Canonical Research Designs I: Difference-in-Differences

Paul Goldsmith-Pinkham

March 18, 2021

Revisiting Research Design

- Recall my attempt at a definition:
  - A *(causal) research design* is a statistical and/or economic statement of how an empirical research paper will estimate a relationship between two (or more) variables that is causal in nature – X causing Y.
  - The design should have a description for how some variation in X is either caused by or approximated by a randomized experiment.

Revisiting Research Design

- Recall my attempt at a definition:
  - A *(causal) research design* is a statistical and/or economic statement of how an empirical research paper will estimate a relationship between two (or more) variables that is causal in nature – X causing Y.
  - The design should have a description for how some variation in X is either caused by or approximated by a randomized experiment.

- Dinardo and Lee (2011) have a famous handbook chapter entitled "Program Evaluation and Research Design" where they make a distinction between two types of research designs

and "design-based" *statistical conditions*. When we have some institutional knowledge about the process by which treatment was assigned, and when there can be common agreement about how to represent that knowledge as a statistical statement, we will label that a "D"-condition; "D" for "data-based", "design-driven", or "descriptive". These conditions are better thought of as *descriptions* of what actually generated the data, rather than *assumptions*. By contrast, when important features of the data generating process are unknown, we will have to invoke some conjectures about behavior (perhaps motivated by a particular economic model), or other aspects about the environment. When we do not literally know if the conditions actually hold, but nevertheless need them to make inferences, we will label them "S"-conditions; "S" for "structural", "subjective", or "speculative". As we shall see, inference about program effects will frequently involve a combination of "D" and "S" conditions: it is useful to be able to distinguish between conditions whose validity is secure and those conditions whose validity is not secure.

Note that although we may not know whether "S"-conditions are literally true, sometimes they will generate strong testable implications, and sometimes they will not. And even if there is a strong link between what we know about program assignment and a "D"-condition, a skeptic may prefer to treat those conditions as hypotheses; so we will also consider the testable implications that various "D"-conditions generate.

Revisiting Research Design

- D-condition designs fall clearly into the “PGP” description of a research design – knowledge of the DGP leads to a variation in the data generating our identification

- S-conditions fall less clearly into this context (as we will discuss).
  - The relationship between X and Y can be clearly articulated, but how it is potentially approximated by a random experiment is less obvious

- This issue will become clear as we discuss our first topic

Estimating causal effects in real settings

- In many applications, we want to estimate the effect of a policy across groups
- However, the policy assignment is *not* necessarily uncorrelated with group characteristics
- How can we identify the effect of the policy without being confounded by these level differences?

Estimating causal effects in real settings

- In many applications, we want to estimate the effect of a policy across groups
- However, the policy assignment is *not* necessarily uncorrelated with group characteristics
- How can we identify the effect of the policy without being confounded by these level differences?

Difference-in-differences!
(DinD)

First, a warning

- This literature has had a certain amount of upheaval over the past 5-6 years

- Tension: provide context for how people currently and historically have studied diff-in-diff
  - But also elaborate on concerns identified in recent papers

- The key issues boil down into two questions:
  1. *What is the counterfactual estimand?*
     - Does your estimator map to your estimand? (e.g. “Are you getting at what you meant to?”)
  2. *What are your structural assumptions and their implications?*
     - Do you need to assume functional forms? (e.g. “Is this really something that has an experimental analog”?)

- Papers have both pointed out issues but also provided solutions to almost all of the problems that they’ve raised, so not something that should prevent you from using these tools

Basic setup

- Assume we have n units (i) and T time periods (t)
- Consider a binary policy D_{it}, and we are interested in estimating its effect on outcomes Y_{it}
- The inherent problem is that D_{it} is *not* necessarily randomly assigned
- The historical key (and parametric) assumption underlying of the potential outcomes model (one version):

$$Y_{it}(D_{it}) = \alpha_i + \gamma_t + \tau_i D_{it}$$
$$\text{s.t. } Y_{it}(1) - Y_{it}(0) = \tau_i$$

- Implication? In the absence of the treatment, the Y_{it} across units **evolve in parallel** – their γ_t are identical. Absent the policy, units may have different *levels* (α_i) but their changes would evolve in parallel
  - This is a key (parametric!) identifying assumption
  - $Y_{it}(0) - Y_{i,t-k}(0) = \gamma_t - \gamma_{t-k}$, $Y_{it}(0) - Y_{jt}(0) = \alpha_i - \alpha_j$

Basic 2x2 DinD setup

- Recall our typical estimand of interest is the ATE or the ATT:

$$\tau_{ATE} = E(Y_{it}(1) - Y_{it}(0)) = E(\tau_i)$$
$$\tau_{ATT} = E(Y_{it}(1) - Y_{it}(0)) = E(\tau_i | D_{it} = 1)$$

- Since D is not randomly assigned and we only observe one time period, this model is inherently not identified without additional assumptions.
  - Why? D_i could be correlated with α_i
  - Recall that our plug-in estimator approaches need estimates for $E(Y_{it}(1))$ and $E(Y_{it}(0))$
  - Where can we get unbiased estimates?

- With two time periods we can make a lot more progress!

Basic 2x2 DinD setup

- Recall our typical estimand of interest is the ATE or the ATT:

$$\tau_{ATE} = E(Y_{it}(1) - Y_{it}(0)) = E(\tau_i)$$
$$\tau_{ATT} = E(Y_{it}(1) - Y_{it}(0)) = E(\tau_i | D_{it} = 1)$$

- Since D is not randomly assigned and we only observe one time period, this model is inherently not identified without additional assumptions.
  - Why? D_i could be correlated with α_i
  - Recall that our plug-in estimator approaches need estimates for $E(Y_{it}(1))$ and $E(Y_{it}(0))$
  - Where can we get unbiased estimates?

- With two time periods we can make a lot more progress!

