Discussion of Baker, Larker and Wang (2021):

How Much Should We Trust Staggered Difference-in-Differences Estimates?

Pedro H. C. Sant'Anna
Vanderbilt University

LSE-UCL Law & Finance Seminar Series, March 2021

Overview

- Provides a practical, hands-on review of pitfalls of using TWFE linear regression models to study treatment effects in staggered DiD designs.
- Very lucid and transparent discussion of **how** TWFE leads to weird weighted sums in “static” TWFE specifications using Goodman-Bacon (2020)
- Intuitive explanation of **how** Callaway and Sant’Anna (2020), Cengiz, Dube, Lindner, and Zipperer (2019), and Sun and Abraham (2020) bypass several of these challenges.
- Shows that these issues arise often in practice by revisiting three empirical applications in Accounting and Finance.
- **Great paper!**

Overall Opinion

- **I like this paper a lot**, to the point I am happy to ignore their **bad** title.
- Let me save some time by not discussing this further as I have more productive things to say!

Main take-away message on how to avoid problems with TWFE

- **Clearly separate the analysis into three steps:**

  1. **Identification:** What kind of variation are we hoping to exploit? Under which assumptions? What is the main building-block of the analysis?

  2. **Aggregation:** How can we leverage all these group-time comparisons to summarize treatment effects?

  3. **Estimation and inference:** What statistical tools do we use? OR, IPW, DR estimators? Should we use simultaneous confidence bands?

Identification

- Importance of defining the building-blocks of the analysis.
  - What are the (disaggregated) parameters of interest? (e.g. $ATT(g,t)$)
  - What are the parallel trends assumptions you are comfortable with? (e.g., PT based on "never-treated", "not-yet-treated", "last-treated cohort")
  - Are we worried about treatment anticipation?
- This step should match "your story".
  - If you care about ATT's, "always-treated" units do not provide information.
  - In the absence of "never-treated" units, data from periods starting from the time last cohort is treated do not "help" in identifying ATT's .

Aggregation

- How do we summarize the average treatment effects across different groups and periods?
- We can construct weigthed averages of the $ATT(g,t)$'s to highlight TE het in a given dimension.
- But we should have full control of how the weights are constructed so we can give these parameters "proper names".
- Paper emphasizes event-study-type parameters, which are great to summarize TE dynamics wrt event time.
- CS also discuss other aggregation schemes

Estimation of event-study-type parameters

- Under no-anticipation and "uncondtional" parallel trends based on "not-yet-treated", CS shows that

$$ATT(g,t) = \mathbb{E}[Y_t - Y_{g-1} | G_g = 1] - \mathbb{E}[Y_t - Y_{g-1} | D_t = 0, G_g = 0].$$

- SA has related results using last-treated cohort as the comparison group.
- Event-study parameters are defined as

$$\theta^{es}(e) = \sum_{g=2}^{\mathcal{T}} \sum_{t=2}^{\mathcal{T}} 1\{t - g + 1 = e\} P(G_g = 1 | \text{Treated for} \geq \text{e periods}) ATT(g,t),$$

- Estimation follows from replacing population expectations with their sample analogues.
- CS also talks about how to appropriately incorporate pre-treatment covariates to allow for covariate-specific trends.

