Diff-in-diff II

Advanced Applied Econometrics

Felix Weinhardt

Slides based on Scott Cunningham, Paul Goldsmith-Pinkham and Andrew Goodman-Bacon. Special thanks to Renke Schmacker

Differences-in-differences recap

- In diff-and-diff, we compare differential time trends between treatment and control groups
- Control group gives us an estimate of the counterfactual outcome for the treatment group
- Central assumptions
  - Parallel trends
  - SUTVA

Differences-in-differences recap

- In diff-and-diff, we compare differential time trends between treatment and control groups
- Control group gives us an estimate of the counterfactual outcome for the treatment group
- Central assumptions
  - Parallel trends
  - SUTVA
- Parallel pre-trends are important, but no silver bullet
- Let's look at pre-trends more carefully

Testing parallel pre-trends

- Diverging pre-trends can falsify the parallel trends assumption
- Showing parallel pre-trends lends support to the validity of the design (worth doing!)
- However, recent literature pointed out issues with testing pre-trends
  - Parallel trends in what?
  - Pre-testing issues

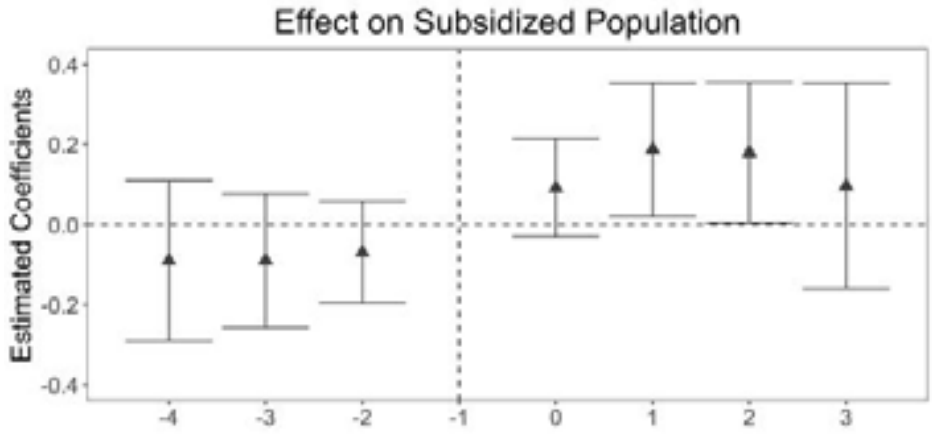
Parallel trends in what?

- How is the outcome specified: in logs or levels?
- If you have parallel pre-trends in logs, they are unlikely to hold in levels and vice versa
- OLS does not have the invariance property that, e.g., quantile regression has
- Roth and Sant'Anna (2021):

  "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."

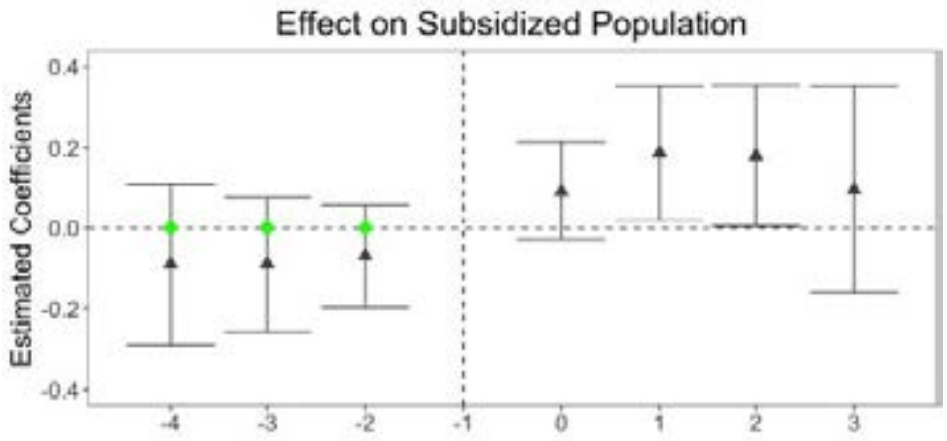
Pre-testing issues (Roth 2020)

- "Event study" visualization of pre-trends
- Zero is in the 95% CIs, so we can't reject parallel trends
- But a sizable pre-trend is in the 95% CI as well
- With low power, we cannot rule out an important alternative hypothesis
- With highly precise pre-trends, the problem is less relevant



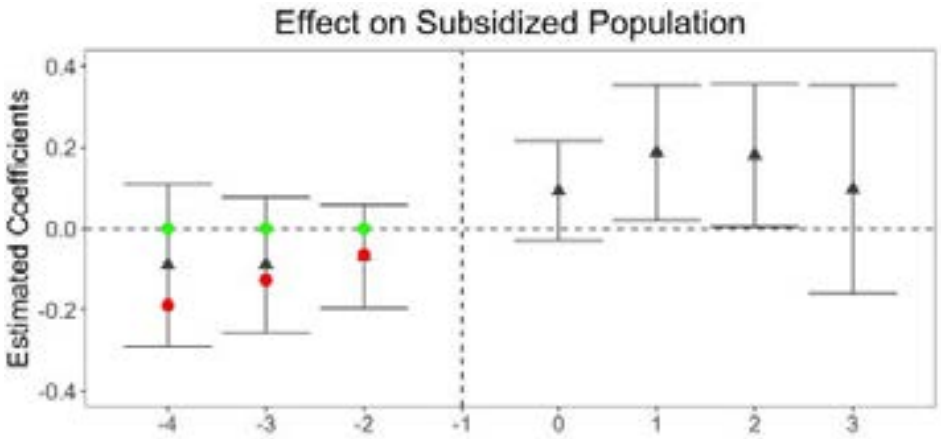
Pre-testing issues (Roth 2020)

- "Event study" visualization of pre-trends
- Zero is in the 95% CIs, so we can't reject parallel trends
- But a sizable pre-trend is in the 95% CI as well
- With low power, we cannot rule out an important alternative hypothesis
- With highly precise pre-trends, the problem is less relevant



