Handout 2. Bias associated with treatment effect heterogeneity in staggered DID studies

The problem

The use of the not-yet-treated units as controls has become quite popular in the program evaluation literature, particularly when there are no never-treated units. For example, in staggered designs, where *all* units (e.g., states or countries) under study implement a certain policy but not all at once, there are no never-treated units. In such cases, by exploiting the differences in the timing of implementation, researchers rely on the not-yet-treated as effective control units. Baker et al. (2022) find that, from 2000 to 2019, top five finance and top five accounting journals published or accepted 751 papers with DID designs. Around 50 percent of these papers rely on staggered designs.

Recent literature warns about a potential bias in staggered designs in the presence of treatment effect heterogeneity (Goodman-Bacon, 2021; Baker et al., 2022; Callaway and Sant'Anna, 2021; Sun and Abraham, 2021; de Chaisemartin and D'Haultfoeuille, 2020). This bias indirectly involves the use of the not-yet-treated units but not when they are used as controls as described in the preceding. Goodman-Bacon (2021) shows that, in staggered designs, the average treatment effect from the canonical DID estimator is in fact the weighted average of all possible two-group/two-period DID estimators. Specifically, for a case where the treated units are divided into early-treated and late-treated, he shows that this general estimator is the weighted average of the DID estimators from its four constituent 2×2 comparisons:

1. Early-treated (as treatment) vs. never-treated (as control);
2. Late-treated (as treatment) vs. never-treated (as control);
3. Early-treated (as treatment) vs. late-treated (as control); and
4. Late-treated (as treatment) vs. early-treated (as control).[1]

The weights are proportional to the size of the groups as well as the treatment variance – i.e., variance of the treatment dummy – in each (two-group) comparison.

The first two comparisons, individually, are typical DID designs in the simplest form. Combined, they form an event study design with only two events. Their weighted average, though an *average* and potentially a problem, which will be discussed shortly, is not the main source of concern that is brought up in this literature. The third comparison is where the not-yet-treated – i.e., the late-treated – are used as controls. Inclusion of this comparison in the weighted average, though does complicate the interpretation,

[1] Comparisons 1, 2, 3, and 4 correspond to Panels a, b, c, and d of Figure 2 in Goodman-Bacon (2021), respectively.

is not the main source of concern either. The fourth comparison essentially uses treated units as controls and as such could potentially become a major source of bias. In some previous studies, this bias has been so large that has reversed the sign of the estimated treatment effect (Baker et al., 2022). In what follows, I describe how and when this bias occurs.

Treatment effect heterogeneity across time (dynamic treatment effects)

Goodman-Bacon (2021) explains, "when already-treated units act as controls, *changes* in their treatment effects over time get subtracted from the [DID] estimate. This negative weighting only arises when treatment effects vary over time, in which case it typically biases regression [DID] estimates away from the sign of the true treatment effect." In other words, the validity of the DID estimator is predicated on the assumption that the control group is not affected by the policy or program within the study period; in this case, the outcome of the control group – i.e., the early-treated units – is affected precisely by the program and within the study period. This bias occurs regardless of whether the change in the treatment effects of the early-treated units over time is positive – i.e., when the policy has a cumulative effect – or negative – i.e., when the effect of the policy fades away over time. This bias is equal to zero only for the special cases in which dynamic treatment effects are absent – i.e., when there are no *changes* in the treatment effects of the early-treated units over time.[2]

Another way to think about this bias is as a failure of the common trends assumption. The common trends assumption dictates that the control and the treatment groups must have parallel trends in the "pre" period. In this case, the pre period is the length of time between when the early-treated received treatment and when the late-treated received treatment. During this period, the two groups must have parallel trends. However, if the outcome of the early-treated group follows a trajectory that is unparallel to that of the late-treated group – as a result of the program and during this period – then the assumption does not hold. This is the cause of the bias associated with any comparison that uses the early-treated as control.

Treatment effect heterogeneity across units

While the bias discussed in the preceding concerns treatment effect heterogeneity across time, the literature also discusses treatment effect heterogeneity across units. The latter could also be a source of potential bias but of a different kind. It is worth mentioning that treatment effect heterogeneity across

[2] The Goodman-Bacon (2021) decomposition makes the source of this type of bias abundantly clear. The first two comparisons correspond to equation (11a) in Goodman-Bacon (2021). The third comparison corresponds to (11b), and the fourth comparison corresponds to (11c) – the only one with an additional bias term to capture the changes in the treatment effects of the already-treated units.

units (or across time) refers to heterogeneity among group that are treated at different times and not heterogeneity among units within a certain group. In the case of heterogeneity across time, the policy or program has a varying effect over time. In the case of heterogeneity across units, groups of units that are treated at different times experience different treatment effects. As elaborated by Sun and Shapiro (2022), for example, the effect of Medicare on health care expenditures may be different in each of the fifty U.S. states. Of course, Medicare was not implemented all at once in all fifty states – i.e., staggered adoption. Nevertheless, the idea here is that, even if Medicare were implemented in all fifty states at the same time, the effect of the policy itself would still differ by state.

