Difference-in-Differences

Paul Goldsmith-Pinkham

February 12, 2019

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)

Basic setup

- Assume we have N firms (i) and T time periods (t)
- Consider a binary policy D, and we are interested in estimating its effect on Y
- Key assumption underlying (parametric) version of the potential outcomes model:

$$E(Y_{it}(D = 0)|i, t) = \alpha_i + \gamma_t$$

- Implication? In the absence of the treatment, the firms' Y **evolve in parallel** – their γ_t are identical
  - This is the key identifying assumption – absent the policy, firms may have different *levels* (α_i) but their changes would evolve in parallel.

Basic setup

- If 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 firm i's outcomes – bigger firms are assigned the treatment, and so it looks like firms are "caused" to be bigger

- What if there are two time periods? We can make a lot more progress! Let D_{it} denote whether firm i is treated in period t.
  - We assume that $t = 0$ no one is treated, while in $t = 1$, some are treated. This is a big assumption! When could this fail?

- Then, our estimating equation can be written as

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

where $\delta = E(Y_{it}(D_{it} = 1) - Y_{it}(D_{it} = 0)|i, t)$, our estimand of interest. We have assumed this is constant (e.g. not time varying).

Basic Setup

- Now consider the simple first differences for firms a and b where firm a was not affected and firm b was:

$$E(Y_{it}|i = a, t = 1) - (Y_{it}|i = a, t = 0) = \gamma_1 - \gamma_0 \tag{2}$$
$$E(Y_{it}|i = b, t = s) - (Y_{it}|i = b, t = s') = \gamma_1 - \gamma_0 + \delta \tag{3}$$

- Hence, we can identify δ by taking the second difference between these groups:

$$\delta = E(Y_{it}|i = b, t = s) - (Y_{it}|i = b, t = s') \tag{4}$$
$$- E(Y_{it}|i = a, t = 1) - (Y_{it}|i = a, t = 0). \tag{5}$$

- A simple linear regression following Equation 1 will identify this parameter exactly.

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

Multiple time periods in basic setup

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? For time periods $t \in [T_0, T_2]$, where the policy occurs at period $T_1 + 1$:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=T_0, t \neq T_1}^{T_2} \delta_t D_{it} + \epsilon_{it}, \tag{6}$$

One of the coefficients is fundamentally unidentified because of α_i – all coefficients test the *relative* effect to period T_1. Pre-test is joint test of insignificance of coefficients before policy

Key point about inference before examples

- **You must cluster on the unit of policy implementation**
- If the policy variation is implemented at the industry level, you cannot cluster at the firm level
- If the policy variation is implemented at the firm level, you cannot use robust standard errors

See Bertrand, Duflo and Mullainathan (2004)

Three Cases of DinD

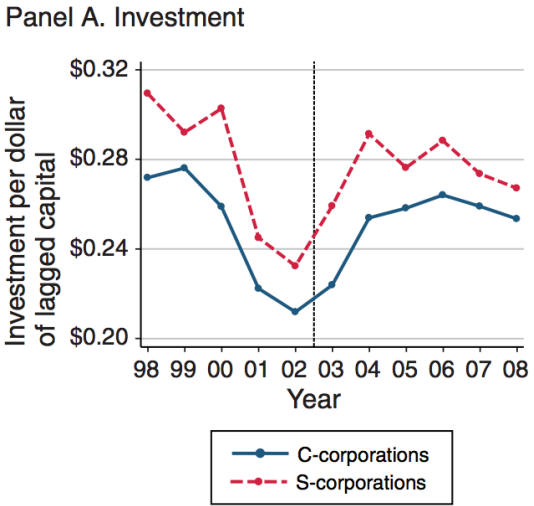
- 1 treatment timing, 1 treated (and 1 control) group
  - Yagan (AER, 2015)
- 1 treatment timing, Continuous treatment
  - Berger, Turner and Zwick (R&R at JF, 2019)
- Many time period treatment, 1 treated (and 1 control) group
  - Jeffers (2018)

