Callaway and Sant'anna DD estimator

a story of differential timing and heterogeneity

SCOTT CUNNINGHAM

MAR 08, 2021 · PAID

49 29 3 Share

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

One of the more exciting papers in econometrics over the last year that I have had the pleasure to read is Callaway and Sant'Anna's forthcoming article in the Journal of Econometrics, "Difference-in-differences with multiple time periods". It is exciting in part because once you have taken the time to carefully read Andrew Goodman-Bacon's piece on the biases of twoway fixed effects (TWFE), you will be desperate for something that could possibly replace that OLS estimator. Had these papers not come out so closely in time, it's possible I might simply have not known the best path forward.

So what is this paper about? It is about a model that when used in situations that are the most common situations researchers face when studying policy, you can estimate the average treated on the treated (ATT) and all its descendent parameters, some which you've never even thought about before. Let's start at the top.

History of thought: Abadie (2005)

I always find it helpful for my mind to find the parent and grandparents of a paper in econometrics to the best I can. The parent of this paper is a 2005 article in the Review of Economic Studies by Alberto Abadie at MIT entitled "Semiparametric Difference-in-differences Estimators". In that paper, Abadie examines a situation

© 2024 scott cunningham · Privacy · Terms · Collection notice
Substack is the home for great culture

> "A good way to do econometrics is to look for good natural experiments and use statistical methods that can tidy up the confounding factors that nature has not controlled for."

This is the heart of Abadie's semiparametric paper — use natural experiments, and fix up the confounders. You used Abadie's estimator when randomization was not possible and treatment was selective on observable covariates. Abadie's paper proposed a method that estimates the ATT, and one of the assumptions it needs is what he calls "conditional parallel trends", which is to say, that the parallel trends hold for a treatment and comparison group only after conditioning on some matrix X. We may also think this situation arises because we think covariates are imbalanced between the treatment and comparison groups because they model the conditional selection into treatment itself. And so to address this, they absorb all of the X information into a single scalar — the propensity score. The second assumption, therefore, is the common support assumption which states that units from the treatment and control both are present along the propensity score's distribution.

It's a straightforward estimator in many ways and with only three steps. The first step is to compute each unit's mean value before and after a treatment. The second is to estimate a propensity score to weight each of those units. And the third is compare the weighted changes in the treatment group to that of the comparison group. The inference he proposes will take into account that the second step estimated the propensity score. And the estimator itself, which estimates the ATT, is:

$$E\left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{d_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}\right]$$

There are a few pieces to this that I'd like you to see before we move on to Callaway and Sant'Anna. First, notice how the numerator in the first term is a difference between any period (t) and a "base" (b) period. I will call this the "long difference" because Callaway and Sant'Anna use that nomenclature in their paper. It is simply the before and after difference that we first take when calculating any DD statistic.

The issues with modeling selection into treatment based on X covariates is captured in the propensity score which often has the benefit of reducing dimensionality and

therefore avoiding the curse of dimensionality. But here we also exploit another benefit of the propensity score — it can be used as an inverse probability weight on the long difference. When this is done, we get the second difference in the DD design.

The reason you see a b subscript on the X term is because unlike many regression based estimation tools which accommodate time varying X variables, Abadie (2005) will require you use X to estimate a propensity score into treatment itself. And that requires using the baseline value of covariates to predict the conditional probability of treatment. This is a unique way of conceiving of covariates in some respects — it's almost like a conditional independence assumption, but without the implicit invoking of randomization.

You will as with all propensity score applications need to pursue diagnostics, such as ensuring common support. But, once you have done so, each unit's long difference will be weighted either by a 1 if the treatment group or a propensity score ratio multiplied by -1. Imagine a large column of these sums averaged, and you will have imagined an unbiased estimator of the ATT.

Bacon decomposition

When is Abadie (2005) used though, and why do you not hear about it very often? It's used primarily with longitudinal panel data like Lalonde's 1986 NSW job trainings program. I say that one because in that dataset, a group of people were treated at some time period, and a group of other people were not treated at that time period. The treatment, in other words, did not "roll out" hitting groups at different points in time. Rather, it hit all of them at the same time. This is what Goodman-Bacon (2019) has called the simple 2x2 DD — it involves a group treated, and an untreated group for comparison, and treatment occurs at only one time period.