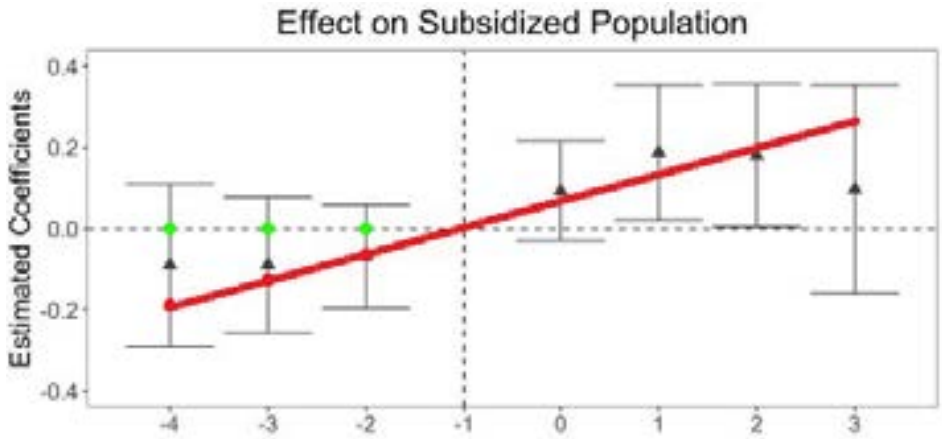
Pre-testing issues (Roth 2020)

- "Event study" visualization of pre-trends
- Zero is in the 95% CIs, so we can't reject parallel trends
- But a sizable pre-trend is in the 95% CI as well
- With low power, we cannot rule out an important alternative hypothesis
- With highly precise pre-trends, the problem is less relevant



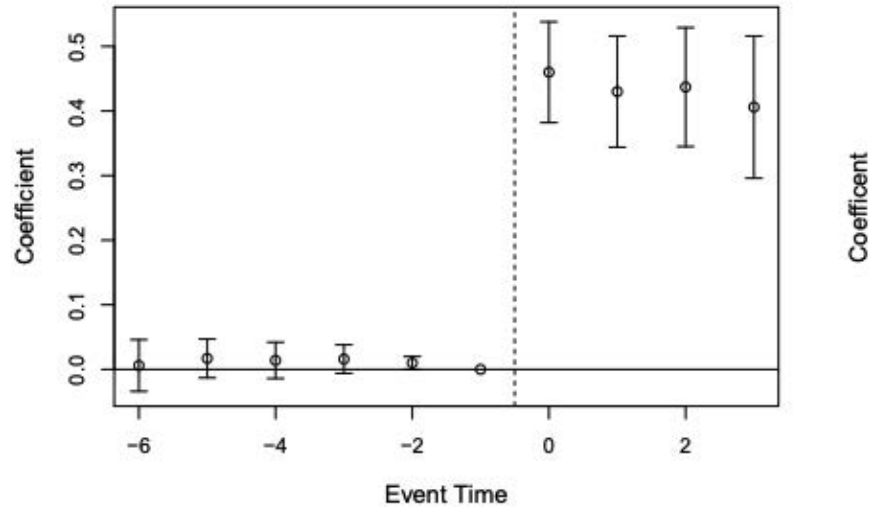
Pre-testing issues (Roth 2020)

- "Event study" visualization of pre-trends
- Zero is in the 95% CIs, so we can't reject parallel trends
- But a sizable pre-trend is in the 95% CI as well
- With low power, we cannot rule out an important alternative hypothesis
- With highly precise pre-trends, the problem is less relevant



Pre-testing issues (Roth 2020)

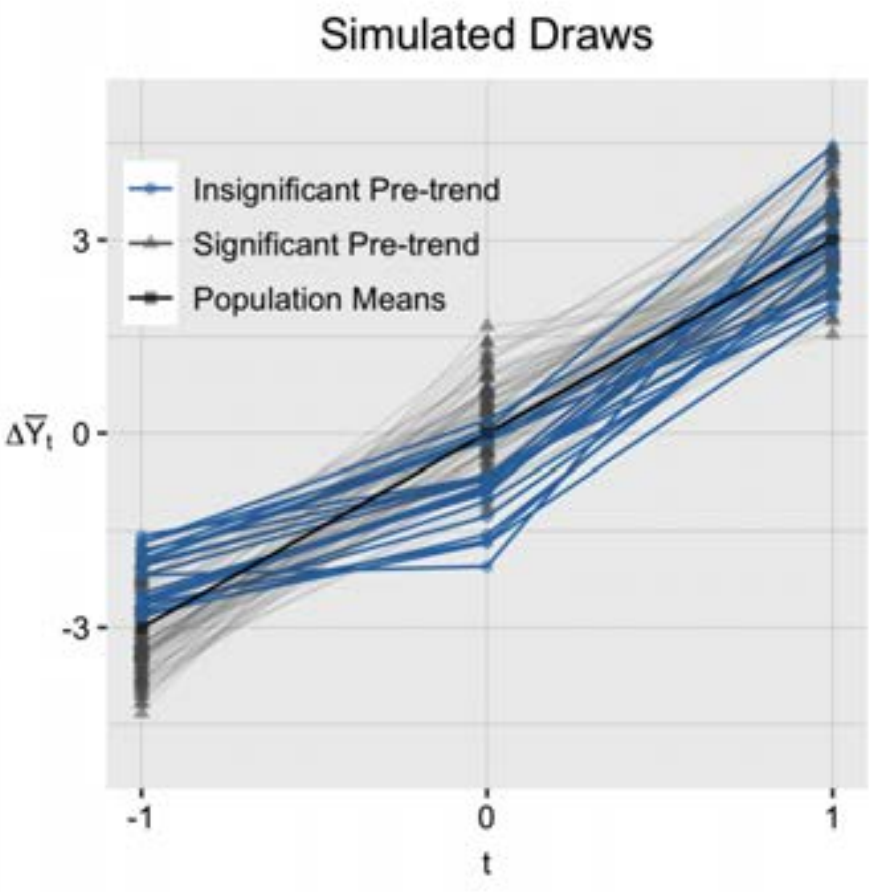
- "Event study" visualization of pre-trends
- Zero is in the 95% CIs, so we can't reject parallel trends
- But a sizable pre-trend is in the 95% CI as well
- With low power, we cannot rule out an important alternative hypothesis
- With highly precise pre-trends, the problem is less relevant

