Observe that  $\Delta D_i$  is one for switchers and zero for stayers.  
That is, the stayers are the control group! Thus,

$$\hat{\beta} = \underbrace{(\bar{Y}_{Switchers,2} - \bar{Y}_{Switchers,1})}_{\text{Change for switchers}} - \underbrace{(\bar{Y}_{AT,2} - \bar{Y}_{AT,1})}_{\text{Change for always treated}}$$

- Problem: if the treatment effect for the always-treated grows over time, that will enter  $\hat{\beta}$  negatively!

## Second Intuition for Negative Weights

- The Frisch-Waugh-Lovell theorem says that we can obtain the coefficient  $\beta$  in  $Y_{it} = \alpha_i + \phi_t + D_{it}\beta + \epsilon_{it}$  by the following two-step procedure.
- First, regress the treatment indicator  $D_{it}$  on the FEs (a linear probability model):
$$D_{it} = \tilde{\alpha}_i + \tilde{\phi}_t + \tilde{\epsilon}_{it}$$
- Then run a univariate regression of  $Y_{it}$  on  $D_{it} - \hat{D}_{it}$  to obtain  $\beta$ .  
→ Thus, 
$$\beta = \frac{\text{Cov}(Y_{it}, D_{it} - \hat{D}_{it})}{\text{Var}(D_{it} - \hat{D}_{it})} = \frac{E(Y_{it}(D_{it} - \hat{D}_{it}))}{\text{Var}(D_{it} - \hat{D}_{it})}$$
- However, it's well known that the linear probability model for  $D_{it}$  may have predictions outside the unit interval. If  $\hat{D}_{it} > 1$  even though unit  $i$  is treated in period  $t$ , then  $D_{it} - \hat{D}_{it} < 0$ , and thus  $Y_{it}$  gets negative weight.

# Not just negative but weird...

The literature has placed a lot of emphasis on the fact that some treatment effects may get negative weights

- But even if the weights are non-negative, they might not give us the most intuitive parameter
- For example, suppose each unit  $i$  has treatment effect  $\tau_i$  in every period if they are treated (no dynamics). Then  $\beta$  gives a weighted average of the  $\tau_i$  where the weights are largest for units treated closest to the middle of the panel
- It is not obvious that these weights are relevant for policy, even if they are all non-negative!

# Issues with dynamic TWFE

- Sun and Abraham (2021) show that similar issues arise with dynamic TWFE specifications:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{k \neq 0} \gamma_k D_{i,t}^k + \varepsilon_{i,t},$$

where  $D_{i,t}^k = 1 \{t - G_i = k\}$  are “event-time” dummies.

- Like for the static spec,  $\gamma_k$  may put negative weight on treatment effects after  $k$  periods for some units
- SA also show that  $\gamma_k$  may be “contaminated” by treatment effects at lags  $k' \neq k$

## Dynamic TWFE - Continued

- The results in SA suggest that interpreting the  $\hat{\gamma}_k$  for  $k = 1, 2, \dots$  as estimates of the dynamic effects of treatment may be misleading
- These results also imply that pre-trends tests of the  $\gamma_k$  for  $k < 0$  may be misleading – could be non-zero even if parallel trends holds, since they may be “contaminated” by post-treatment effects!

## Dynamic TWFE - Continued

- The results in SA suggest that interpreting the  $\hat{\gamma}_k$  for  $k = 1, 2, \dots$  as estimates of the dynamic effects of treatment may be misleading
- These results also imply that pre-trends tests of the  $\gamma_k$  for  $k < 0$  may be misleading – could be non-zero even if parallel trends holds, since they may be “contaminated” by post-treatment effects!
- The issues discussed in SA arise if dynamic path of treatment effects is heterogeneous across adoption cohorts
  - Biases may be less severe than for “static” specs if dynamic patterns are similar across cohorts

# New estimators (and estimands!)

- Several new (closely-related) estimators have been proposed to try to address these negative weighting issues
- The key components of all of these are:
  1. Be precise about the target parameter (estimand) – i.e., how do we want to aggregate treatment effects across time/units
  2. Estimate the target parameter using only “clean-comparisons”



## Example – Callaway and Sant’Anna (2020)

- Define  $ATT(g, t)$  to be ATT in period  $t$  for units first treated at period  $g$ ,

$$ATT(g, t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

## Example – Callaway and Sant’Anna (2020)

- Define  $ATT(g, t)$  to be ATT in period  $t$  for units first treated at period  $g$ ,

$$ATT(g, t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- Under PT and No Anticipation,  $ATT(g, t)$  is identified as

$$ATT(g, t) = \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Change for cohort } g} - \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Change for never-treated units}}$$

- Why?