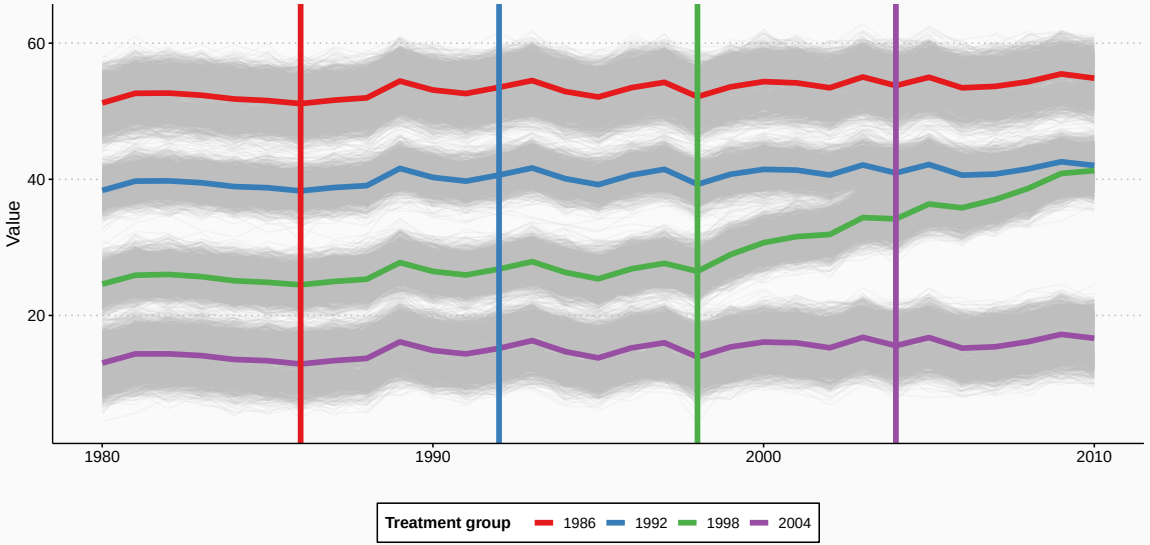
Inference

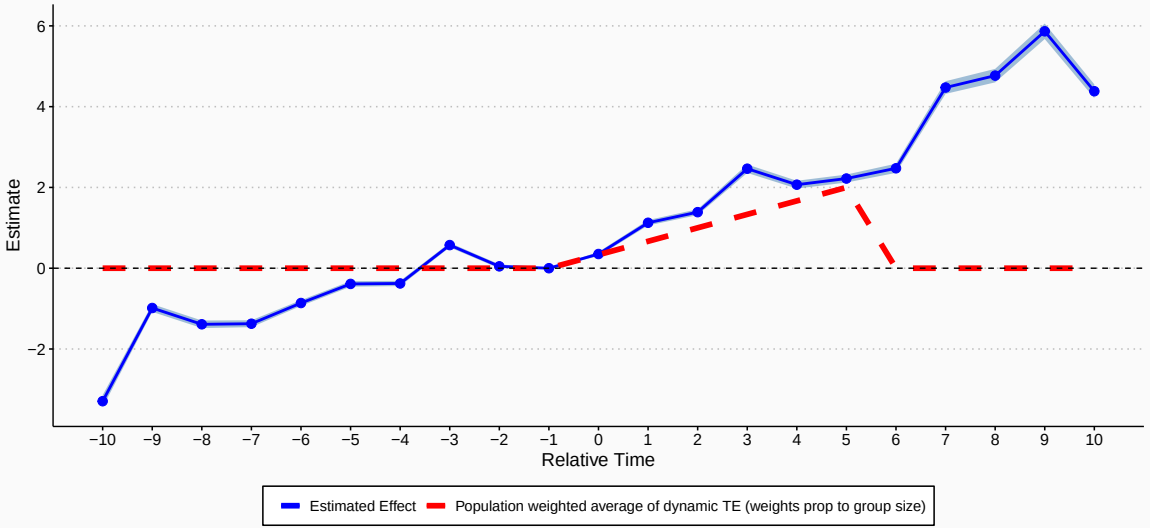
- This paper focus on pointwise inference procedures
  - $\widehat{\theta}(e) \pm 1.96 \times \widehat{SE}(e)$
- When we care about the evolution of average treatment effects wrt event time, this is not really appropriate
- With 10 event-times, you are essentially conducting 10 different hypothesis tests.
- Multiple testing problem!
- Solution: Simultaneous confidence intervals
  - $\widehat{\theta}(e) \pm$ bootstrapped simultaneous critical values $\times \widehat{SE}(e)$