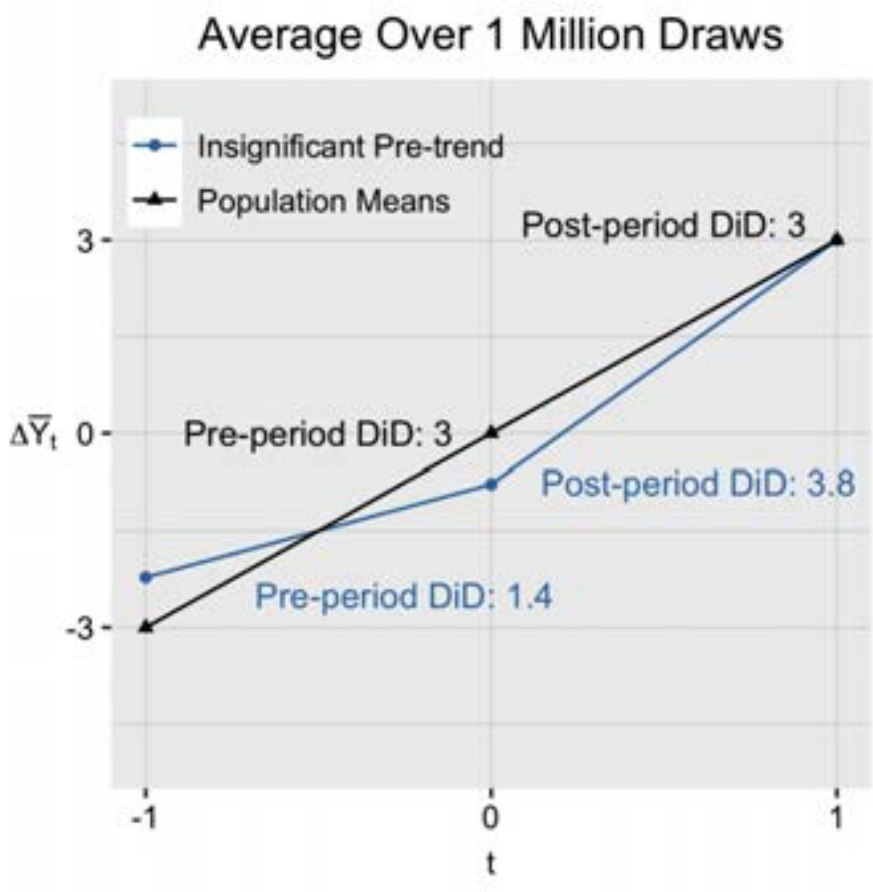


Outcome: Medicaid Eligibility

Pre-testing issues (Roth 2020)

- By selecting on pre-trends that "pass", will tend to choose baseline realizations that satisfy pre-trends, but induce *bias* in the effect





Pre-testing issues (Roth 2020)

- 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

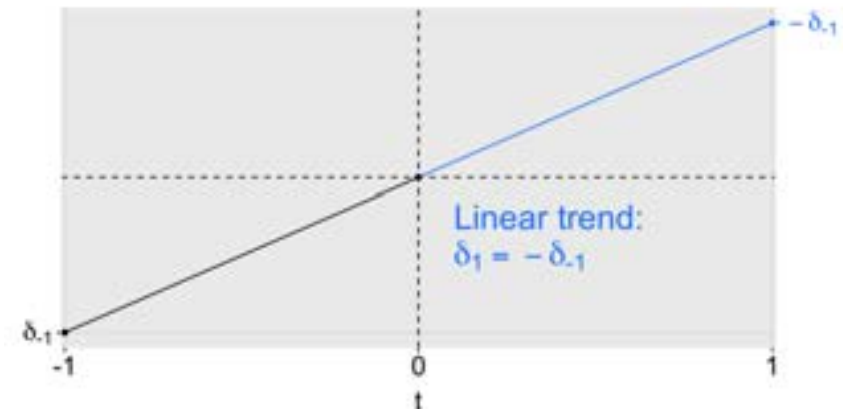
Rambachan and Roth (2020) suggestion

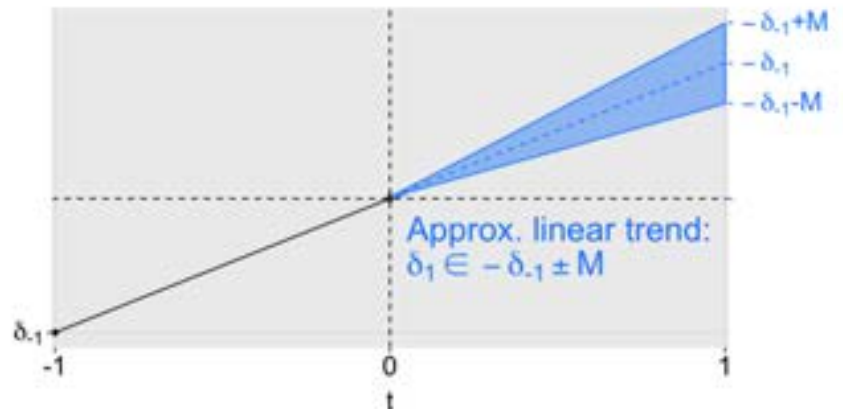
- Intuitive proposed solution for robustness. Note the post and pre