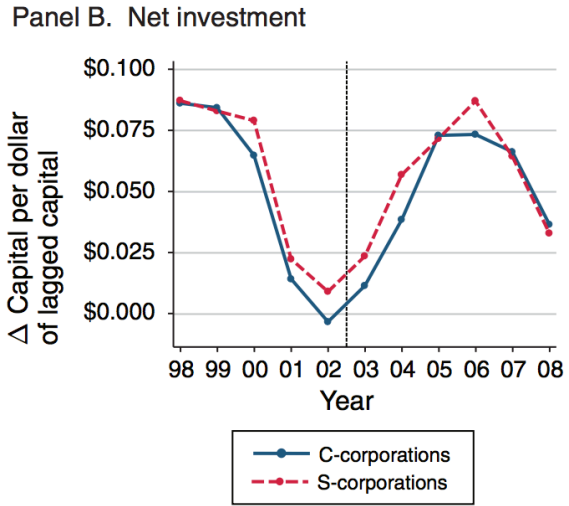
Yagan (2015)

- 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):

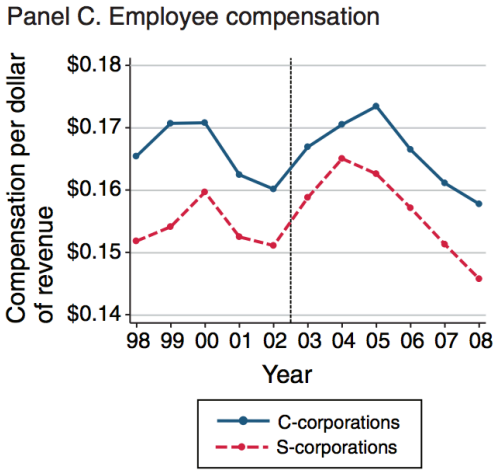
  The identifying assumption underlying this research design is not random assignment of C- versus S-status; it is that C- and S-corporation outcomes would have trended similarly in the absence of the tax cut.

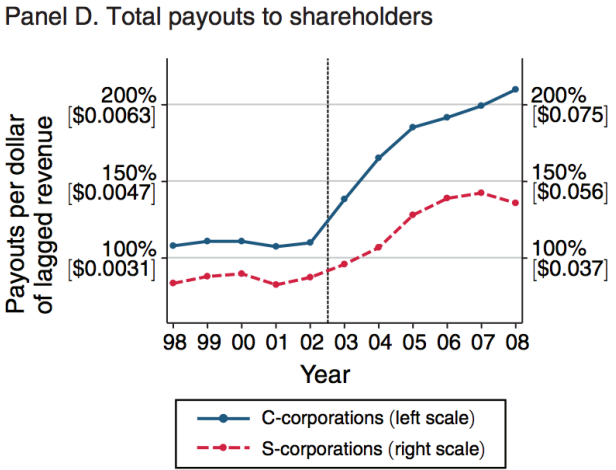
Investment Effects (none)





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)

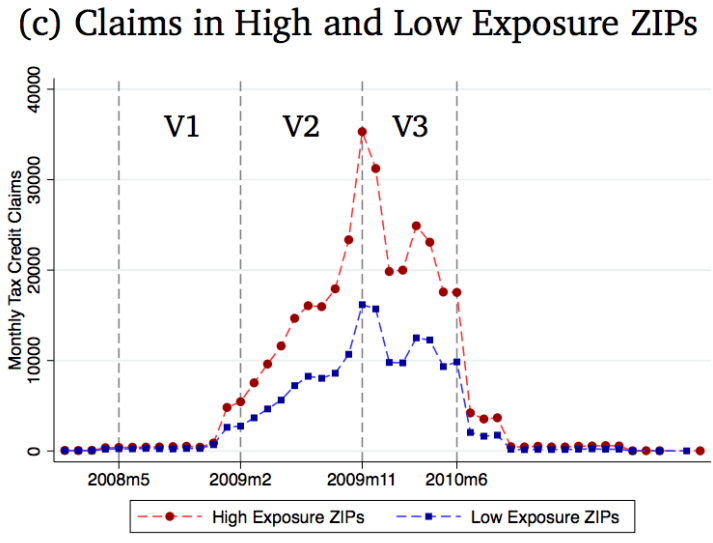
- This paper studies the impact of temporary fiscal stimulus (First-Time Home Buyer tax credit) on housing markets

- 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:

