CODERS' CORNER

CENTRE FOR THE STUDY OF AFRICAN ECONOMIES

UNIVERSITY OF OXFORD

STAGGERED DIFFERENCE-IN-DIFFERENCE

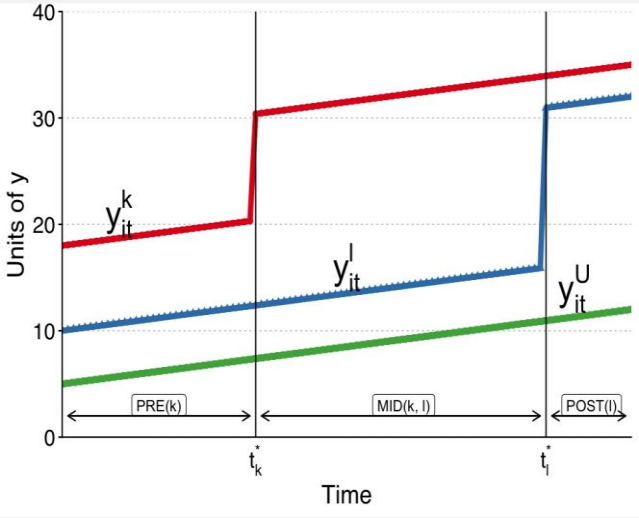
Difference in difference estimators have been a popular empirical strategy for researchers since the "credibility revolution" [Angrist and Pischke, 2008] and they remain incredibly popular. Empirics somewhat got ahead of the theory, and the econometrics has caught up only recently. In doing so, it has discovered some potential problems with the canonical approach. Here, we will focus on issues with differential treatment timing across units as discussed in Goodman-Bacon [2018].

Consider the familiar two-way fixed effect (TWFE) specification:

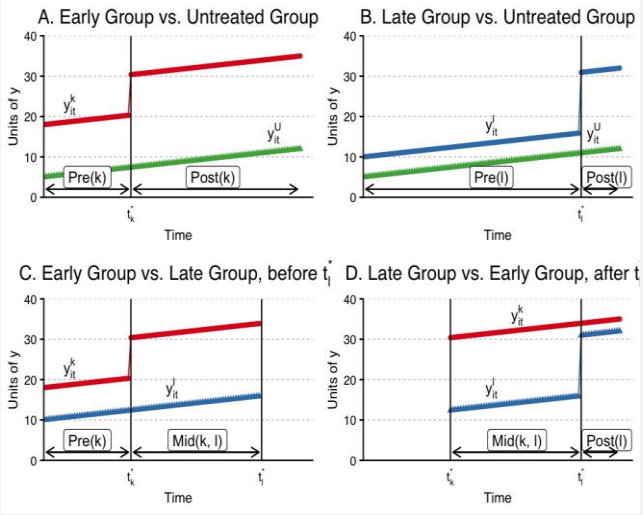
$$yit = \beta \cdot Tit + ai + \tau t + \varepsilon it \quad (1)$$

y_{it} is some outcome of interest, a_i and τ_t are unit (i) and time (t) fixed effects, T_{it} is treatment and ε_{it} is some idiosyncratic error. The coefficient of interest is, as always, β. Now suppose that treatment T_{it} turns on at different times for different units i. Then Goodman-Bacon [2018] shows that $\hat{\beta}$ will be a weighted average of all possible (2x2) *traditional* DiD estimators. A traditional DiD estimator is a 2-period, 2-group 'experiment' where one group is treated in the second period. An example with three treatment groups is given in the figure below.

(a) Overall



(b) Split into each 'experiment'



Colour image of the graph in Goodman-Bacon [2018] as presented in Andrew Baker's slides.

The red group is treated early, blue group late and the green group is never treated. The left-hand figure shows the overall time series of outcomes, and the right-hand figure shows the four 2×2 'experiments' that contribute to $\hat{\beta}$. $\hat{\beta}$ as recovered from a classic two-way fixed effect regression as in equation 1, is a weighted average of each treatment effect where weighted are a function of (i) the size of the sub-sample (ii) relative size of treatment and control units (iii) the timing of treatment in the sub-sample. Note that already-treated units can act as a control (see panel C and D) and now we will need four parallel trends assumptions — one for each experiment.

The Stata command '*bacondecomp*' can be used to provide a decomposition of where the variation in the overall treatment effect is coming from, by plotting each 2×2 estimator against its weight. It does *not* give a recommendation of how to proceed — see references for some suggestions on that.

Example

As an example, I ran the decomposition on constructed data (see attached do file). Units in this data are treated at random times within a 20-period window. To highlight one known issue, I've allowed the treatment effect to increase over time[1]. I run the following code in Stata to implement the Goodman-Bacon decomposition.

```
Use example_data.dta, clear ssc install
bacondecomp

// regression (naive) reghdfe y treat,
Absorb(id t)

// bacon-decomposition xtset id t
Bacondecomp y treat, ddetail

Graph export becondecomp_example.png, width(2000) replace.
```

This generates the following output.