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

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

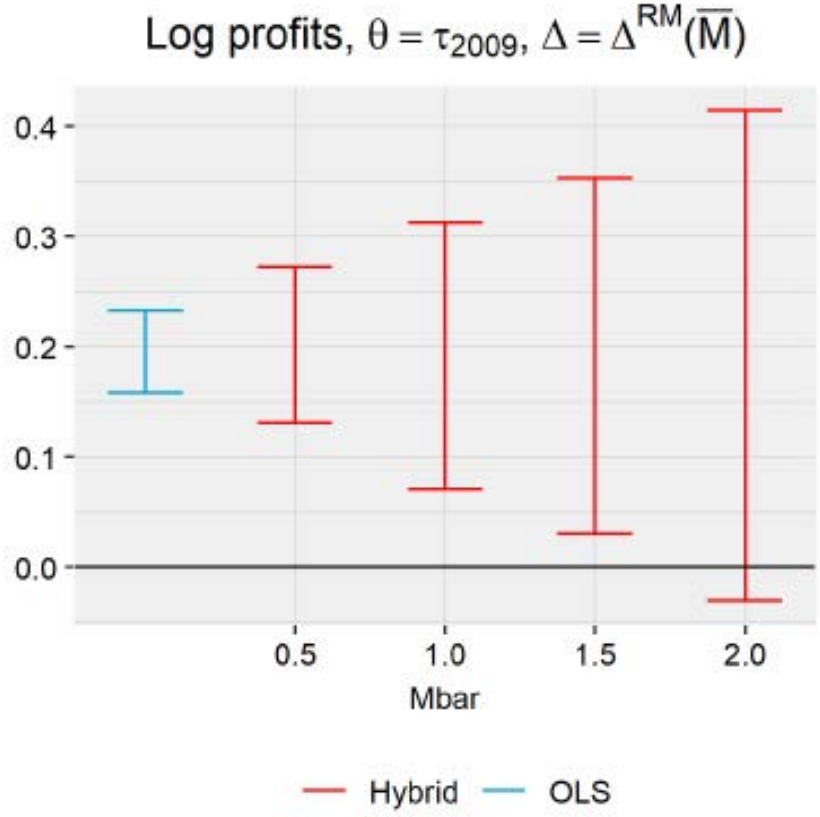
  effects:
- parallel trends assumes these δ are zero. But pre-trends may not be zero.
  - R&R say: we can use the info from our pre-trends to bound post-trend
  - Use a *smoothness* assumption, M, on the second derivative. E.g. simple case:





Rambachan and Roth (2020) suggestion

Using the honestdid package, we can find the largest M such that the 95% CI excludes zero while requiring $|\delta_1^+| \leq M|\delta_1^-|$



We can rule out a null effect unless we allow for violations of parallel trends that are twice as large than the max in the pre-period!

Checks for DD Design

- Very common for readers and others to request a variety of "robustness checks" from a DD design
- Think of these as along the same lines as the leads and lags we already discussed
  - Falsification test using data for prior periods (already discussed)
  - Falsification test using data for alternative control group
  - Falsification test using alternative "placebo" outcome that should not be affected by the treatment

Alternative control group – DDD

TABLE 3—DDD ESTIMATES OF THE IMPACT OF STATE MANDATES ON HOURLY WAGES

Location/year	Before law change	After law change	Time difference for location
A. *Treatment Individuals: Married Women, 20–40 Years Old:*			
Experimental states	1.547 (0.012) [1,400]	1.513 (0.012) [1,496]	−0.034 (0.017)
Nonexperimental states	1.369 (0.010) [1,480]	1.397 (0.010) [1,640]	0.028 (0.014)
Location difference at a point in time:	0.178 (0.016)	0.116 (0.015)	
Difference-in-difference:	−0.062 (0.022)		
B. *Control Group: Over 40 and Single Males 20–40:*			
Experimental states	1.759 (0.007) [5,624]	1.748 (0.007) [5,407]	−0.011 (0.010)
Nonexperimental states	1.630 (0.007) [4,959]	1.627 (0.007) [4,928]	−0.003 (0.010)
Location difference at a point in time:	0.129 (0.010)	0.121 (0.010)	
Difference-in-difference:	−0.008: (0.014)		
DDD:	**−0.054** **(0.026)**		

DDD in Regression

$$
\begin{aligned}
W_{ijt} &= \alpha + \beta_1 X_{ijt} + \beta_2 \tau_t + \beta_3 \delta_j + \beta_4 D_i + \beta_5 (\delta \times \tau)_{jt} \\
&\quad + \beta_6 (\tau \times D)_{ti} + \beta_7 (\delta \times D)_{ij} + \beta_8 (\delta \times \tau \times D)_{ijt} + \varepsilon_{ijt}
\end{aligned}
$$

- The DDD estimate is the difference between the DD of interest and a placebo DD (which is supposed to be zero)
- If the placebo DD is non-zero, it might be difficult to convince the reviewer that the DDD removed all the bias
- If the placebo DD is zero, then DD and DDD give the same results but DD is preferable because standard errors are smaller for DD than DDD

Standard errors in DD strategies

- If you are using panel data, outcomes and the treatment tend to be severely autocorrelated
- Using robust standard errors will lead to overrejection of the H0 due to downward biased standard errors (see Bertrand, Duflo, Mullainathan 2004)
- **If you have enough clusters, you must cluster on the unit of policy implementation.**
  - If the policy is implemented at the industry level, you should cluster at the industry and not at the firm level
- In the Card and Krueger study (1 treated, 1 control state), clustering at the state level is infeasible

Standard errors in DD strategies – few clusters

- Often, we have less than 42 clusters (MHE) in our diff-in-diff
- Solutions to calculate standard errors with few clusters have been proposed but this is an active field of research
- Highly context specific and often involves making trade-offs between (strong) assumptions
  - Donald and Lang (2007) two-step procedure
  - Conley and Taber (2010) if you have many control groups and few treated groups
  - (Wild) Cluster bootstrap (e.g., Cameron, Gelbach, Miller 2008)
  - Recent work by Andreas Hagemann

Plan for today

- Different cases of diff-in-diff
- Issues with staggered timing and TWFE
- Synthetic controls