2×2 DinD estimation

	t =0	t = 1
$D = 0$	$\gamma_0 + \alpha_i$	$\gamma_1 + \alpha_i$
$D = 1$	$\gamma_0 + \alpha_i + \tau_i$	$\gamma_1 + \alpha_i + \tau_i$

- Now consider the within unit difference:

$$Y_{i1} - Y_{i0} = (\gamma_1 - \gamma_0) + \tau_i(D_{i1} - D_{i0})$$

- Hence

$$E(Y_{i1} - Y_{i0}|D_{i1} - D_{i0} = 1) - E(Y_{i1} - Y_{i0}|D_{i1} - D_{i0} = 0) = E(\tau_i|D_{i1} - D_{i0} = 1)$$

- Wait, you say, that's a lot more notation than I was expecting.
  - Simplifying assumption: treatment only goes one way in period 1
  - "absorbing adoption", e.g. $D_{i0} = 0$

$$E(Y_{i1} - Y_{i0}|D_{i1} = 1) - E(Y_{i1} - Y_{i0}|D_{i1} = 0) = \underbrace{E(\tau_i|D_{i1} = 1)}_{ATT}$$

An aside on our simplifying assumption

- The choice of focusing on take-up of a policy, such that $D_{i1} \geq D_{i0}$, is well-grounded in many policy settings
- However, there are cases where policies turn on, and then turn off, and this can vary across units
- This can be challenging and potentially problematic with heterogeneous effects
- Need to think carefully about whether D_i turning on is identical (but opposite sign) to D_i turning off
  - Hull (2018) working paper on mover designs discusses this
- For today, will ignore this issue

Estimation using linear regression

- A simple linear regression will identify $E(\tau_i | D_{i1} = 1)$ with two time periods:

$$Y_{it} = \alpha_i + \gamma_t + D_{it}\beta + \epsilon_{it} \tag{1}$$

- This setup is sometimes referred to as the Two-way Fixed Effects estimator (TWFE)

- Note: we could have also estimated τ directly:

$$\hat{\tau} = n^{-1} \sum_i \underbrace{D_i(Y_{i1} - Y_{i0})}_{\Delta\overline{Y}_1} - \underbrace{(1 - D_{i1})(Y_{i1} - Y_{i0})}_{\Delta\overline{Y}_0}$$

  - Intuitively, we generate a counterfactual for the treatment using the changes in the untreated units: $E(Y_{i1} - Y_{i0} | D_i = 0)$

- Necessary: two time periods! What if we have more?

Multiple time periods in basic setup

- Let's consider a policy that occurs all at t_0 (e.g. single timing rolled out to treated units)

- More time periods helps in several ways:
  1. If we have multiple periods *before* the policy implementation, we can partially test the underlying assumptions
     - Sometimes referred to as "pre-trends"
  2. If we have multiple periods *after* the policy implementation, we can examine the timing of the effect
     - Is it an immediate effect? Does it die off? Is it persistent?
     - If you pool all time periods together into one "post" variable, this estimates the average effect. If sample is not balanced, can have unintended effects!

- How do we implement this?

$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=1, t \neq t_0}^{T} \delta_t D_{it} + \epsilon_{it},$$

  - One of the coefficients is fundamentally unidentified because of α_i
  - All coefficients measure the effect *relative* to period t_0.

Pre-testing and structural assumptions

- Note that for the above model, we made a stronger assumption about trends
  - The Dinardo and Lee "S-assumptions" start to bite
  - We assumed that $Y_{it}(d) - Y_{i,t-k}(d) = \gamma_t - \gamma_{t-k}$ for all k and d
  - This is testable pre-treatment (hence the pre-test)

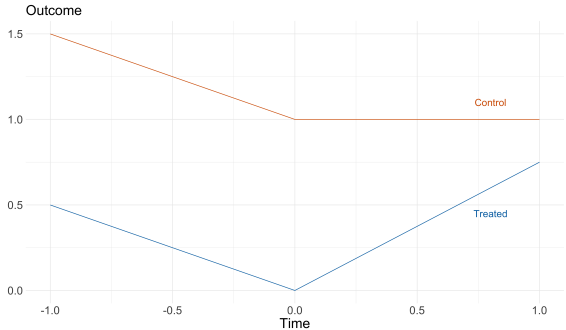
- This is very powerful and has helped spark the growth in DinD regressions
  - Visual demonstrate of "pre-trends" helps support the validity of the design
  - Worth doing!

- Two key issues:
  1. Pre-testing can cause statistical problems
  2. What does parallel trends even mean?

Pre-testing issues (Roth 2020)

- Consider $T = 3$ and think about what a pre-trend test is trying to do

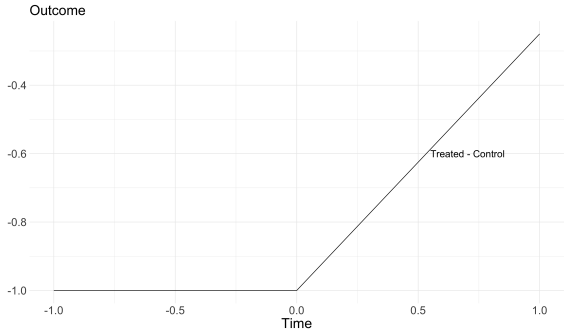
- Testing whether the difference relative to $t = 0$ for $t = -1$ is significant



Pre-testing issues (Roth 2020)

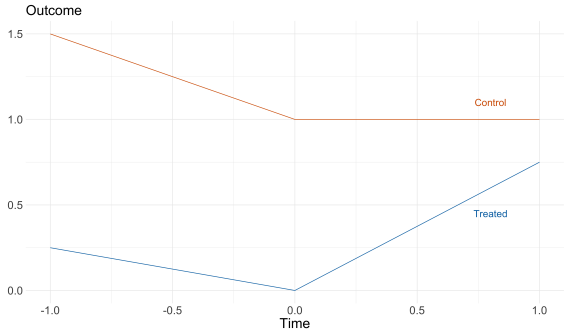
- Consider $T = 3$ and think about what a pre-trend test is trying to do

- Testing whether the difference relative to $t = 0$ for $t = -1$ is significant



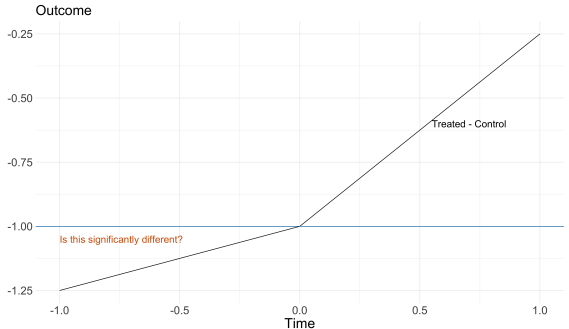
Pre-testing issues (Roth 2020)