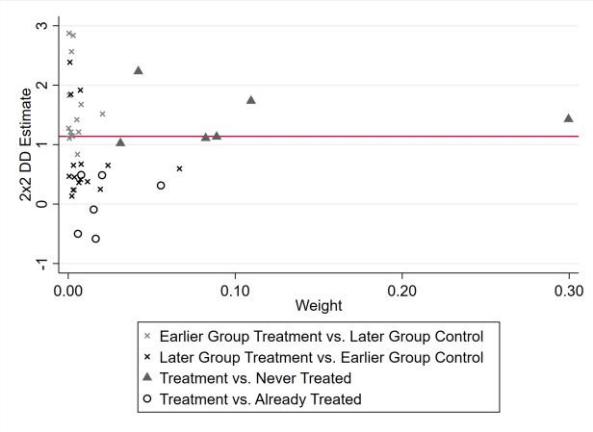
(a) Stata output

```
. bacondecomp y treat , ddetail
Calculating treatment times...
Calculating weights...
Estimating 2x2 diff-in-diff regressions...

Diff-in-diff estimate: 1.139

DD Comparison                Weight      Avg DD Est
------------------------------------------------------
Earlier T vs. Later C        0.061           1.433
Later T vs. Earlier C        0.165           0.601
T vs. Never treated          0.654           1.432
T vs. Already treated        0.121           0.142
------------------------------------------------------
T = Treatment; C = Control
```

(b) Decomposition graph



The left-hand panel shows the naive TWFE regression estimate of 1.139[2]. The overall estimate is then split into four categories giving the weight of each category and its average estimate. "Earlier T vs. Later C" considers units treated early against to-be (but not yet) treated units, "Later T vs. Earlier C" considers units treated later against already treated units, "T vs Never treated" considers treated units vs. those that are never treated and "T vs Already treated" considers treated units vs. those that are always treated.

The 1st and 3rd group are often considered the cleanest, and indeed show high and similar estimates. Groups where previously treated units are used as controls (groups 1 and 4) can be somewhat suspect - and indeed show low estimates.

[1] To be precise treatment is 0.7 plus 0.1 times each period of treatment, so if you have been treated for three periods, the treatment effect in the fourth period is $0.7 + 3 * 0.1 = 1$.

[2] You can check this is significant with the SE of your choosing by running "*reghdfe y treat, absorb(id t) vce(r)*" for example.

The graph in the right-hand figure plots each 2 × 2 estimate against its corresponding weight and sheds some light on potential issues.

Two things jump out from this.

- First, the estimates that used previously treated units as controls (the dark x and circle) are in general much lower than those using untreated units.
- Second, one single 2×2 'experiment' accounts for 30% of the overall estimate - which warrants further investigation.

So, the standard TWFE estimate of DiD with staggered treatment may not give the correct estimate — and Goodman-Bacon gives us a nice way to decompose $\hat{\beta}$ and diagnose potential issues. Although Goodman-Bacon doesn't give a solution, it does hint at the potential sources of good variation. Subsequent work has looked at explicit ways such variation can be leveraged, for example see[3]: Abraham and Sun [2018], Athey and Imbens [2021], Callaway and Sant'Anna [2020], Cengiz et al. [2019], Deshpande and Li [2019], Strezhnev [2018]. See also this recent paper by Kirill Borusyak, Xavier Jaravel, and Jann Spiess.

References

Sarah Abraham and Liyang Sun. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *arXiv preprint arXiv:1804.05785, 2018.*

Joshua D Angrist and J¨orn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.

Susan Athey and Guido W Imbens. Design-based analysis in difference-in-differences settings with staggered adoption. *Journal of Econometrics*, 2021.

Kirill Borusyak, Xavier Jaravel, and Jann Spiess. Revisiting event study designs: Robust and efficient estimation.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 2020.

Doruk Cengiz, Arindrajit Dube, Attila Lindner, and Ben Zipperer. The effect of minimum wages on low-wage jobs.

The Quarterly Journal of Economics, 134(3):1405–1454, 2019.

Scott Cunningham. *Causal inference: The mixtape*. Yale University Press, 2021.

Manasi Deshpande and Yue Li. Who is screened out? application costs and the targeting of disability programs.

American Economic Journal: Economic Policy, 11(4):213–48, 2019.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. Technical report, National Bureau of Economic Research, 2018.

Anton Strezhnev. Semiparametric weighting estimators for multi-period differencein-differences designs. In *Annual Conference of the American Political Science Association, August*, volume 30, 2018.

[3] A better starting point, rather than jumping straight into papers, may be the blog post or the slides from Andrew Baker, the "Difference in differences" section of Cunningham [2021], or Andrew Goodman-Bacon's slides.

Luke Milsom, DPhil candidate in the Department of Economics, University College, Oxford

10 June 2021