Cases of DinD

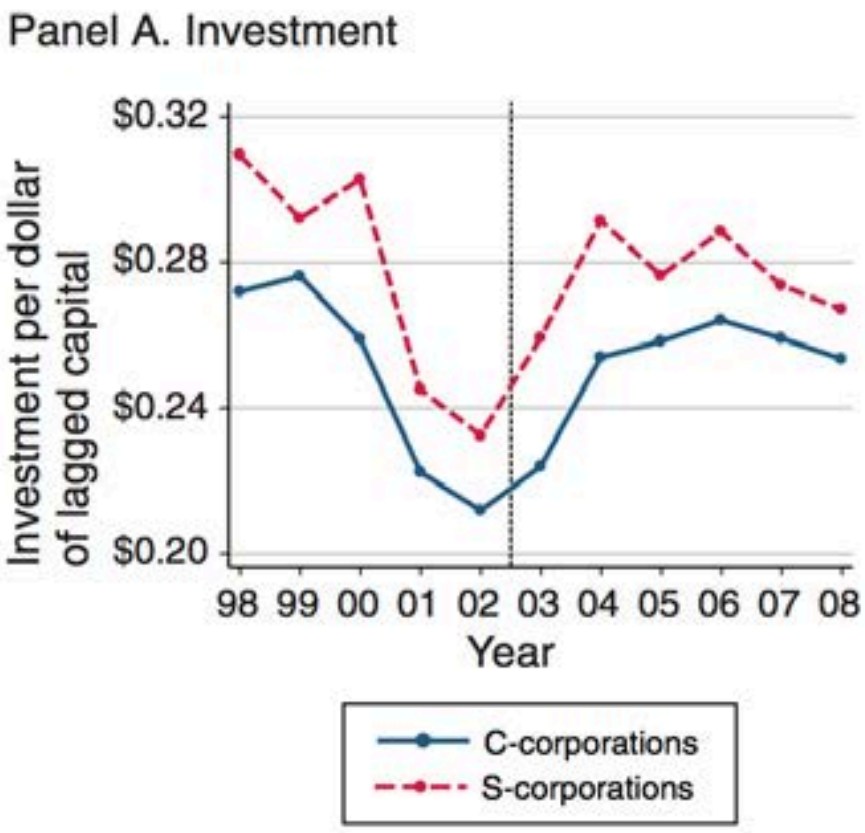
- 1 treatment timing, Binary treatment, 2 periods
  - Card and Krueger (AER, 1994)
- 1 treatment timing, Binary treatment, T periods
  - Yagan (AER, 2015)
- 1 treatment timing, Continuous treatment
  - Berger, Turner and Zwick (JF, 2020)
- Staggered treatment timing, Binary treatment
  - Bailey and Goodman-Bacon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

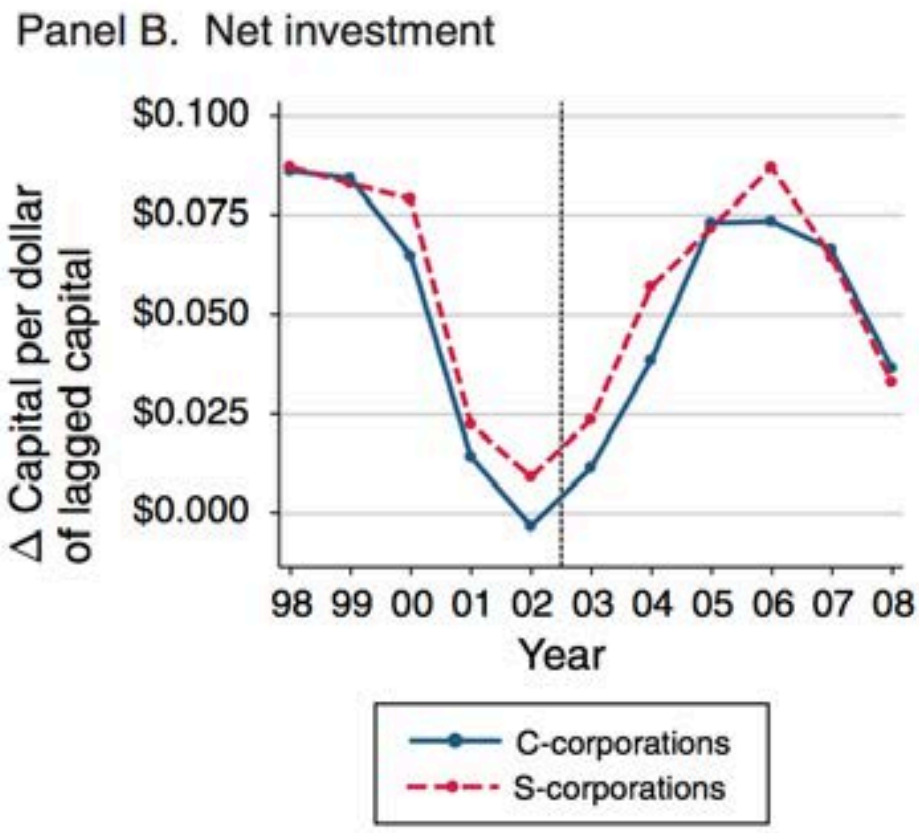
Yagan (2015)

- 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):

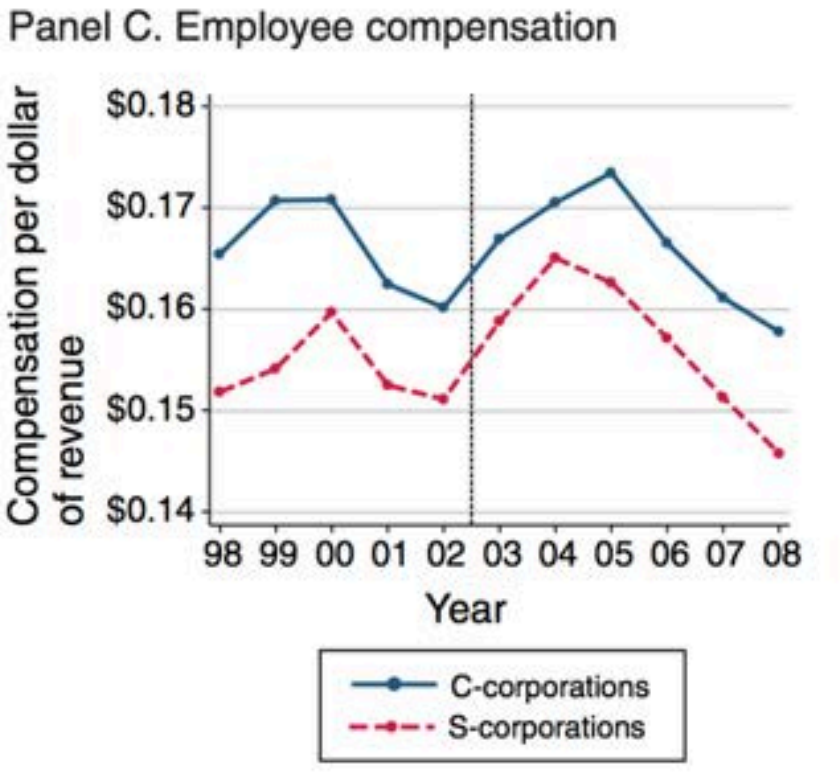
  The identifying assumption underlying this research design is not random assignment of C- versus S-status; it is that C- and S-corporation outcomes would have trended similarly in the absence of the tax cut.