But this is not the most common situation when people like economists use DD for some research application. The more common situation is when units select into treatment at different points in time, usually because of evolving state legislation across the country. Think of the minimum wage for instance. It is not merely the federal minimum wage hikes that determine the minimum wages that firms can pay their workers. Most of the variation in the minimum wage comes from the autonomous decision making of state and local policymakers. Because the treatment rolls out across the country hitting states at different points in time, Abadie (2005)

and its Stata package -absdid- are actually not appropriate, because as you saw in that equation above, there was no term describing any particular group's unique treatment date. So what will we do?

Goodman-Bacon (2019) is a powerfully helpful article for revealing that the twoway fixed effects estimator (panel fixed effects with year dummies) is biased — potentially quite biased — when there is differential timing (i.e., the rollout I've been saying) and heterogenous treatment effects over time. His paper isolates the many different ways in which twoway fixed effects shapes the estimates, often for no good reason theoretically. You learn for instance that you must satisfy a complex variance weighted parallel trends assumption for each group. You learn that the variance of treatment itself is a weight on each group's respective 2x2. And you learn that the variance of treatment will oddly enough be maximized at 0.25 only when the share of time that a group spends in the treatment group is 1/2 the time. Given we are talking about panel data, that means groups at the middle of the panel will be weighted higher than those at either end, ceteris paribus.

Goodman-Bacon (2019) goes further than that though. He shows that even with this complex variance weighted parallel trends assumption equaling zero, you can still be in hot water if you estimate the ATT with twoway fixed effects. That's because in the probability limit, as the number of units grows to infinity, the TWFE estimator is equal to the sum of the following three terms:

$$\widehat{\delta}_{twfe} = VWATT + VWPT - \Delta ATT$$

Now, at best, TWFE will identify the VWATT, which in my mind is not the ATT and therefore since it's only unique to the panel length you chose, it is likely not something you can simply pass over to your policymakers or employers and say "We've done it. We know what this policy will do." That is because the VWATT is not really a policy parameter. Rather, it's a policy parameter shaped by the limitations of the estimator caused by choosing a particular panel length, and it is unclear whether that really is or is not helpful in the times of hard decisions and scarce resources.

But even then, if treatment effects do not evolve over time, TWFE will only identify VWATT if the variance weighted parallel trends assumptions hold. But what if they evolve over time? What if long run treatment effects are more pronounced than short ones? What if some of the people graduated during a recession pandemic and others

didn't? Wouldn't we expect that the treatment effects may vary by cohort or even vary over time? We can't and shouldn't rule it out just because we feel comfortable using a tool like TWFE. That's the very definition of gullibility. Gullible econometrics is mostly harmful, not harmless, when inferring causality from non experimental data.

Callaway and Sant'Anna (2020)

Entering stage left is Brant Callaway at the University of Georgia and his former teacher and coauthor Pedro Sant'Anna at Vanderbilt. They were colleagues and students of Andrew Goodman-Bacon, and I imagine sometimes the three of them talking about their respective projects. Perhaps each one was listening closely to subtle things the other was saying. Elsewhere in the world, several others were working on this problem of differential timing, such as Kirill Borusyak and the French team of Chaisemartin and D'HaultfEuille. So there was definitely something in the air because spread over the globe, many econometricians were digging deeper into the methodology of difference-in-differences under heterogenous treatment effects and differential timing.

I will discuss Callaway and Sant'Anna (CS, 2020) because it is the one I know best and for this inaugural entry, I thought I should do something I love, and I love one of the authors, Pedro, very much, and I love this paper, too, very much. The paper is an econometrics paper in the sense that its argument is laid out in three parts: identification, estimation and inference. So let's start at the top of each of them.