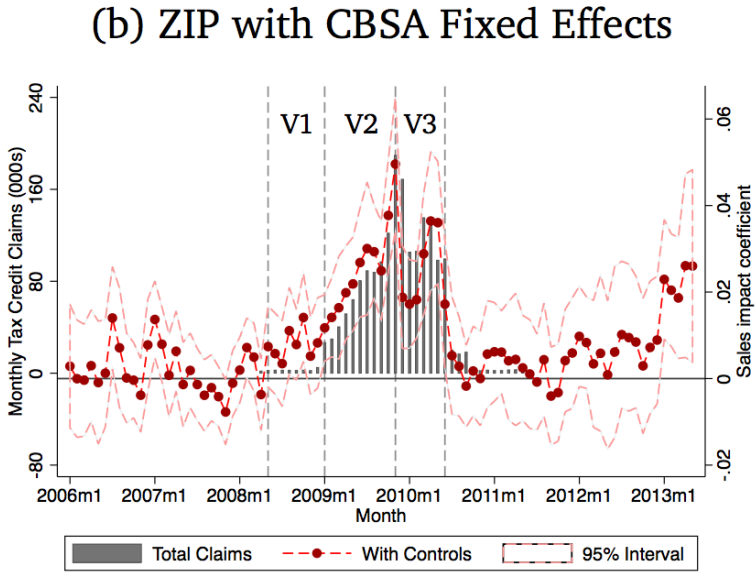
    The key threat to this design is the possibility that time-varying, place-specific shocks are c related with our exposure measure.

- 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.

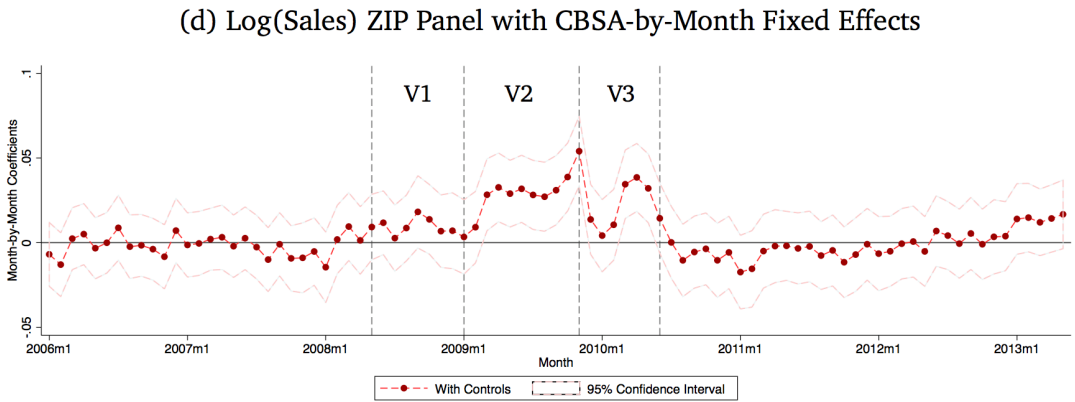
First stage: Binary approximation



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.

$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=T_0, t \neq T_1}^{T_2} \delta_t D_{it} + \epsilon_{it}, \tag{7}$$

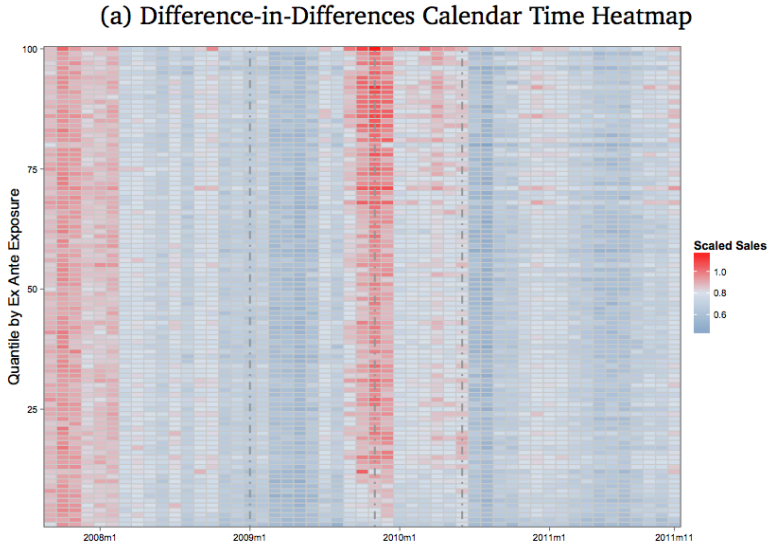
- However, if it is continuous, we are making an additional strong functional form assumption that the effect of D_{it} on our outcome is linear.