- Consider $T = 3$ and think about what a pre-trend test is trying to do
- Testing whether the difference relative to $t = 0$ for $t = -1$ is significant
- Unconditionally, this is reasonable. However, Roth (2020) highlights that this is a form of *pre-testing*, and that low power in detecting pre-trends can be problematic



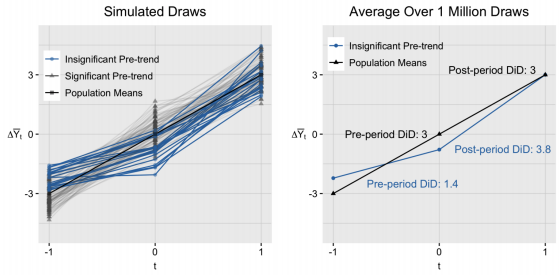
Pre-testing issues (Roth 2020)

- Consider $T = 3$ and think about what a pre-trend test is trying to do
- Testing whether the difference relative to $t = 0$ for $t = -1$ is significant
- Unconditionally, this is reasonable. However, Roth (2020) highlights that this is a form of *pre-testing*, and that low power in detecting pre-trends can be problematic



Pre-testing issues (Roth 2020)

- Consider $T = 3$ and think about what a pre-trend test is trying to do
- Testing whether the difference relative to $t = 0$ for $t = -1$ is significant
- Unconditionally, this is reasonable. However, Roth (2020) highlights that this is a form of *pre-testing*, and that low power in detecting pre-trends can be problematic
- By selecting on pre-trends that "pass", will tend to choose baseline realizations that satisfy pre-trends, but induce *bias* in the effect



How to interpret this caution?

- First, don't panic. Examining pre-trend is still important diagnostic
- Important to realize that selecting your design based on pre-trend is *constructing* your counterfactual
  - Pre-tests will cause you to potentially contaminate your design
- Suggested solution from Roth (2020): incorporate robustness to pre-trends into your analysis. Rambachan and Roth (2020) present results on testing sensitivity of DinD results to pre-trends
  - Brief intuition follows

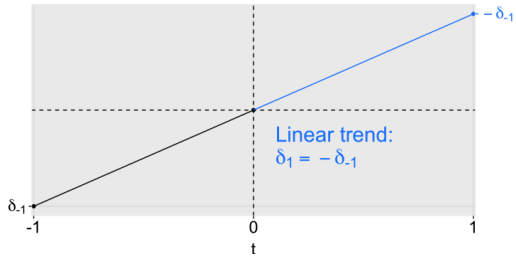
Rambachan and Roth (2020) suggestion

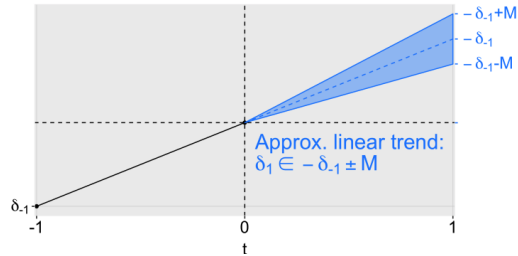
- Intuitive proposed solution for robustness. Note the post and pre effects:

$$\mathbb{E}\left[\hat{\beta}_1\right] = \tau_{ATT} + \underbrace{\mathbb{E}\left[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,1}(0) - Y_{i,0}(0) \mid D_i = 0\right]}_{\text{Post-period differential trend} =: \delta_1},$$

$$\mathbb{E}\left[\hat{\beta}_{-1}\right] = \underbrace{\mathbb{E}\left[Y_{i,-1}(0) - Y_{i,0}(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_{i,-1}(0) - Y_{i,0}(0) \mid D_i = 0\right]}_{\text{Pre-period differential trend} =: \delta_{-1}}.$$

- parallel trends assumes these δ are zero. But pre-trends may not be zero.
  - R&R say: we can use the info from our pre-trends to bound post-trend
  - Use a *smoothness* assumption, *M*, on the second derivative. E.g. simple case:





This approach adds more work but also more validity

- Need to select M, and will likely have less strong results
- However, very powerful way to address concerns about pre-trends
- Code for applying this technique is availabe in R:
  `https://github.com/asheshrambachan/HonestDiD`

Parallel trends in what?

- A known issue that was historically not formalized is the question of what the outcome is specified as: logs, or levels?

- Hopefully it's clear that if something satisfies pre-trends in logs, it seems unlikely to satisfy in levels

- Recall that this is the issue of *invariance* we discussed with quantile treatment effects
  - In our parametric setting, if there are time trends in the outcomes, the parallel trends are likely not to hold for all transformations of the variables.
  - That could be problematic if you wanted to be agnostic about the model!

- Roth and Sant'Anna (2021) directly discuss this issue. Their suggestion:

  Our results suggest that researchers who wish to point-identify the ATT should justify one of the following: (i) why treatment is as-if randomly assigned, (ii) why the chosen functional form is correct at the exclusion of others, or (iii) a method for inferring the entire counterfactual distribution of untreated potential outcomes.

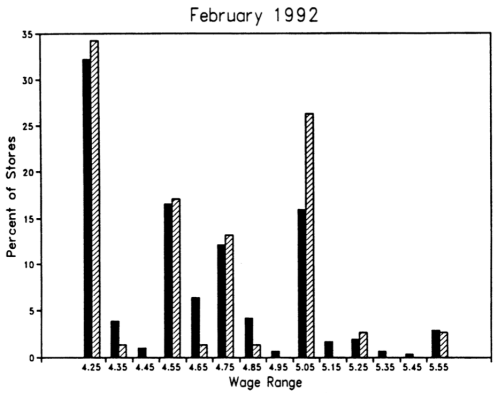
Cases of DinD

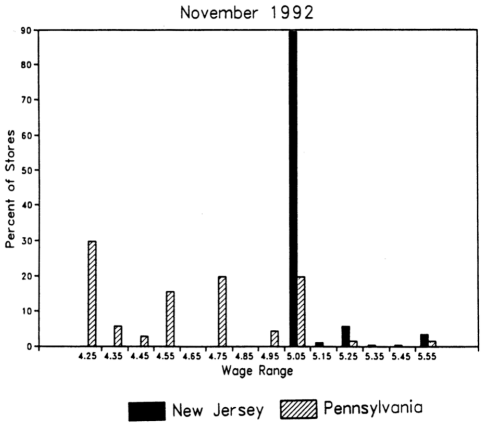
- 1 treatment timing, Binary treatment, 2 periods
  - Card and Krueger (AER, 1994)
- 1 treatment timing, Binary treatment, T periods
  - Yagan (AER, 2015)
- 1 treatment timing, Continuous treatment
  - Berger, Turner and Zwick (JF, 2020)
- Staggered treatment timing, Binary treatment
  - Bailey and Goodman-Bacon (AER, 2015)