When treatment effect heterogeneity across units exists, Goodman-Bacon (2021) claims that the DID estimator "lies along the bias/variance tradeoff: the weights deliver efficiency by potentially moving the point estimate away from, say, the sample ATT." He suggests that the added efficiency may not be worth the bias, "particularly when [the estimator] differs strongly from a given parameter of interest, which occurs when treatment effect heterogeneity is correlated with treatment timing." On the other hand, this type of bias may be small "if there is little variation in treatment timing, if the untreated group is very large, or if some timing groups are very large." This means, this type of bias could be avoided by ensuring that the control group consists of only the never-treated.

Solutions

Goodman-Bacon (2021) decomposition and diagnostic can help measure the extent of the bias due to treatment effect heterogeneity, whether it is across time or across units. However, this approach only applies to balanced panels and does not allow for covariates. There are now several alternative estimators that address treatment effect heterogeneity, including those of Sun and Abraham (2021), Callaway and Sant'Anna (2021), and de Chaisemartin and D'Haultfoeuille (2020). These alternative estimators rely on avoiding comparisons of late- versus early-treated units. As Baker et al. (2022) note, in essence, all three methods are built upon the fundamental notion that treated units should not be compared to "inappropriate controls" (e.g., comparisons of late- versus early-treated firms). These methods have another common denominator: they are based on some type of event-study design – i.e., they break down the treatment effect for cohorts that are treated at different times. Callaway and Sant'Anna (2021), for example, estimate what they refer to as "group-time average treatment effects." In their design, units that are treated in time t and units that are treated in $t+1$ are in separate groups. Using an event-study design that is conceptually similar to that of Callaway and Sant'Anna (2021), Sun and Abraham (2021) also estimate "cohort-specific average treatment effects on the treated," k periods after the first treatment. Not surprisingly, they find that the event-study design solves only treatment effect heterogeneity across units

and not across time. In their suggested design, the control group consists only of the never-treated or the not-yet-treated but the latter are never used as treatment units. Other studies –e.g., Cengiz et al. (2019)– have introduced similar approaches to address the bias associated with treatment effect heterogeneity.

Conclusions

In short, in staggered designs, treatment effect heterogeneity becomes a potential source of bias when there exists a group that takes two different roles: once acts a treatment group and once as a control. The units within the group must obviously receive the treatment at some point – i.e., they are either early-treated or not-yet-treated. Thus, the bias could be avoided all together by having a control group that consists of *only* the never-treated. This is clearly not an option for many studies because sometimes there are no never-treated units. In such cases, treatment effect heterogeneity across time, which seems to be a more critical source of bias given the attention it has received in the literature, could be avoided by not using units that have been treated – i.e., early-treated or already-treated units – as controls. In other words, only the not-yet-treated can be used as controls. An event study design could then help with the bias associated with treatment effect heterogeneity across units that receive treatment at different times.

Implications

Many Government of Canada innovation (and other) programs do follow a staggered treatment adoption pattern. As such, treatment effect heterogeneity could pose as a source of bias, as discussed in the preceding. To avoid this potential bias, one must not include treated firms of any kind in the control group. That is, a firm acts either as a treatment unit or as a control unit, but never both. Also, to the extent that data allows – e.g., when there are sufficient observations in each year – in addition to the canonical DID, one should apply event study designs when treatment adoption is staggered.

References:

Baker, A. C., Larcker, D. F., & Wang, C. C. (2022). How much should we trust staggered difference-in-differences estimates? *Journal of Financial Economics*, 144(2), 370-395.

Callaway, B. and P. H. Sant'Anna. (2021). Difference-in-Differences with multiple time periods. *Journal of Econometrics* 225(2), 200-230.

Cengiz, D., A. Dube, A. Lindner, and B. Zipperer. (2019). The Effect of Minimum Wages on Low-Wage Jobs. *The Quarterly Journal of Economics* 134 (3), 1405-1454.

Handout 2. Bias associated with treatment effect heterogeneity in staggered DID studies

de Chaisemartin, C. and X. D'Haultfoeuille. (2020). Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *American Economic Review* 110 (9), 2964-2996.

Goodman-Bacon, A. (2021). Difference-in-Differences With Variation in Treatment Timing. *Journal of Econometrics* 225(2), 254-277.

Sun, L. and S. Abraham. (2021). Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. *Journal of Econometrics* 225(2), 175-199.

Sun, L. and J. Shapiro. (2022). A Linear Panel Model with Heterogeneous Coefficients and Variation in Exposure. *Journal of Economic Perspectives* 36(4), 193-204.