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)

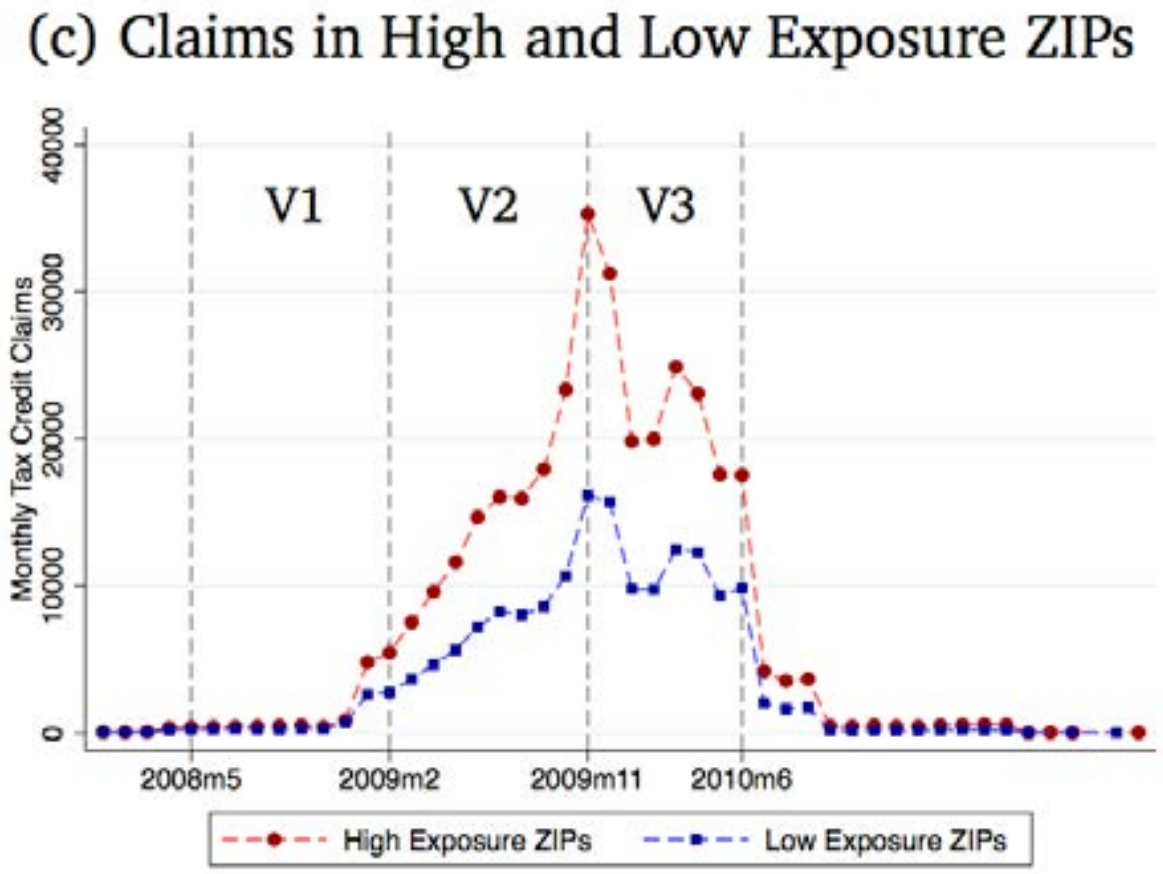
- This paper studies the impact of temporary fiscal stimulus (First-Time Home Buyer tax credit) on housing markets

- 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:

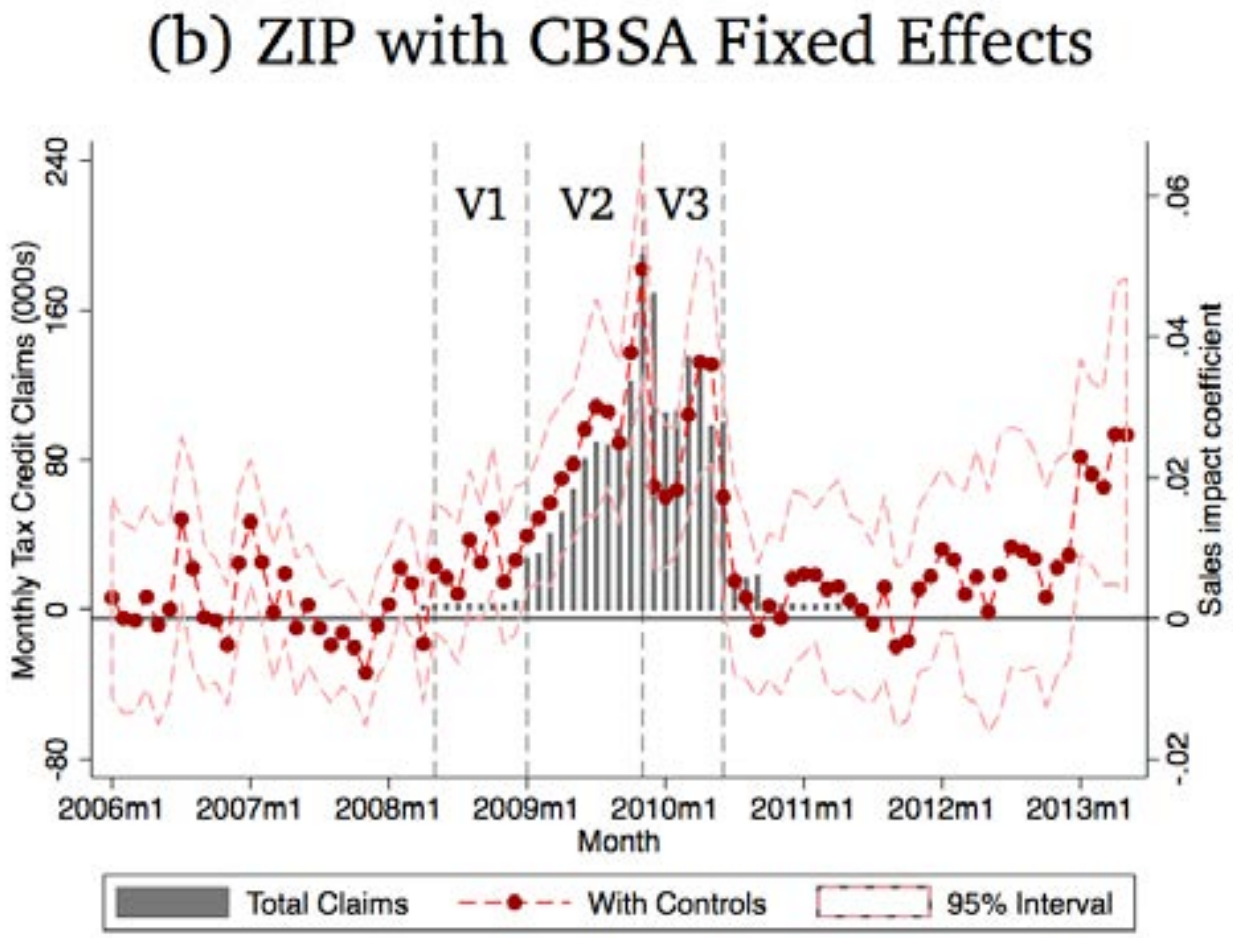
    The key threat to this design is the possibility that time-varying, pla
    specific shocks are correlated with our exposure measure.