Card and Krueger (1994)

- Card and Krueger (1994) study the impact of New Jersey increasing the minimum wage 4.25 to 5.05 dollars an hour on April 1, 1992
- Key question is what impact does this have on employment?
  - Need a counterfactual for NJ, and use Pennsyvania as a control
- Collected data in 410 fast food restaurants
  - Called places and asked for employment and starting wage data
  - Sample data from Feb 1992 and Nov 1992
- Hence, D_i is NJ vs PA, and $t = 0$ is Feb 1992 and $t = 1$ is Nov 1992

Stark Effect on Wages in Card and Krueger (1994)





Effect on Employment in Card and Krueger (1994)

- Despite a large increase in wages, seemingly no negative impact on employment
  - In fact, marginally significant positive impact
- Looking at raw data, this positive impact is driven by a decline in PA
  - This decline is reasonable if you think that PA is a good counterfactual, since 1992 is in the middle of a recession
- A second comparison can be run with stores whose starting wage in pre-period was above treatment cutoff
  - These stores perform similarly to PA

Variable	Stores by state: PA (i)	NJ (ii)	Difference, NJ − PA (iii)	Stores in New Jersey[a]: Wage = $4.25 (iv)	Wage = $4.26–$4.99 (v)	Wage ≥ $5.00 (vi)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)	19.56 (0.77)	20.08 (0.84)	22.25 (1.14)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)	20.88 (1.01)	20.96 (0.76)	20.21 (1.03)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)	1.32 (0.95)	0.87 (0.84)	−2.04 (1.14)

Key considerations for thinking about Card and Krueger (1994)

- The treatment can't really be thought of as randomly assigned
  - Treatment is completely correlated within states
  - As a result, any within-state correlation of errors will be correlated with treatment status

- Given the limited number of states, time periods, and treatments, more valuable to view this as a case study
  - Under strong parametric assumptions, can infer causality!
  - Card acknowledges (Card and Krueger interview with Ben Zipperer):

So the great advantage of a quasi-experiment or natural experimental like minimum wage is that it's a real intervention. It's real firms that are all affected. You get part of the general equilibrium effect. That's pretty important for understanding the overall story. The disadvantage is that someone can always say, well, it isn't truly random. And the number of units might be small. So you might only have two states. At some abstract level, there's only two degrees of freedom there. And so that's a problem.

Yagan (2015)

- Yagan (2015) tests whether the 2003 dividend tax cut stimulated corporate investment and increased labor earnings

- Big empirical question for corporate finance and public finance

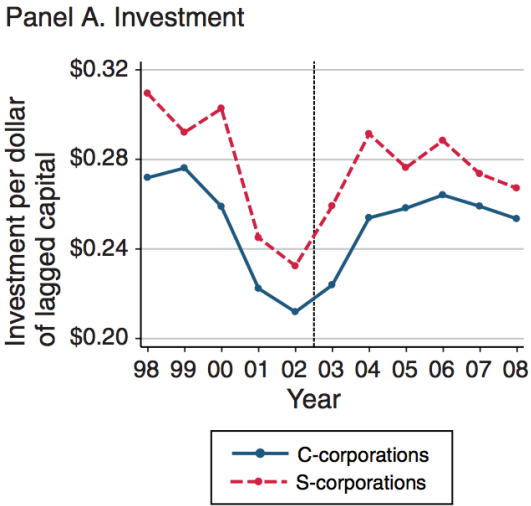
- No direct evidence on the real effects of dividend tax cut
  - real corporate outcomes are too cyclical to distinguish tax effects from business cycle effects, and economy boomed

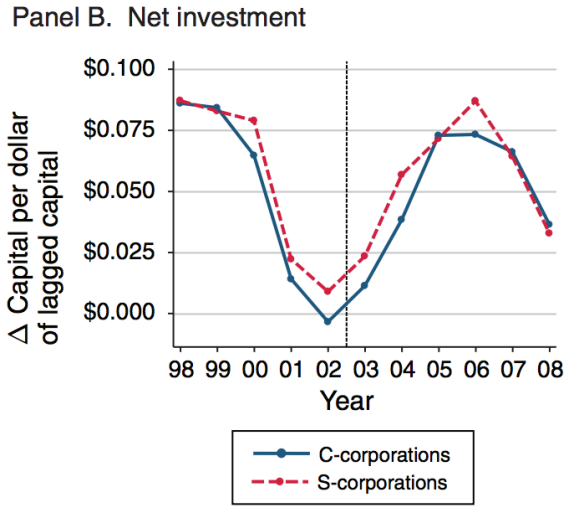
- Paper uses distinction between “C” corp and “S” corp designation to estimate effect
  - Key feature of law: S-corps didn’t have dividend taxation

- Identifying assumption (from paper):

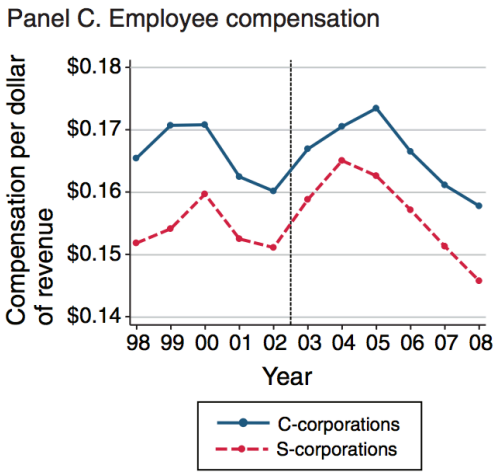
  The identifying assumption underlying this research design is not random assignment of C- versus S-status; it is that C- and S-corporation outcomes would have trended similarly in the absence of the tax cut.

