SO YOU'VE BEEN TOLD TO DO MY DIFFERENCE-IN-DIFFERENCES THING: A GUIDE

Andrew Goodman-Bacon

10/9/2019

Goodman-Bacon (2018) analyzes the two-way fixed effects difference-in-differences (DD) estimator when there is variation in *when* treatment status turns on.[1] Some findings show what comparisons in the data actually give rise to the estimate (whether or not that estimate is "good"):

1. The DD estimator is a weighted average of all 2x2 estimators in the data. A 2x2 DD has two groups (a treatment group whose treatment status changes and a control group whose treatment status does not change) and two time periods (pre-treatment and post-treatment). Every unit is part of the control group in *some* 2x2 DD terms.
2. The weights come from the size of each subgroup (what share of units—states, counties, firms, workers, etc—are in the treatment and control group for a given pair, *and* what share of time periods are used in a given 2x2 subsample), and the variance of treatment (how close to the beginning/end of the subsample window does treatment turn on. The weights depend on the sample.
3. Estimates can change across specifications because the weights change, the 2x2 DD terms change, or both. A Oaxaca-Blinder-Kitagawa decomposition measures how much of the coefficient change comes from each source.
4. Controls can introduce new (almost certainly unintended) identifying variation: comparisons between units with the same treatment variable but different *predicted* treatment based on the covariates.

Some findings show how to interpret DD estimates in theory:

5. Differential trends in counterfactual outcomes in a given timing group generate bias in proportion to how much weight that group gets as a treatment group versus a control group. Groups treated in the middle of the panel matter most, groups treated at the beginning or

[1] Newest version is here:
https://cdn.vanderbilt.edu/vu-my/wp-content/uploads/sites/2318/2019/07/29170757/ddtiming_7_29_2019.pdf

the end can actually act like control groups (a positive trend leads to negative bias). The weights are calculable and can be used in a balance test derived from the estimator.

6. When treatment effects change (monotonically) over time, the DD estimate is biased away from the sign of the true effects.[2] It happens because the treatment effects themselves put units on a differential trend. They are no longer good control units terms that compare "switchers" to "already treated" units are biased.
7. When treatment effects are constant, DD gives a variance-weighted average treatment effect. This is not the same as the average treatment effect on the treated. Groups treated in the middle get more weight than their sample shares imply (because they have high variance), groups treated at the beginning or end of the panel get less weight than their sample shares imply.

A Stata package, `bacondecomp` (Goodman-Bacon, Goldring, and Nichols 2019), calculates the decomposition, makes a scatter plot of the 2x2 DDs against the weights (see Figure 6 in the main paper), and can store the weights and 2x2 DDs for subsequent calculations.

Here are a series of questions and comments that researches have had themselves or have had put to them by editors/referees/seminar audiences, etc. I provide some plain responses and advice, leaving most technical detail to the paper itself.

[2] Other recent papers call the same result "negative weighting" (Abraham and Sun 2018, Borusyak and Jaravel 2017, de Chaisemartin and D'HaultfŒuille 2018, Strezhnev 2018). There are no negative weights in my decomposition theorem. They are based on sample shares and variances, which are always positive. Negative weights refer to the way that a regression DD coefficient—even with identical common trends in counterfactual outcomes—weights together theoretical treatment effect parameters.

1. "Is DD wrong?"

Not in general. The DD research design—comparing outcomes for groups whose treatment status changes to groups whose treatment status does not change—still can be a good idea. The DD *specification*—estimating the coefficient a single post-treatment dummy—is a bad idea when your treatment effects vary over time (get bigger with time since treatment). In this case, just summarize your findings in a different way—event-study or a linear trend-break, for instance.

2. "Should I do an event-study?"

Yes. In many cases, event-study will be right and you can trust a flat pre-period and clear post-treatment changes. This is especially true when you have a large group of untreated units because that puts less weight on the "problematic" 2x2 DDs that use already treated units as controls. Abraham and Sun (2018) analyze the event-study specification itself (no other paper does). They show that even with *no* untreated units, event-study is fine when treatment effects change as long as the pattern of effects is the same for all "treatment cohorts". Event-study can break down when groups treated at different times have treatment effects of different "shapes" (like a different slope change for earlier versus later units, for example).