- 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.

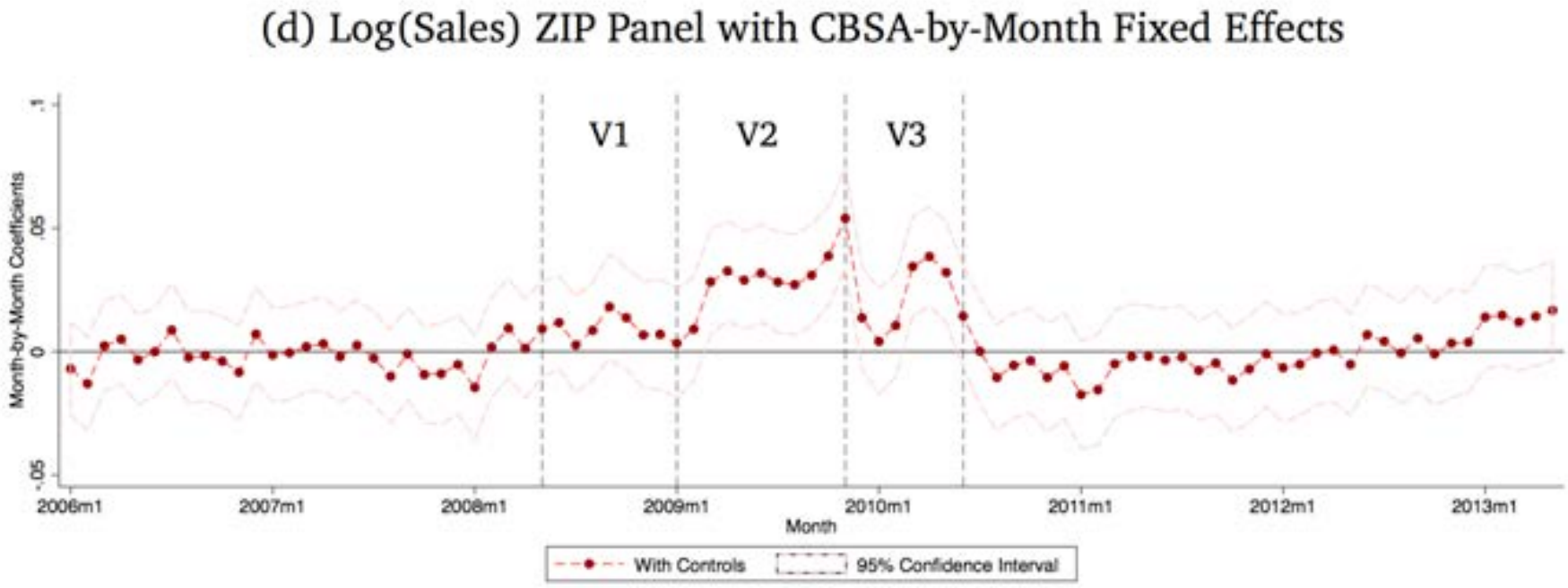
First stage: Binary approximation



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.