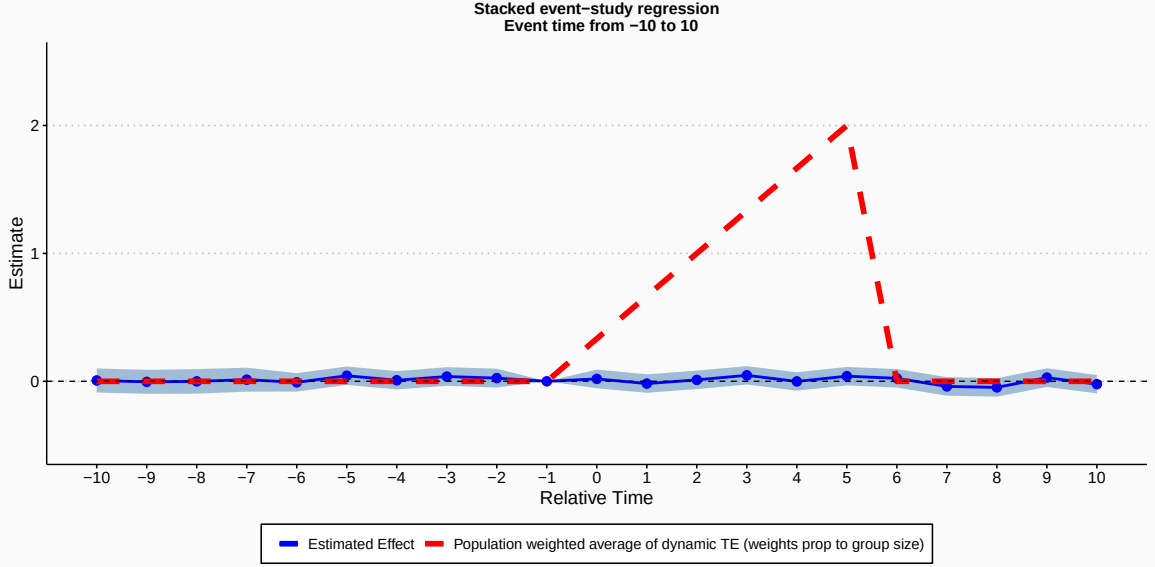
More work is needed to better understand stacked regressions