3. "What should I do if my estimates vary over time?"

Present event-studies and discuss those magnitudes, subject to the caveat above. If you need to impose some restrictions to gain precision, group the event-study coefficients (see Table 2 in Bailey and Goodman-Bacon (2015)) or fit a trend break—$D_{it} \times (t - t_i^*)$--instead of a constant DD coefficient if that is what the event-study results indicate (see Figure 7 and Table 2 in Goodman-Bacon and Schmidt (2019)).[3]

[3] See Roth (2019) on statistical issues with conditioning analytical choices on specification tests like pre-trend tests.

Many recent papers propose alternative estimators all trying to ensure that there are only "good" comparisons between "switchers" and "*not* treated" units (see Baker 2019). All of them compare outcomes for a group whose treatment turns on to outcomes for a group whose treatment will turn on in the future. The challenges are that this limits the follow up window (because the control units get treated eventually) and only works for earlier treated units (because the last group has no future switchers to compare to).

Deshpande and Li (2017) provide one of the clearest examples of this in their "stacked DD" estimates. They discuss how to physically structure your dataset to obtain these kinds of "good" DD comparisons from a single regression. If you do this, you will have a shorter-run DD estimate (or event-study estimate) that you know does not come from comparisons to already treated groups. Callaway and Sant'Anna (2018) provide a flexible reweighting estimator whose benefit is a clear up front discussion of the treatment effect parameter you actually want. Compare your effects using two-way fixed effects to these estimators to be sure you are not focusing on a biased estimate.

4. "Should I do the reweighted balance test?"

Yes. It reflects the actual sources of variation in the estimator, but is not as noisy as another a priori reasonable approach, (just testing joint balance across all timing groups). You always want the most precise balance tests possible. Get the "balance weights" for each timing group using `bacondecomp`: sum the decomposition weight for all terms where group k is the treatment group and subtract the sum of the decomposition weights for all terms where it is the control group. Make a dummy that equals one for untreated units and any timing groups for whom the balance weight is negative. Do a cross-sectional regression of each covariate on that dummy, weighting by the balance weight. The *t*-test is a "reweighted balance test."

5. "What do I learn from the DD decomposition scatter plot?"

This is produced automatically by `bacondecomp`.

a. Which 2x2 DDs matter most? Look at the 2x2 DDs to the right on the x-axis. They get the most weight and you can describe in your test how many terms account for "most" of the weight. It may only be a few. In this case, why not also show (or at least explore) estimates from that specific, simple, DD comparison alone?

b. Why kinds of comparisons matter most (ie. have the most weight)? This is part of the output of `bacondecomp`. Also check how much weight is on each timing group. This may be useful in relation to specific context (ie. one treatment year matters in context).

c. How do the average DDs vary across types of comparisons? Check to see if the comparisons to "already treated" groups are very different than the other comparisons. That average combined with the weight on these terms gives you a measure of how biased the regression coefficient is.[4]

d. You can take out the bias from time-varying effects by hand. Average the 2x2 DDs *except* for those that use already treated units as controls weighting by the decomposition weights. This point estimate will not be biased by time-varying treatment effects (although it will generally be skewed toward short-run effects because the window between different treatment groups is often small so there are more chances to estimate the immediate effects than the longer-run ones).

6. "What does 'within' mean when I do `bacondecomp` with controls?"

[4] You can only see this in a model with no controls (and the ddetail option). This is a limitation of the derivation of the decomposition with controls.

Controls change the nature of the treatment variable from a dummy to a dummy with covariates partialled out. Those covariate evolve differently across units that share the same actual treatment dummy. To some extent (see equation (25)), these differences drive the coefficient estimate you get from the controlled DD regression. The within component tells you weight on this new variation and what the coefficient on it is. To get rid of this part of your estimate, collapse your controls to averages at the timing-group-by-year level. If "within" variation (purely from the controls) matters meaningfully for your estimate, note that in your write up.