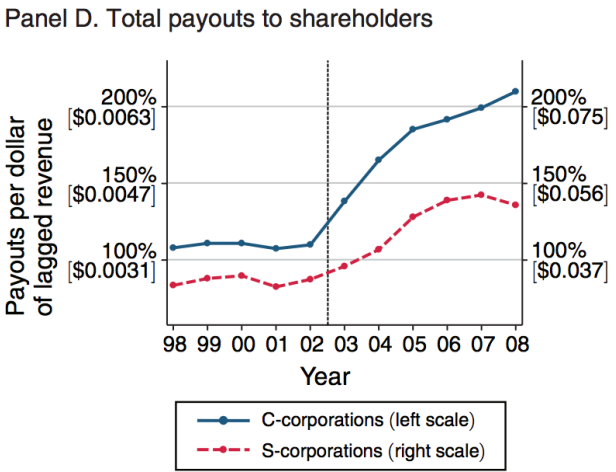
Investment Effects (none)





Employee + Shareholder effects (big)





Key Takeaway + threats

- Tax reform had zero impact on differential investment and employee compensation
- Challenges orthodoxy on estimates of cost-of-capital elasticity of investment
- What are underlying challenges to identification?
  1. Have to assume (and try to prove) that the only differential effect to S- vs C-corporations was through dividend tax changes
  2. During 2003, could other shocks differentially impact?
     - Yes, accelerated depreciation – but Yagan shows it impacts them similarly.
- Key point: you have to make *more* assumptions to assume that zero **differential** effect on investment implies zero **aggregate** effect.

Berger, Turner and Zwick (2019)

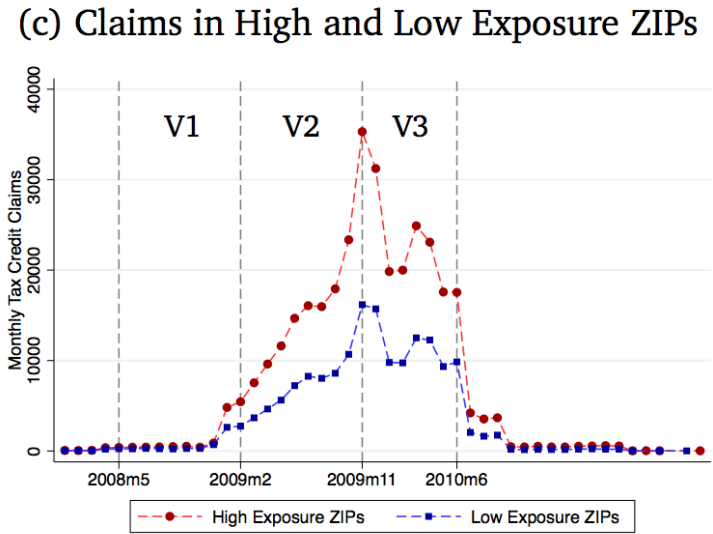
- This paper studies the impact of temporary fiscal stimulus (First-Time Home Buyer tax credit) on housing markets

- Policy was differentially targetted towards first time home buyers
  - Define program exposure as "the number of potential first-time homebuyers in a ZIP code, proxied by the share of people in that ZIP in the year 2000 who are first-time homebuyers"
  - The design:

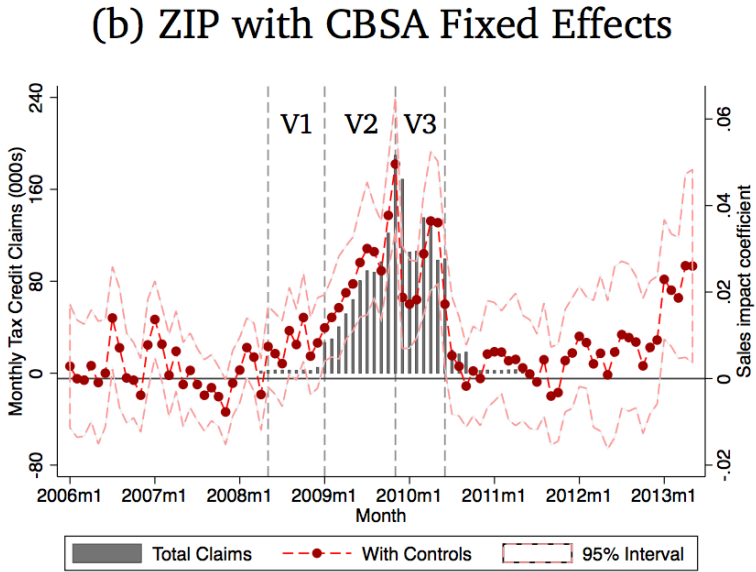
    The key threat to this design is the possibility that time-varying, place-specific shocks are c related with our exposure measure.

- This measure is **not** binary – we are just comparing areas with a low share vs. high share, effectively. However, we have a dose-response framework in mind – as we increase the share, the effect size should grow.

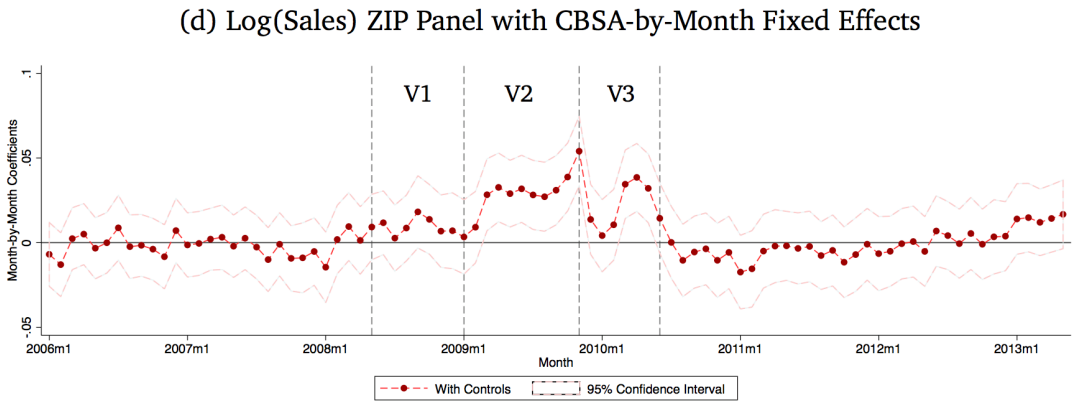
First stage: Binary approximation



First stage: Regression coefficients



Final Outcome: Regression coefficients



Binary Approximation vs. Continuous Estimation

- Remember our main equation did not necessarily specify that D_{it} had to be binary.