The key concept in CS is the group-time ATT. The group-time ATT is a unique ATT for a cohort of units treated at the same point in time. So if Florida and Arkansas both pass legislation in 2005, then we call them the 2005 group, or cohort. If ten more states are selected in 2006, we call them the 2006 group. And so forth. The group-time ATT can also be thought of as a dynamic term. Maybe we want the 2005 group's ATT in 2006, or maybe we want in 2007, or maybe even further out. For each group there are T-gt ATT parameters (where T is the last date of the panel and gt is the treatment date for thet group). Over many groups, the sheer number of group-time ATTs that we can conceive of is the sum of all those T-gt ATTs for every group. So even though I find the group-time ATT parameter very interesting and intuitive, it also immediately causes me to realize that I am about to be confronted with a number of parameter estimates that far exceed the one that I usually focus on in any paper, often labeled nothing more than "DD estimate". CS will handle this though by

both identifying the assumptions needed for a consistent estimate of the group-time ATT, and then provide a simple way to aggregate all of them into fewer and simpler parameters. But I get ahead of myself. Let's dig a little deeper first.

First, as we did with Abadie (2005), we must estimate the propensity score. But because we have multiple treatment dates for multiple groups, there is a unique propensity score for every group. As with Abadie's estimator, though, we will be using pre-treatment (or baseline) covariates in our propensity score calculation, not time varying ones.

Whenever you estimate a conditional probability of some binary event occurring, you are always using at least two types of units for this estimation: you are using units in the treatment group, and you are using units in the comparison group. So this means you always have to compare the treatment group *to someone* in a different group. In Abadie's setup, that's the untreated group, because recall, we only had treatment occurring at one point in time. And so it was naturally the case that the comparison group would simply be that large reservoir of untreated units.

But we do not have the luxury of a large reservoir of untreated units necessarily in many applications with multiple time periods and differential timing. Who will be the comparison group for each group? Well we know who it *won't* be. You will never use as your comparison group in the CS estimator those groups who had previously been treated. Just as Goodman-Bacon (2019) showed us, it is a sin to use an already treated group as a comparison group. The dynamics of their treatment can curdle the milk and so we avoid it at all cost.

So who does CS use as its comparison when calculating a propensity score that absorbs all of the information in X to create implicit pairings of units in the treatment and comparison groups? There are two options: we may use a pool of units as our comparison group who never are treated during the duration of the panel. Or we may use a pool of units who have simply *not yet* been treated by the time of treatment.

Consider four groups: Group A is treated in 2005; Group B is treated in 2007; Group C in 2009; and Group D in 2010. We can calculate using CS group-time ATTs for A, B and C, but not for D. And the reason we cannot calculate the group-time ATT for Group D is because by the time we reach 2010, there are no longer in our data any untreated groups. So, depending on your application, you may either be estimating

group-time ATT based on a pool of never-treated units, or you may be estimating a subset of all possible group-time ATTs (excluding the last group who can only in the end function as a comparison group for the earlier ones).

Let's look first at the parameter of interest. This was the same parameter that Abadie was focused on, only that his application only had one treatment group. Now we have several. The parameter is expressed as:

$$ATT(g,t) = E[Y_t^1 - Y_t^0 | G_g = 1]$$

Now we must identify this parameter, but how? Well before we get to the how, let's think about the assumptions needed in the first place. There are four assumptions that must be met if we are to use CS. They are 1) that the sampling of the data is panel or repeated cross-sectional data; 2) that only conditional on X will we be able to maintain parallel trends; 3) the treatment must only turn on; it cannot turn on and then off again. They call this irreversible treatment. And 4) for some range, it must be the case that the treatment and comparison groups have units with the same approximate propensity score. This fourth and last assumption is the common support assumption that we discussed earlier with Abadie (2005).