## Example – Callaway and Sant’Anna (2020)

- Define  $ATT(g, t)$  to be ATT in period  $t$  for units first treated at period  $g$ ,

$$ATT(g, t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- Under PT and No Anticipation,  $ATT(g, t)$  is identified as

$$ATT(g, t) = \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Change for cohort } g} - \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Change for never-treated units}}$$

- Why? This is a two-group two-period comparison, so the argument is the same as in the canonical case!

# Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1} | G_i = g] - E[Y_{it} - Y_{i,g-1} | G_i = \infty]$$

# Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$

- Apply definition of POs to obtain:

$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

# Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$

- Apply definition of POs to obtain:

$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Use No Anticipation to substitute  $Y_{i,g-1}(\infty)$  for  $Y_{i,g-1}(g)$ :

$$E[Y_{it}(g) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

# Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$

- Apply definition of POs to obtain:

$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Use No Anticipation to substitute  $Y_{i,g-1}(\infty)$  for  $Y_{i,g-1}(g)$ :

$$E[Y_{it}(g) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Add and subtract  $E[Y_{it}(\infty)|G_i = g]$  to obtain:

$$E[Y_{it}(g) - Y_{it}(\infty)|G_i = g] + \\ [E[Y_{it}(\infty) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]]$$

# Proof of Identification Argument

- Start with

$$E[Y_{it} - Y_{i,g-1}|G_i = g] - E[Y_{it} - Y_{i,g-1}|G_i = \infty]$$

- Apply definition of POs to obtain:

$$E[Y_{it}(g) - Y_{i,g-1}(g)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Use No Anticipation to substitute  $Y_{i,g-1}(\infty)$  for  $Y_{i,g-1}(g)$ :

$$E[Y_{it}(g) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]$$

- Add and subtract  $E[Y_{it}(\infty)|G_i = g]$  to obtain:

$$E[Y_{it}(g) - Y_{it}(\infty)|G_i = g] + \\ [E[Y_{it}(\infty) - Y_{i,g-1}(\infty)|G_i = g] - E[Y_{ig}(\infty) - Y_{i,g-1}(\infty)|G_i = \infty]]$$

- Cancel the last term using PT to get  $E[Y_{it}(g) - Y_{it}(\infty)|G_i = g] = ATT(g, t)$



## Example – Callaway and Sant’Anna (2020)

- Define  $ATT(g, t)$  to be ATT in period  $t$  for units first treated at period  $g$ ,

$$ATT(g, t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

## Example – Callaway and Sant’Anna (2020)

- Define  $ATT(g, t)$  to be ATT in period  $t$  for units first treated at period  $g$ ,

$$ATT(g, t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- Under PT and No Anticipation,

$$ATT(g, t) = \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Change for cohort } g} - \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Change for never-treated}}$$

## Example – Callaway and Sant’Anna (2020)

- Define  $ATT(g, t)$  to be ATT in period  $t$  for units first treated at period  $g$ ,

$$ATT(g, t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g]$$

- Under PT and No Anticipation,

$$ATT(g, t) = \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Change for cohort } g} - \underbrace{E[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Change for never-treated}}$$

- We can then estimate this with sample analogs:

$$\widehat{ATT}(g, t) = \underbrace{\widehat{E}[Y_{it} - Y_{i,g-1} | G_i = g]}_{\text{Sample change for cohort } g} - \underbrace{\widehat{E}[Y_{it} - Y_{i,g-1} | G_i = \infty]}_{\text{Sample change for never-treated}}$$

where  $\widehat{E}$  denotes sample means.

# Aggregation schemes

- If have a large number of observations and relatively few groups/periods, can report  $\widehat{ATT}(g, t)$ 's directly.
- If there are many groups/periods, the  $\widehat{ATT}(g, t)$  may be very imprecisely estimated and/or too numerous to report concisely

# Aggregation schemes

- In these cases, it is often desirable to report sensible averages of the  $\widehat{ATT}(g, t)$ 's.