$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=T_0, t \neq T_1}^{T_2} \delta_t D_{it} + \epsilon_{it}, \tag{2}$$

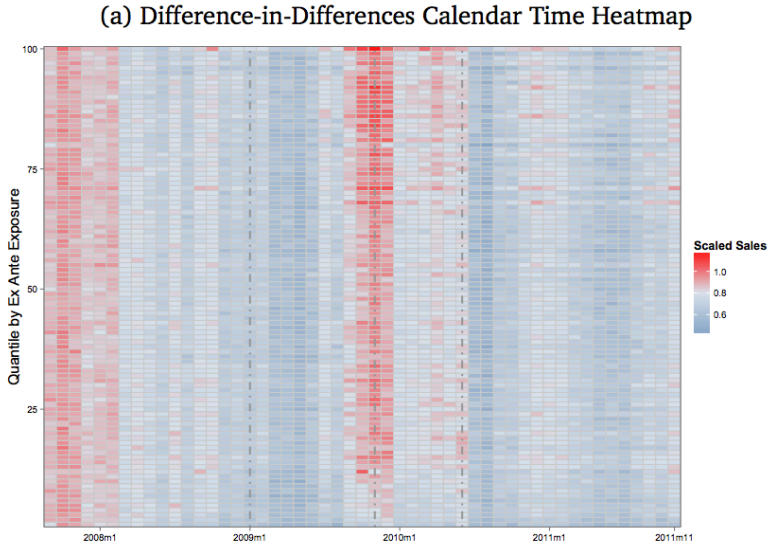
- However, if it is continuous, we are making an additional strong functional form assumption that the effect of D_{it} on our outcome is linear.

- We make this linear approximation all the time in our regression analysis, but it is worth keepping in mind. It is partially testable in a few ways:
  - Bin the continuous D_{it} into quartiles $\{\widetilde{D}_{itk}\}_{k=1}^{4}$ and estimate the effect across those groups:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=T_0, t \neq T_1}^{T_2} \sum_{k=1}^{4} \delta_{t,k} \widetilde{D}_{it,k} + \epsilon_{it}. \tag{3}$$

  - What does the ordering of $\delta_{t,k}$ look like? Is it at least monotonic?

Berger, Turner and Zwick implementation of linearity test



Takeaway

- When you have a continuous exposure measure, can be intuitive and useful to present binned means “high” and “low” groups
- However, best to present regression coefficients of the effects that exploits the full range of the continuous measure so that people don’t think you’re data mining
- Consider examining for non-monotonicities in your policy exposure measure
- This paper is still has only one “shock” – one policy time period for implementation

Bailey and Goodman-Bacon (2015)

- Paper studies impact of rollout of Community Health Centers on mortality
  - Idea is that CHCs can help lower mortality (esp. among elderly) by providing accessible preventative care

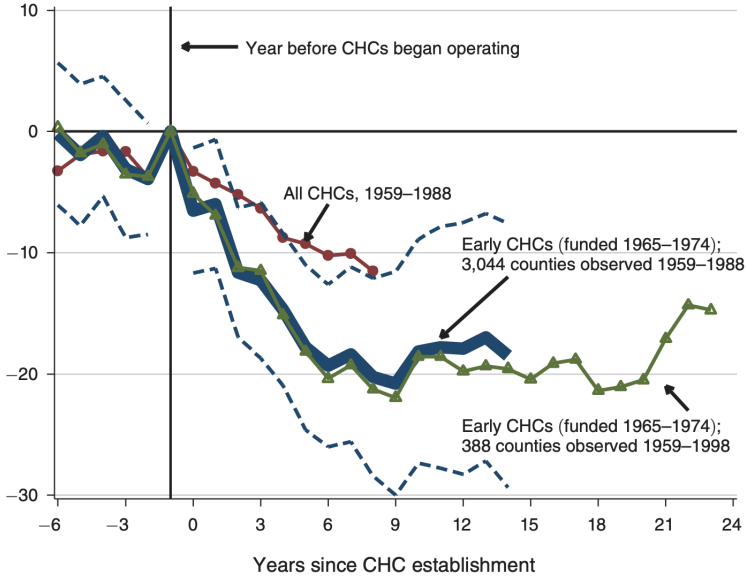
- Exploit timing of implementation of CHCs

  Our empirical strategy uses variation in when and where CHC programs were established to quantify their effects on mortality rates. The findings from two empirical tests support a key assumption of this approach—that the timing of CHC establishment is uncorrelated with other determinants of changes in mortality.

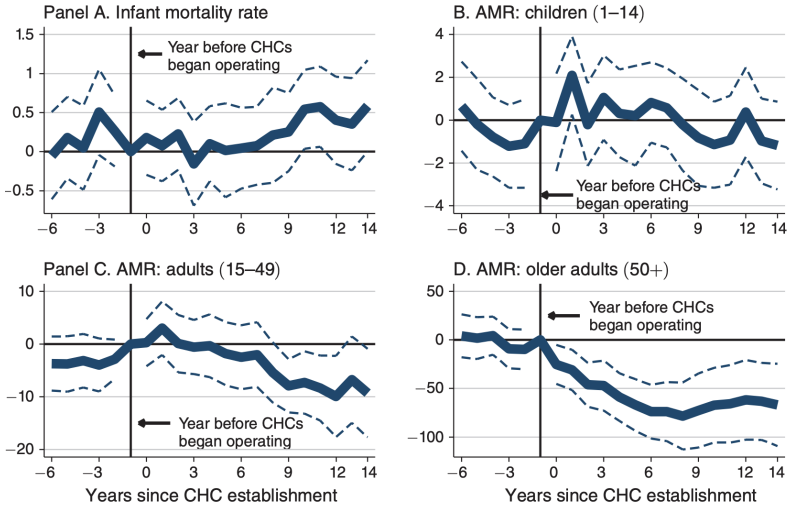
- Issue is that CHCs tend to be done in places

- Since CHCs are started in different places in different time periods, we estimate effects in *event-time*, e.g. relative to initial rollout.