With these assumptions, CS managed to discover and invent an estimator that when used could yield an unbiased and consistent estimate of each group's individual group-time ATT. The equation for it is here:

$$ATT(g,t) = E\left[\left(\frac{G_g}{E[G_g]} - \frac{\frac{\hat{p}(X)C}{1-\hat{p}(X)}}{E\left[\frac{\hat{p}(X)C}{1-\hat{p}(X)}\right]}\right)(Y_t - Y_{g-1})\right]$$

Notice how similar it is to Abadie's 2005 estimator, which I reproduce again here.

$$E\left[\frac{Y_t - Y_b}{Pr(D_t = 1)} \times \frac{d_t - Pr(D = 1|X_b)}{1 - Pr(D = 1|X_b)}\right]$$

Notice how each has a "long difference", for instance. Notice how each has this negative weight on the propensity score ratio. But beyond that it gets a little harder to see exactly how they are similar. And that is because the Abadie estimator is

estimating a single ATT for one group — the only group that is treated. But CS is estimating several ATTs. Specifically, an ATT for a group at any point in time we want to consider it. And that equation above Abadie's, two equations back, does it.

In that equation, though, we have new terms. For instance, we have Gg. Gg is a dummy variable equalling one if the unit is in group g, where g is the year of the units' common treatment date. We have an average of Gg in the first term's denominator. So if there are 5 states treated out of 10, then E[Gg] = 0.5.

And C is a dummy equalling one if it's in the comparison group. Notice how when Gg=1, C=0 and vice versa. Also notice that you are taking the long difference using only the base year as the before period. You don't even use, in other words, the lengthy number of years prior to treatment to calculate the group-time ATTs. You will for an event study, but you won't for the treatment effect itself.

After this, it's all smooth sailing. They propose a kind of bootstrapping procedure, for instance, which will conduct asymptotically valid inference that can adjust for autocorrelation and clustering. And they propose aggregating these individual ATTs by group and time into fewer, but still interpretable, parameters.

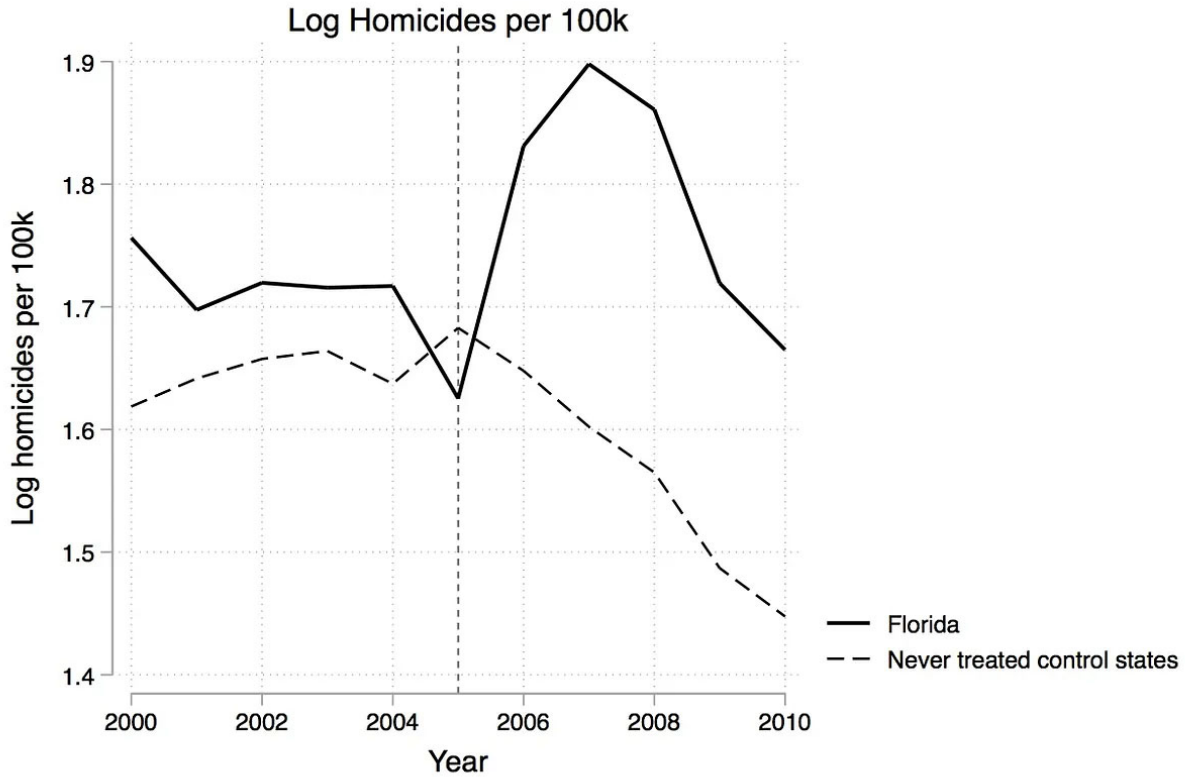
Cheng and Hoekstra's castle doctrine study