7. "Why does the output of `bacondecomp` differ with and without controls?"

With controls I have only derived the DD decomposition down to a "two-group" comparison level (the regression coefficient relating two timing groups across the whole panel), not a "2x2 DD" comparison level (the regression coefficient relating two timing groups only on the subset of periods where one switches). The command cannot, therefore, decompose the estimate into controlled terms at the same level of detail.

8. "What do I learn from comparing specifications?"

Do the Oaxaca-Blinder-Kitagawa decomposition by running `bacondecomp` twice and calculating the pieces (old 2x2 DDs times change in the weights, new weights times change in the 2x2 DDs, change in the 2x2 DDs times change in the weights). Abstracting from bias, if your coefficient changes because of the weights, your regression is giving you a different estimand. It need not mean that your results are sensitive because your design is invalid. If they change because of the 2x2 DDs, this is more consistent with bias.

Scatter the 2x2s and weights (saved in `bacondecomp`'s `stub()` option) from different specifications against each other. This will show you where different analytic choices have the

most bite. The example in the paper shows that weighting matters a lot but only because it increases the influence of California.

9. "How do I do the pre-trend specification?"

The clearest way to implement this approach is if you have many periods before any unit is treated. Suppose the first treatment date is t_1^*. On a sample up to but not including t_1^*, regress Y_{it} on all the fixed effects and control variables and the interaction of timing group dummies with a linear calendar time trend. Construct residuals from this regression for the *whole sample*. Use these as the outcome in your main DD regression.[5] Do not include unit-specific linear time trends (or at the very least, do not only do this).

Conclusion

DD is still a common, clever, research design with an incredibly long history. Keep doing it. New work in this area shows us where estimates come from, how to interpret them, and how to do a better job implementing the same kinds of underlying research designs researchers have used for years.

[5] Taking residuals is the same as subtracting the estimated trends from Y_{it}.

Abraham, Sarah, and Liyang Sun. 2018. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." *Working Paper*.

Bailey, Martha J., and Andrew Goodman-Bacon. 2015. "The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans." *American Economic Review* 105 (3):1067-1104.

Baker, Andrew. 2019. "Difference-in-Differences Methodology." https://andrewcbaker.netlify.com/2019/09/25/difference-in-differences-methodology/.

Borusyak, Kirill, and Xavier Jaravel. 2017. "Revisiting Event Study Designs." *Harvard University Working Paper*.

Callaway, Brantly, and Pedro Sant'Anna. 2018. "Difference-in-Differences With Multiple Time Periods and an Application on the Minimum Wage and Employment." *Working Paper*.

de Chaisemartin, C.;, and X. D'HaultfŒuille. 2018. "Two-way fixed effects estimators with heterogeneous treatment effects." *Working Paper*.

Deshpande, Manasi, and Yue Li. 2017. "Who Is Screened Out? Application Costs and the Targeting of Disability Programs." *National Bureau of Economic Research Working Paper Series* No. 23472. doi: 10.3386/w23472.

Goodman-Bacon, Andrew. 2018. "Difference-in-Differences with Variation in Treatment Timing." *National Bureau of Economic Research Working Paper Series* No. 25018. doi: 10.3386/w25018.

Goodman-Bacon, Andrew, Thomas Goldring, and Austin Nichols. 2019. "bacondecomp: Stata module for Decomposing difference-in-differences estimation with variation in treatment timing." *Stata Command*.

Goodman-Bacon, Andrew, and Lucie Schmidt. 2019. "Federalizing Benefits: The Introduction of Supplemental Security Income and the Size of the Safety Net." *National Bureau of Economic Research Working Paper Series* No. 25962. doi: 10.3386/w25962.

Roth, Jonathan. 2019. "Pre-test with Caution: Event-study Estimates After Testing for Parallel Trends." *Working Paper*.

Strezhnev, Anton. 2018. "Semiparametric Weighting Estimators for Multi-Period Difference-in-Differences Designs." *Working Paper*.