Negative effect on mortality



Negative effect on mortality, particularly among elderly



Key takeaways

- Since the policy changes are staggered, we are less worried about effect driven by one confounding macro shock.
- Easier to defend story that has effects across different timings
  - Also allows us to test for heterogeneity in the time series
- Still makes the exact same identifying assumptions – parallel trends in absence of changes

But a big issue emerges when we exploit differential timing

- We have been extrapolating from the simple pre-post, treatment-control setting to broader cases
  - multiple time periods of treatment

- In fact, in some applications, the policy eventually hits everyone – we are just exploiting differential timing.

- If we run the "two-way fixed effects" model for these times of DinD

$$y_{it} = \alpha_i + \alpha_t + \beta^{DD} D_{it} + \epsilon_{it} \tag{4}$$

  what comparisons are we doing once we have lots of timings?

- Key point: is our *estimator* mapping to our *estimand*?

- Well, what's our estimand?

What is our estimand with staggered timings?

- There are a huge host of papers touching on this question
- Callaway and Sant'anna (2020) propose the following building block estimand:

$$\tau_{ATT}(g,t) = E(Y_{it}(1) - Y_{it}(0)|D_{it} = 1 \forall t \geq g), \tag{5}$$

  the ATT in period t for those units whose treatment turns on in period g.
  - In the 2x2 case, this was exactly our effect!
  - This paper assumes absorbing treatment, but can be weakened in other papers (de Chaisemartin and d'Haultfoeuille (2020) discuss this)
- It seems very reasonable that for our overall estimand, we want some weighted combinated of these ATTs
- Callaway and Sant'anna (2020) highlight two ways to identify the above estimand:
  1. Parallel trends of treatment group with a group that is "never-treated"
  2. Parallel trends of treatment group with the group of the "not yet treated"
- Using these estimands, C&S provide a very natural set of potential ways to aggregate these estimands up

Wait what happened to TWFE?

- It turns out that the logic of the TWFE does not naturally extend to differential timings
  - Recall that from our discussion of linear regression, regression is great because it does a variance weighted approximation:

$$\tau = \frac{E(\sigma_D^2(W_i)\tau(W_i))}{E(\sigma_D^2(W_i))}, \qquad \sigma_D^2(W_i) = E((D_i - E(D_i|W_i))^2|W_i)$$

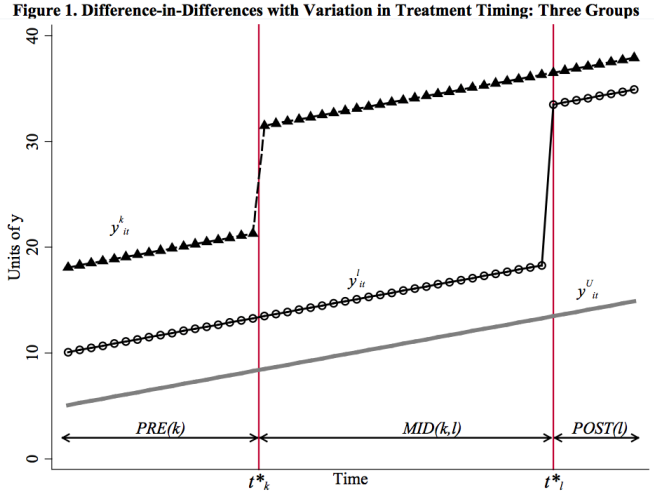
  - It turns out that in the panel setting with staggered timings, these weights are not necessarily positive

- Key insight from several papers: with staggered timings + heterogeneous effects, the TWFE approach to DinD (both using a single pooled estimator, or using an event study) can put large negative weight on certain groups' estimands, and large positive weight on others
  - Serious issue for interpretability
  - Some example papers: Borusyak and Jaravel (2017), de Chaisemartin and D'Haultfœuille (2020), Goodman-Bacon (2019), Sun and Abraham (2020)

- Key point: this is *solveable*. Merely a construct of being overly casual with estimator definition

Goodman-Bacon 2x2 comparisons

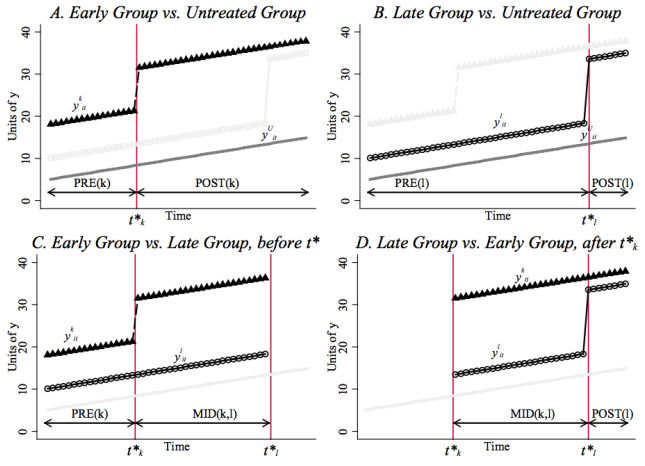
- Consider two staggered treatements and a never-treated group
- What does the TWFE estimator estimate?



Goodman-Bacon 2x2 comparisons

- Four potential comparisons that can be made
- turns out that TWFE DD estimator (pooled) is the weighted average of all 2x2 comparisons
- These weights end up putting a high degree of weight on units treated in the middle of the sample (since they have the highest variance in the treatment indicator!)