I'd like to now take the time to discuss an application of the CS model. This code was created by my coauthor, Yunie Le, an extraordinarily talented PhD student at Claremont Graduate University with whom I have been writing a paper that uses CS. Yunie graciously helped me with the R code for this example, as R is not even my second, or third language. It's my feral language which I can only speak in guttural noises and grunts.

But before we begin, let's discuss the background of the data we will be using. The data comes from a 2013 article by Cheng Cheng and Mark Hoekstra entitled "Does Strengthening Self Defense Law Deter Crime or Escalate Violence?" The paper concerns a series of legislation that passed from 2005 to 2009 in the United States. The legislation was called "castle doctrine" reforms, but in Florida it was called the Stand Your Ground law. These laws removed principles from the English common law that had dominated American life for centuries concerning self-defense. Historically, if a person was in danger, they had a responsibility to retreat, unless they were at home, in which case they retreat to the wall of their home and no further.

Then and only then were they allowed to defend themselves with lethal force. These castle doctrine laws — called "castle" because the home is one's castle — did a few things. First, they allowed a person to use lethal force in places other than the home. Second, they removed any civil liability that stemmed from killing an opponent. And third, they gave the presumption of reasonable fear to the shooter. When you combine all three, you get a significant decline in the opportunity cost of killing another human being by allowing it to happen in places previously prohibited and by reducing the criminal and civil liabilities associated with killing someone.

The paper is a traditional DD paper in that it uses TWFE in a DD design to estimate a single parameter which we now know is the VWATT. The problem is there's some reason to think that the treatment effects evolve over time. Consider this picture of Florida when it passes Stand Your Ground.



There are good reasons here to think that Florida would've had parallel trends. For instance, the trends before (excluding that big drop in 2003) seem to have a similar slope. But notice how homicides grow from 2005 to 2006 and then again into 2007. This is suggestive of dynamic treatment effects, but we don't want to overstate either because technically we do not observe the true ATT. Even with variance weighted parallel trends equalling zero, TWFE will still be biased if there are dynamic

treatment effects with differential timing. So we use CS instead.

Now CS in this situation is going to estimate a lot of group-time ATTs. There's five groups for instance each of whom gets treated in 2005, 2006, 2007, 2008 and 2009. The panel ends in 2010, so that's 5 ATT(g,t) for the first group, 4 for the second, 3 for the third, 2 for the fourth and 1 for the last — a total of 15 group-time ATTs. Let's look at the code, and then the output one by one.

```
library(readstata13)
library(ggplot2)
library(did) # Callaway & Sant'Anna

castle <- data.frame(read.dta13('https://github.com/scunning1975/mixtape/raw/master/castle.dta'))
castle$effyear[is.na(castle$effyear)] <- 0 # untreated units have effective year of 0

# Estimating the effect on log(homicide)
atts <- att_gt(yname = "l_homicide", # LHS variable
tname = "year", # time variable
idname = "sid", # id variable
gname = "effyear", # first treatment period variable
data = castle, # data
xformla = NULL, # no covariates
#xformla = ~ l_police, # with covariates
est_method = "dr", # "dr" is doubly robust. "ipw" is inverse probability weighting. "reg" is regression
control_group = "nevertreated", # set the comparison group which is either "nevertreated" or "notyettreated"
bstrap = TRUE, # if TRUE compute bootstrapped SE
biters = 1000, # number of bootstrap iterations
print_details = FALSE, # if TRUE, print detailed results
clustervars = "sid", # cluster level
panel = TRUE) # whether the data is panel or repeated cross-sectional

# Aggregate ATT
```

```
agg_effects <- aggte(atts, type = "group")
summary(agg_effects)

# Group-time ATTs
summary(atts)

# Plot group-time ATTs
ggdid(atts)

# Event-study
agg_effects_es <- aggte(atts, type = "dynamic")
summary(agg_effects_es)

# Plot event-study coefficients
ggdid(agg_effects_es)
```