- 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:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=T_0, t \neq T_1}^{T_2} \sum_{k=1}^{4} \delta_{t,k} \widetilde{D}_{it,k} + \epsilon_{it}. \tag{8}$$

  - 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

Jeffers (2018)

- Paper studies impact of restricting labor mobility on entrepreneurial activity and capital investment.

- Exploit timing of changes in enforcibility of non-compete agreements across states

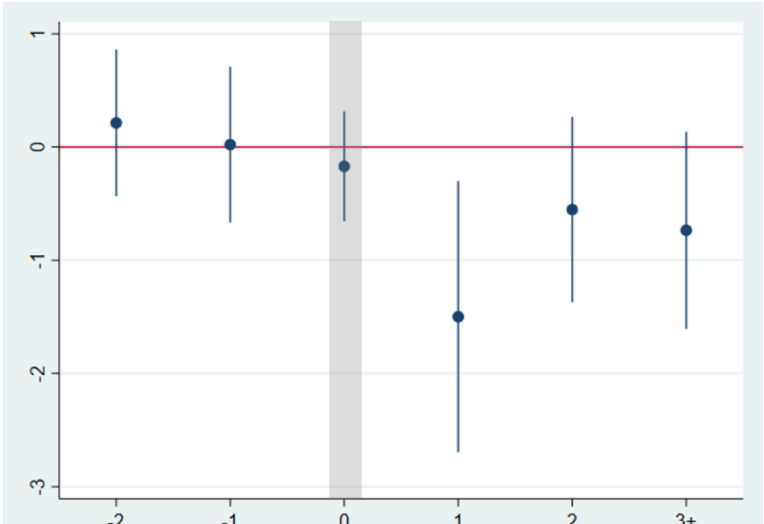
  > *My identification strategy relies on seven state supreme court rulings and one law that modified the enforceability of NCs between 2008 and 2014.*

- Observes individuals at firms using LinkedIn data, and can measure departures.

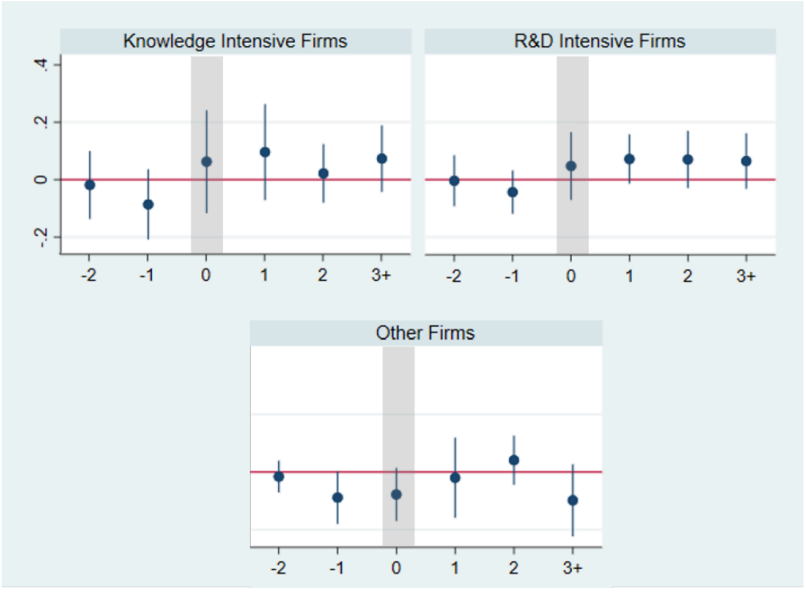
- Since laws are changed in different places in different time periods, we estimate effects in *event-time*, e.g. relative to law changes.