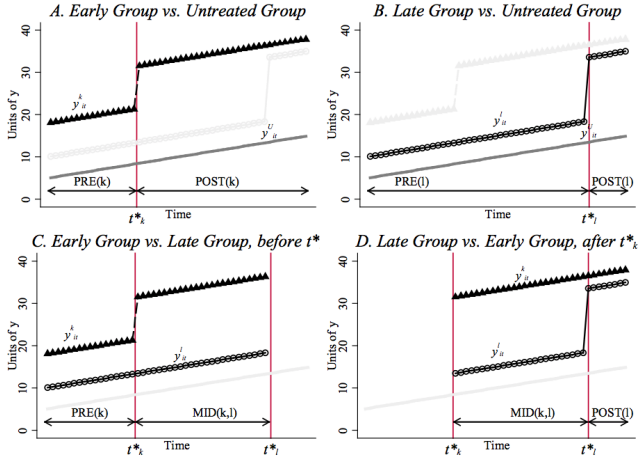
Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



Goodman-Bacon 2x2 comparisons

- The weighting becomes problematic if the effects vary over time – if the effects are instantaneous and time-invariant, the weights are all positive

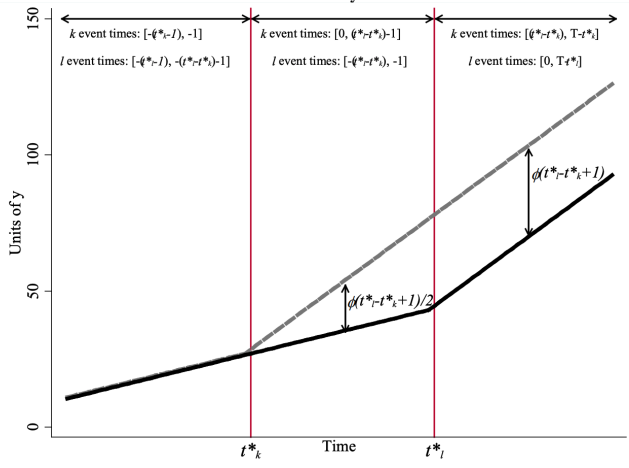
Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



Goodman-Bacon 2x2 comparisons

- The weighting becomes problematic if the effects vary over time – if the effects are instantaneous and time-invariant, the weights are all positive
- However, time-varying effects create bad counterfactual groups, and create negative weights
- Goodman-Bacon provides a way to assess the weights in a given TWFE design

Figure 3. Difference-in-Differences Estimates with Variation in Timing Are Biased When Treatment Effects Vary Over Time



What to do with staggered timing in DinD?

- There's really no reason to use the baseline TWFE in staggered timings
  - A perfect example wherein the estimator does not generate an estimate that maps to a meaningful estimand

- There are several approaches proposed in the literature that are just as good!
  - Sun and Abraham (2020)
  - de Chaisemartin and d'Haultfoeuille (2020)
  - Borusyak and Jaravel (2017)
  - Callaway and Sant'anna (2020)

- These all are robust to this issue. I find Callaway and Sant'anna quite intuitive, but your circumstances may vary slightly. Key piece to keep in mind that differs a bit:
  - Is my treatment absorbing?

- Irrespective of the exact paper, the key point is that we are generating a counterfactual and need to be careful that our estimator does so correctly

Issues to keep in mind:
1) what is inference – e.g. if you do PA vs. NJ, you can't really talk about design based inference. Just sampling. CIte andreas hagmann work
2) Inference: What's your "experiment"? Is it feasible to think about this in a design based framework? Roth and Sant'anna, Athey and Imbens (2018) propose some. Worth going back in time to at least point out the issue from BDM
3) Are pre-trends really pre-trends? Roth says no. Additionally, uniform confidence intervals

Finally, a discussion on inference

- First, let's start with the old school fact that you must know if you are working with panel data and Dind
- **You must cluster on the unit of policy implementation if possible**. See Bertrand, Duflo and Mullainathan (2004)
  - I say "if possible" since clearly in Card and Krueger that is infeasible
- If the policy variation is implemented at the industry level, you should not cluster at the firm level
- If the policy variation is implemented at the firm level, you cannot use robust standard errors

Small clusters

- The Card and Krueger case was too extreme, but there are approaches for dealing with a small number of clusters
- This apporach typically involves bootstrapping, and can handle small number of treated groups relative to the overall population
- See Andreas Hagemann's work for a place to start

Uniform confidence intervals

- Finally, when considering event study graphs, pre-trend graphs should use uniform confidence intervals, rather than pointwise confidence intervals
  - Advocated for by Freyaldenhoven et al (2018):

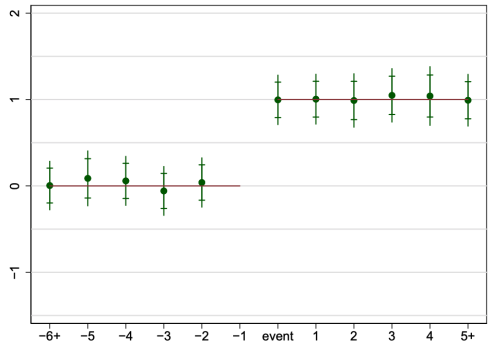
- Code available here thanks to Ryan Kessler:
  `https://github.com/paulgp/simultaneous_confidence_bands`

> Figure 2 shows both pointwise 95% confidence intervals and uniform 95% sup-t confidence bands (Olea and Plagborg-Møller 2019). Applied papers commonly include pointwise confidence intervals in event plots.[12] These permit testing only of preselected pointwise hypotheses. Uniform bands such as those we show here are designed to contain the true path of the coefficients 95% of the time, and are therefore arguably more useful for giving readers a sense of what kinds of pre-trends are consistent with the data.

Uniform confidence intervals

- Finally, when considering event study graphs, pre-trend graphs should use uniform confidence intervals, rather than pointwise confidence intervals
  - Advocated for by Freyaldenhoven et al (2018):

- Code available here thanks to Ryan Kessler:
  `https://github.com/paulgp/simultaneous_confidence_bands`



Conclusion

- Difference in difference is hugely powerful in applied settings
- Does not require random assignment, but rather implementation of policies that differentially impacts different groups and is not confounded by other shocks at the same time.
- Can be a great application of big data, with convincing graphs that highlight your application
- Also allows for partial tests of identifying assumptions
- Worth carefully thinking about what your identifying assumptions are in each setting, and transparently highlighting them.
- Important to note that this always identifies a *relative* affect, and to aggregate, you will typically need a model and additional strong assumptions (see Auclert, Dobbie and Goldsmith-Pinkham (2019) for an example in a macro setting).

My takeaways from new literature

- Beware weak tests of pre-trends. Consider using R&R's partial identification tests to assess robustness of results.

- Do not worry about the new literature on staggered timings if you only have one timing!

- Think carefully about your estimand if you're using a staggered timing DinD – what's your counterfactual in each case?
  - Software exists for many of these papers. This is doable!

- When plotting confidence intervals in event studies, you should plot uniform confidence intervals.