castle_cs.R hosted with ❤ by GitHub view raw

If you want, you can also download it here from my GitHub. But let me tell you what it does first. The name of the R package is called **did**. It estimates the CS estimator using a number of options including the double robust method (by Sant'Anna and Zhao), the inverse probability weighting method (see above), and regression. So look — even here with CS, you can use regression, because ultimately you are merely calculating weighted differences in means. You can also choose whether you will use the never treated or the not yet treated by switching out words in line 15. You can choose the number of bootstraps, and you can choose the level of clustering, as well as whether it's panel or repeated cross-sections. It's everything a growing boy needs. Let's look at some of the output that we get once we run it. I'll reproduce now the code that will produce interesting statistics for us.

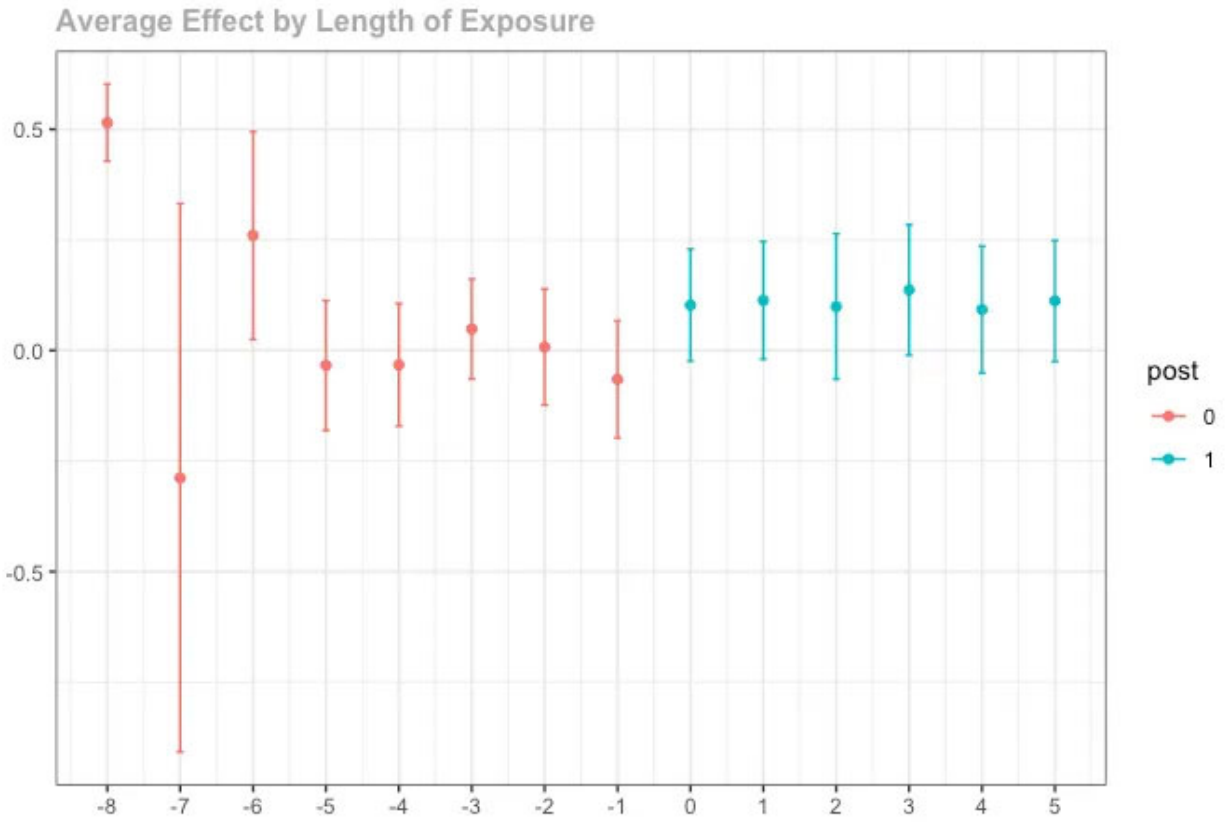
Table 1: Aggregated group-time ATTs

Group	ATT	SE	95% Confidence bands	
All groups	0.1075**	0.0358	0.0373	0.1778
2005	0.0952**	0.0321	0.0221	0.1682
2006	0.1074	0.0540	-0.0155	0.2304
2007	0.1332**	0.0559	0.0058	0.2606
2008	0.1181	0.0571	-0.0120	0.2483

2009	-0.0028	0.0390	-0.0916	0.0860

In the top row, I've labeled it "All groups" because it's the aggregated ATT(g,t) for all groups and all time periods. This is one of the aggregations that CS proposes. As you can see, given the outcome is logged homicides per 100,000, these laws may have caused as much as a 10.8% increase in homicides. The next rows below the horizontal bar show the average ATT per group. So the ATT for Florida (2005) is 0.0952, but the average for the others is growing — up until 2007, and then falls. The biggest effects appear to come from the passage of the laws in 2007 which has a coefficient of 13.3% increase in homicides.

Interestingly, when we run this in TWFE, we find a slightly smaller estimate of between 8-10%. This is interesting in part because the bias created by dynamic treatment effects always pushes the parameter estimate closer to zero away from the true parameter. And that means, not only is TWFE biased, but it's biased towards zero. Yes, it can flip signs, but more than likely it will merely be an effect that is "too small". Well, we see that here with CS. With CS, we get bigger effects than we got with TWFE. Again, this is the logic of what Goodman-Bacon (2019) was telling us.