Negative effect on departures

$$100 * \frac{\#\text{ departures}}{\#\text{ employees}}_{it} = \alpha + \sum_m \beta_m \{\text{treated}_i * \text{m years to treatment}\} + \gamma_i + \theta_{jt} + \epsilon_{it}$$



Positive effect on investment



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

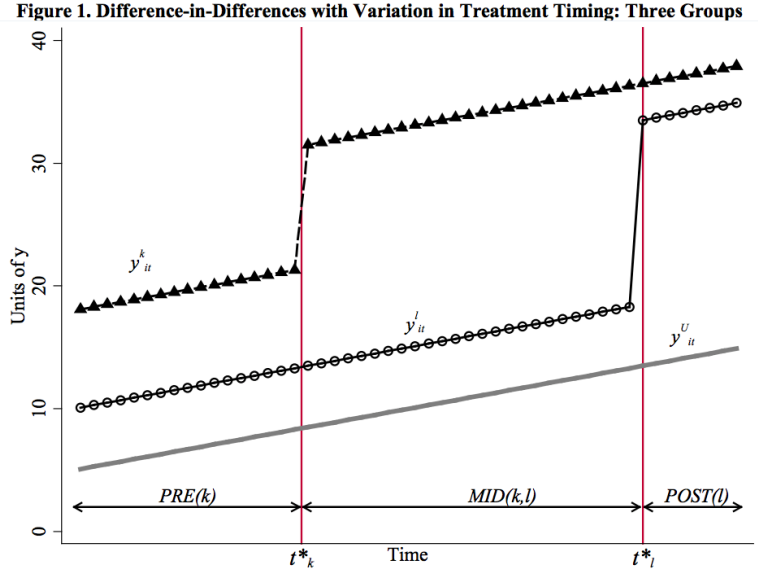
$$y_{it} = \alpha_i + \alpha_t + \beta^{DD} D_{it} + \epsilon_{it} \tag{9}$$

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

- How should we map this estimation

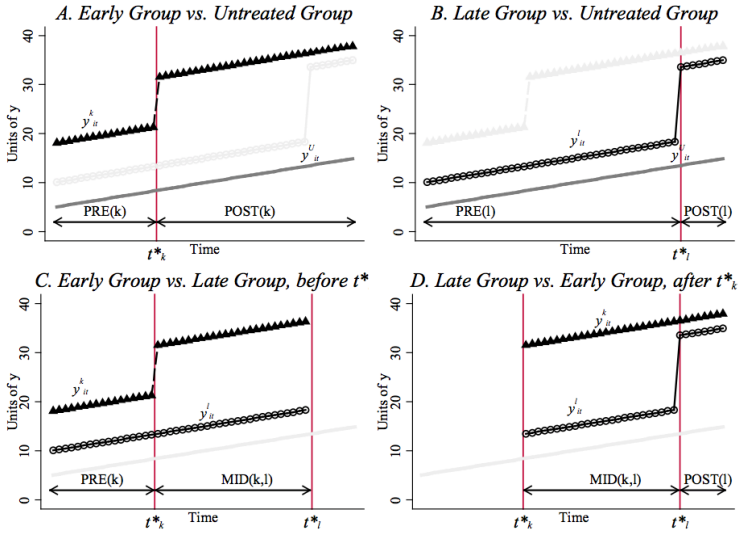
- Goodman-Bacon (2018) talks about exactly this.

Simple example data



Combinations of 2x2 comparisons

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



Weighted combinations of 2x2 comparisons

- It turns out that the TWFE DD coefficient is a variance-weighted average of all 2x2 diff-in-diff comparisons you could form.

- This can be a serious issue if you have effects that vary over time: the treatment effects themselves put already treated units on a differential trend. Using already-treated units as controls will bias your results!

- Goodman-Bacon paper shows how to construct weights to identify which comparisons are getting the largest weight
  - Does it come from comparing to “already-treated” groups?

- Very good twitter thread that discusses the paper here: urlhttps://twitter.com/agoodmanbacon/status/1039126592604303360

Additional papers discussing diff-in-diff

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3148250
http://scholar.harvard.edu/files/borusyak/files/event_studies_may8_website.pdf
https://arxiv.org/abs/1804.05785v1
https://sites.google.com/site/clementdechaisemartin/two_way_FE.pdf?attredirects=0&d=1
https://www.antonstrezhnev.com/s/generalized_did.pdf
http://www.nber.org/papers/w24963

Conclusion + Onward to Bartik

- 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).