# Aggregation schemes

- In these cases, it is often desirable to report sensible averages of the  $\widehat{ATT}(g, t)$ 's.
- One of the most useful is to report event-study parameters which aggregate  $\widehat{ATT}(g, t)$ 's at a particular lag since treatment
  - E.g.  $\hat{\theta}_k = \sum_g \widehat{ATT}(g, t + k)$  aggregates effects for cohorts in the  $k$ th period after treatment
  - Can also construct for  $k < 0$  to estimate “pre-trends”

# Aggregation schemes

- In these cases, it is often desirable to report sensible averages of the  $\widehat{ATT}(g, t)$ 's.
- One of the most useful is to report event-study parameters which aggregate  $\widehat{ATT}(g, t)$ 's at a particular lag since treatment
  - E.g.  $\hat{\theta}_k = \sum_g \widehat{ATT}(g, t + k)$  aggregates effects for cohorts in the  $k$ th period after treatment
  - Can also construct for  $k < 0$  to estimate “pre-trends”
- C&S discuss other sensible aggregations too – e.g., if interested in whether treatment effects differ across good/bad economies, may want to “calendar averages” that pool the  $\widehat{ATT}(t, g)$  for the same year

# Comparisons of new estimators

- Callaway and Sant'Anna also propose an analogous estimator using *not-yet-treated* rather than never-treated units.
- Sun and Abraham (2021) propose a similar estimator but with different comparisons groups (e.g. using last-to-be treated rather than not-yet-treated)
- Borusyak et al. (2021), Wooldridge (2021), Gardner (2021) propose “imputation” estimators that estimate the counterfactual  $\hat{Y}_{it}(0)$  using a TWFE model that is fit using only pre-treatment data
  - Main difference from C&S is that this uses more pre-treatment periods, not just period  $g - 1$
  - This can sometimes be more efficient (if outcome not too serially correlated), but also relies on a stronger PT assumption that may be more susceptible to bias
- Roth and Sant'Anna (2021) show that you can get even more precise estimates if you're willing to assume treatment timing is “as good as random”



# Personal advice

- Don't freak out about this new literature!

# Personal advice

- Don't freak out about this new literature!
- In most cases, using the “new” DiD methods will not lead to a big change in your results (empirically, TE heterogeneity is not *that* large in most cases)
  - The exceptions are cases where there are periods where almost all units are treated – this is when “forbidden comparisons” get the most weight

# Personal advice

- Don't freak out about this new literature!
- In most cases, using the “new” DiD methods will not lead to a big change in your results (empirically, TE heterogeneity is not *that* large in most cases)
  - The exceptions are cases where there are periods where almost all units are treated – this is when “forbidden comparisons” get the most weight
- The most important thing is to be precise about who you want the comparison group to be and to choose a method that only uses these “clean comparisons”

# Personal advice

- Don't freak out about this new literature!
- In most cases, using the “new” DiD methods will not lead to a big change in your results (empirically, TE heterogeneity is not *that* large in most cases)
  - The exceptions are cases where there are periods where almost all units are treated – this is when “forbidden comparisons” get the most weight
- The most important thing is to be precise about who you want the comparison group to be and to choose a method that only uses these “clean comparisons”
- In my experience, the difference between the new estimators is typically not that large – can report multiple new methods for robustness (to make your referees happy!)

## References I

- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess**, “Revisiting Event Study Designs: Robust and Efficient Estimation,” *arXiv:2108.12419 [econ]*, 2021.
- Gardner, John**, “Two-stage differences in differences,” *Working Paper*, 2021.
- Roth, Jonathan and Pedro H. C. Sant’Anna**, “Efficient Estimation for Staggered Rollout Designs,” *arXiv:2102.01291 [econ, math, stat]*, 2021.
- Sun, Liyang and Sarah Abraham**, “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects,” *Journal of Econometrics*, 2021, 225 (2), 175–199.
- Wooldridge, Jeffrey M**, “Two-Way Fixed Effects, the Two-Way Mundlak Regression, and Difference-in-Differences Estimators,” *Working Paper*, 2021, pp. 1–89.