Next let's look at a simple event study plot. All of the data has been reframed in relative event time so that we can see the values of an ATT(g,t) parameter expressed there. Notice how for five years prior, there are essentially no differences between treatment and control group units. The earlier imbalances are likely caused by the low number of units with leads that far back. But, starting immediately in the year of treatment, and each year after, log homicides leveled up.

Perhaps it is because the group-time ATTs expressed in relative time are so flat that the result is so similar to what one finds with what I found using TWFE in my mixtape here. Without dynamics in the ATT over time, TWFE is only biased by variance weighted parallel trends assumption, and while the variance weighted part doesn't migrate over to CS, it will likely be the case that if one breaks down, so will the other.

Concluding remarks

The most popular quasi-experimental design is the difference-in-differences. Each year, there are thousands of new papers around the world that use this methodology to estimate the causal effect of some treatment. They almost all will be exploiting some differential timing thinking it affords them advantages when in fact if they estimate with TWFE, it is the exact opposite.

But by the grace of God, econometricians have been intently at work on this problem and have developed some nice solutions so that the rest of us can keep our day jobs. One of my favorites is the Callaway and Sant'Anna estimator, which is in an R package called **did**, but unfortunately not in Stata. I encourage you to study and ultimately use the above code to do this yourselves in your own work. We must all be vigilant against the urge to use an estimator like TWFE for no other reason than that we know how to type it in to Stata.

49 Likes · 3 Restacks

← Previous Next →

29 Comments

Write a comment...

Vanessa Santos Apr 27, 2021

this is amazing. Thanks so much for walking us through this!

LIKE REPLY SHARE

John Lump May 4, 2021

Thanks so much for writing this up! Is there any work/guidance on what (if any) of this recent TWFE work applies to settings in which treatments are *reversible*? I.e. settings in which people enter into a treated state that they then exit again. I assume the variance weighting structure still applies, but I'm not sure what else would be at work

LIKE REPLY SHARE

3 replies by scott cunningham and others

27 more comments...