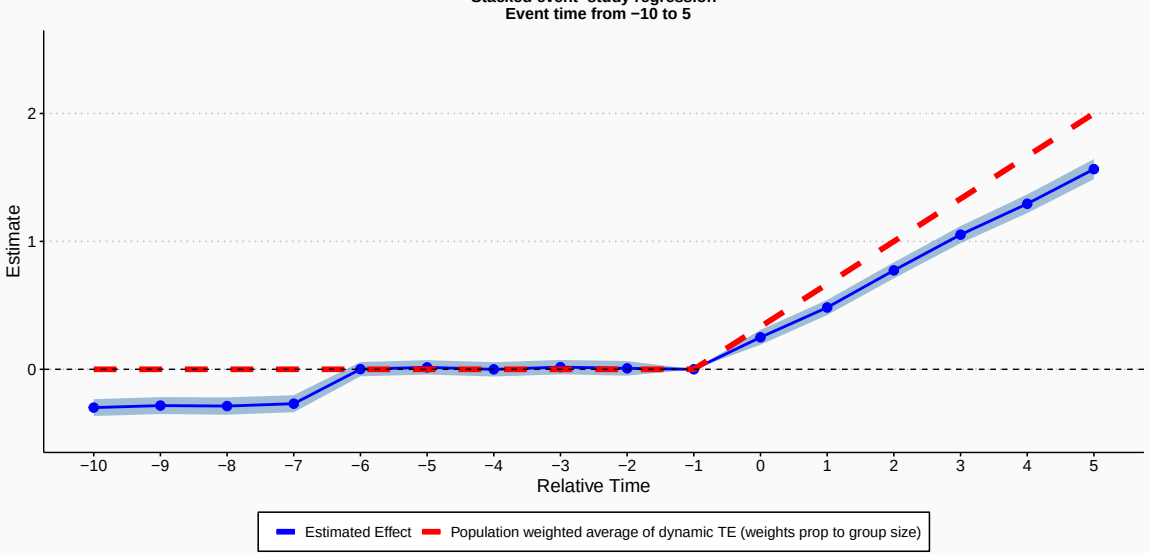
Stacked Regressions - some thoughts

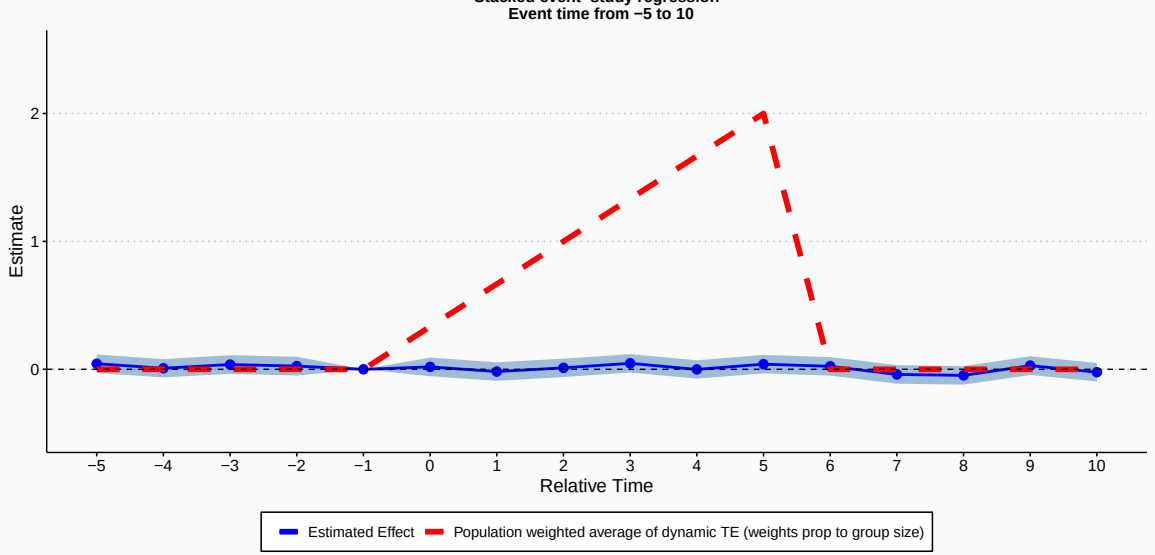
- Do we know the form of the weights that stack-regressions implicitly use?!
- What kind of TE parameter is recovered by it?

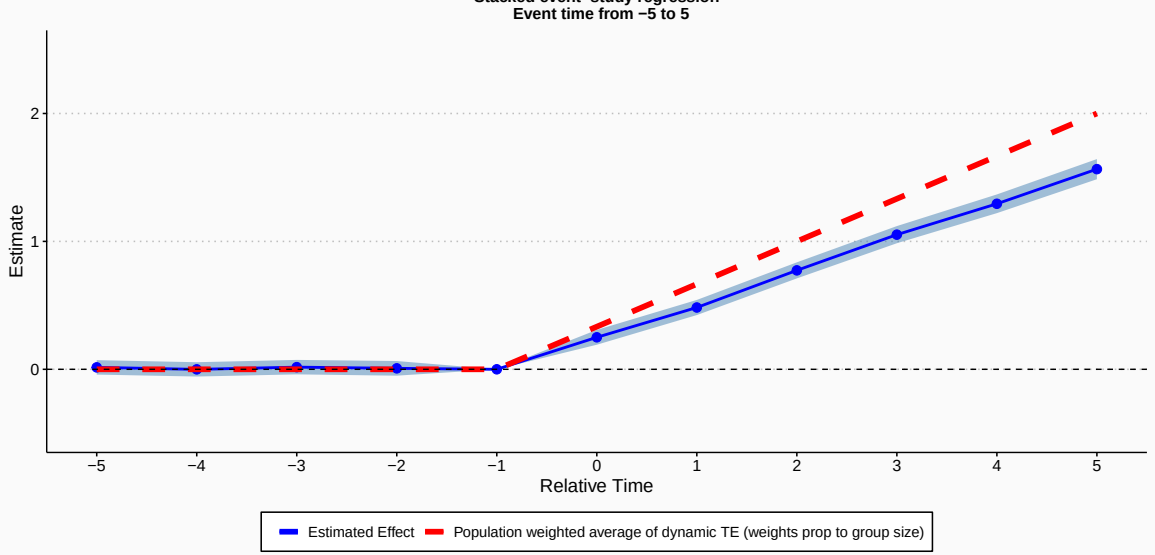

One draw of the DGP with heterogeneous treatment effect dynamics across cohorts
Value
60
40
20
1980
1990
2000
2010
Treatment group
1986
1992
1998
2004

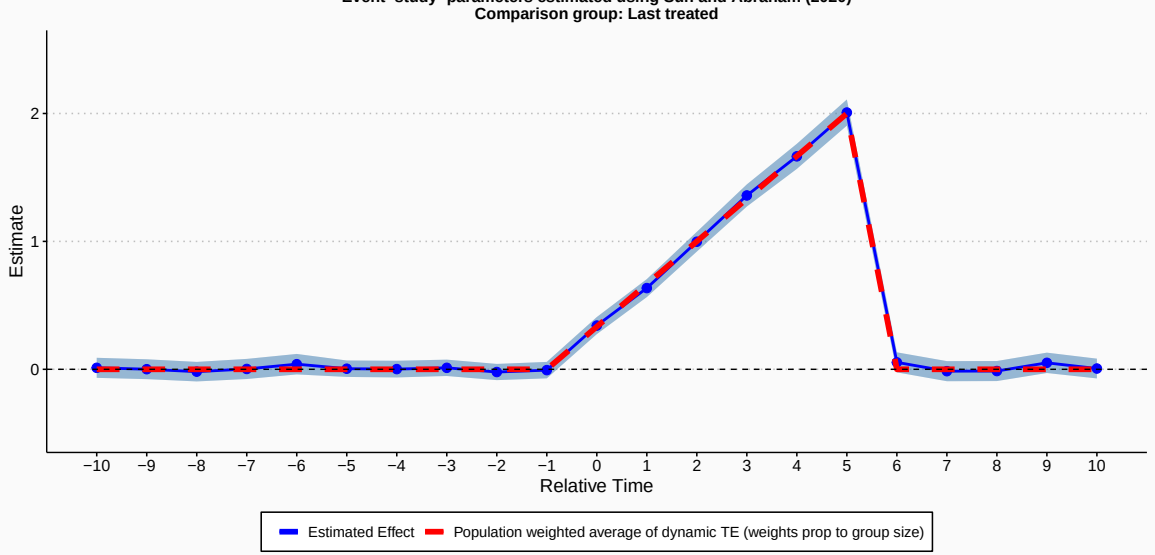

TWFE event−study regression
Binned end−points at −10 and 10
Estimate
6
4
2
0
−2
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
7
8
9
10
Relative Time
Estimated Effect
Population weighted average of dynamic TE (weights prop to group size)

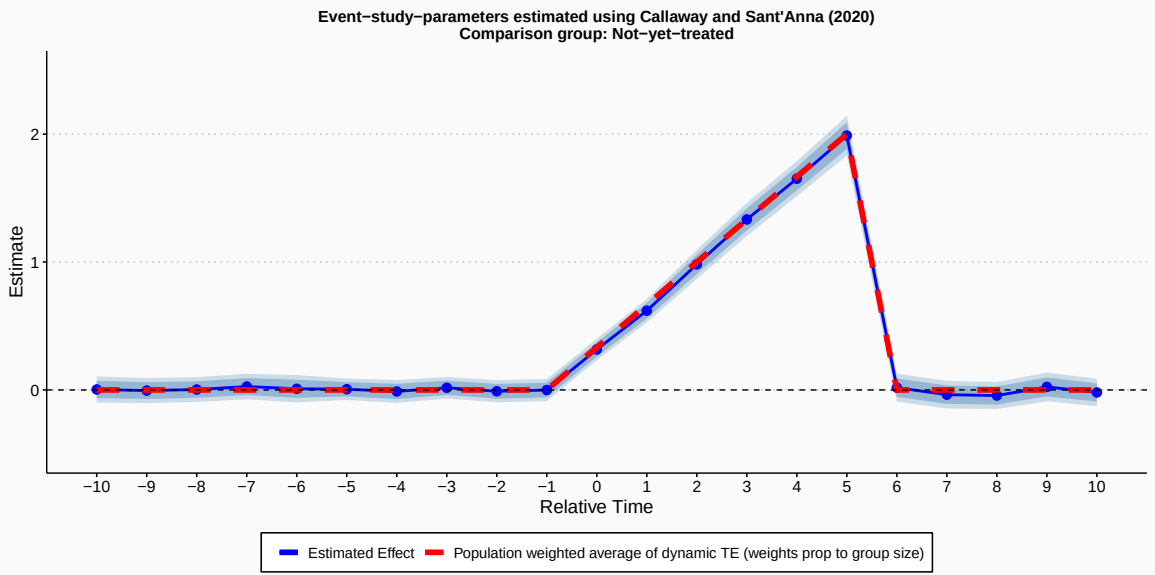

Stacked event−study regression
Event time from −10 to 10
Estimate
2
1
0
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
7
8
9
10
Relative Time
Estimated Effect
Population weighted average of dynamic TE (weights prop to group size)


Stacked event−study regression
Event time from −10 to 5
Estimate
2
1
0
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
Estimated Effect
Population weighted average of dynamic TE (weights prop to group size)


Stacked event−study regression
Event time from −5 to 10
Estimate
2
1
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
7
8
9
10
Relative Time
Estimated Effect
Population weighted average of dynamic TE (weights prop to group size)


Stacked event−study regression
Event time from −5 to 5
Estimate
2
1
0
−5
−4
−3
−2
−1
0
1
2
3
4
5
Relative Time
Estimated Effect
Population weighted average of dynamic TE (weights prop to group size)


Event−study−parameters estimated using Sun and Abraham (2020)
Comparison group: Last treated
Estimate
2
1
0
−10
−9
−8
−7
−6
−5
−4
−3
−2
−1
0
1
2
3
4
5
6
7
8
9
10
Relative Time
Estimated Effect
Population weighted average of dynamic TE (weights prop to group size)



Display both pointwise and simultaneous conf. intervals

Stacked Regressions - some thoughts

- Seems to depend on implementation
- Importance of “balancing on event-time”
  - Only include leads and lags that are well-defined for all observations in your sample
- Even when you do so, it probably recovers some type of variance-based weighted average of ATT’s (this is a conjecture)
- “Better” than standard TWFE as it seems to avoid non-convex/negative weights (this is another conjecture)
- I would still favor CS and SA because we have full control of the weights: Don’t like to let the “estimation method” chooses them for us

Stacked regressions do not explicitly separate “aggregation” and “estimation and inference” steps

Can we do even better than these “modern” DiD procedures?

Can we do better if treatment timing is random?

- All this discussion here rely on the credibility of a PT assumption
- Many times, researchers justify this by claiming that treatment timing is "quasi-random".
- If that is indeed the case, we can do much better than DiD!
  - See my recent paper with Jonathan Roth, "Efficient Estimation for Staggered Rollout Designs."
- If not, need to justify why we believe in parallel trends.
  - See my other recent paper with Jonathan Roth, "When Is Parallel Trends Sensitive to Functional Form?"