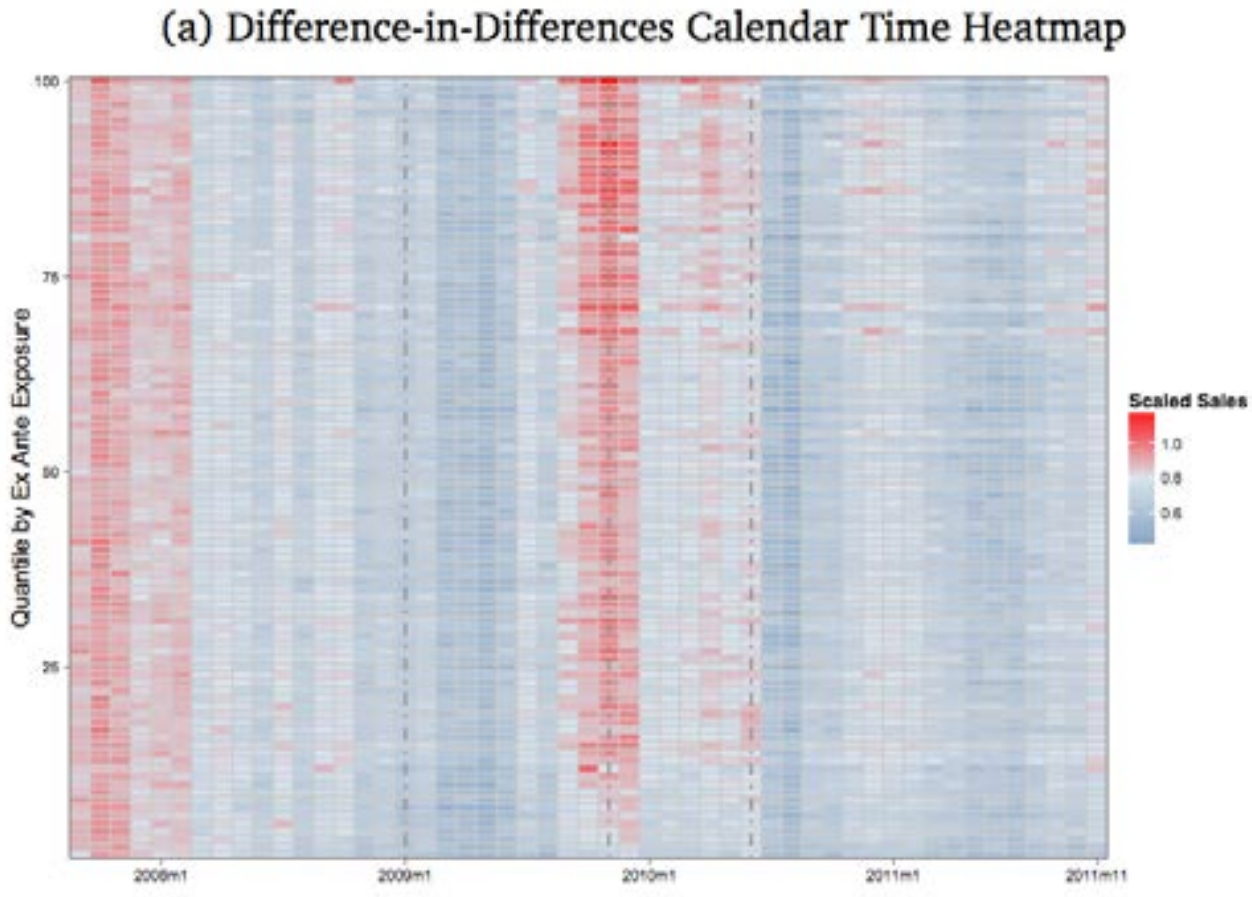
$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=1, t \neq t_0}^{T} \delta_t D_{it} + \epsilon_{it}, \tag{1}$$

- 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:

$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=1, t \neq t_0}^{T} \sum_{k=1}^{4} \delta_{t,k} \widetilde{D}_{it,k} + \epsilon_{it}. \tag{2}$$

  - 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 where there is capacities for it (e.g., medical schools)

- 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

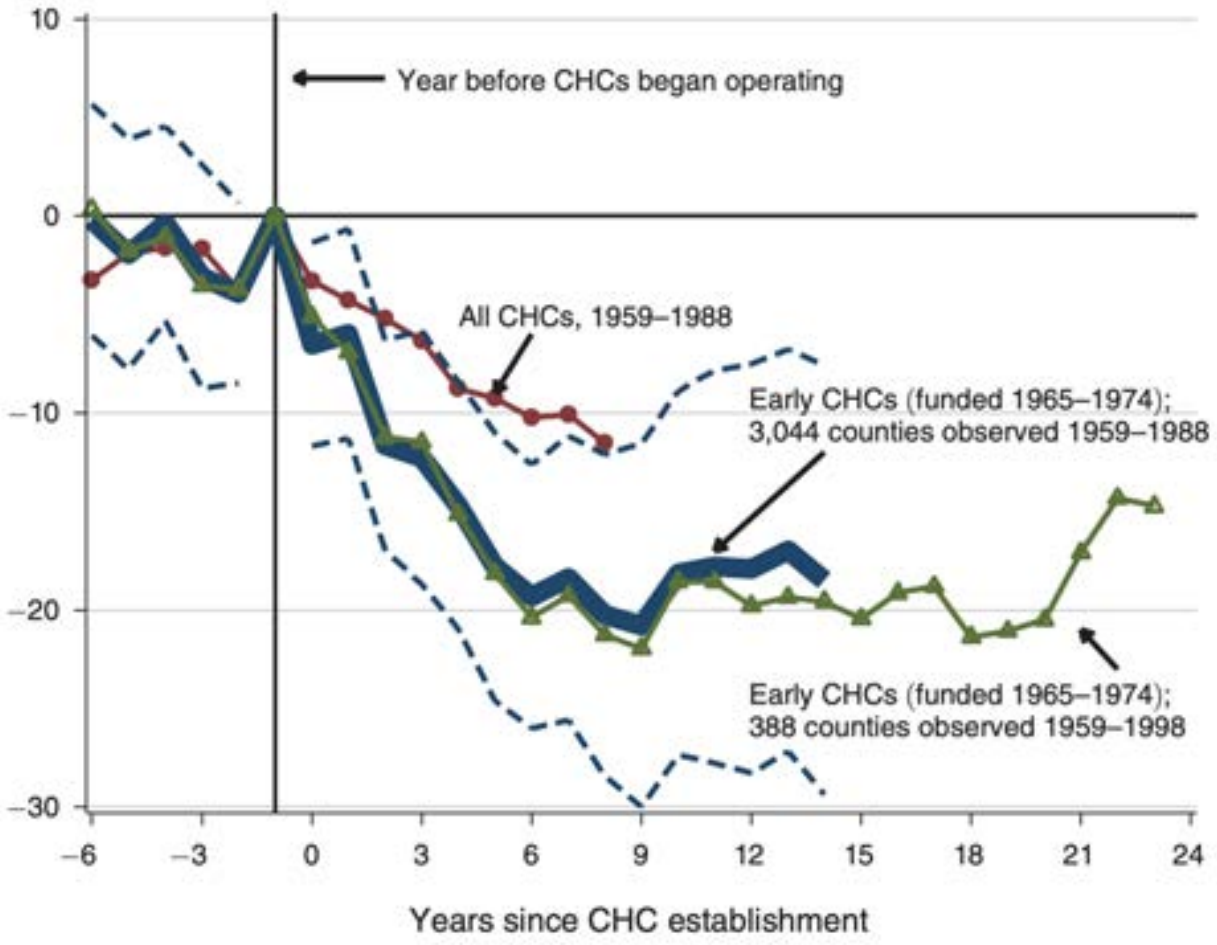
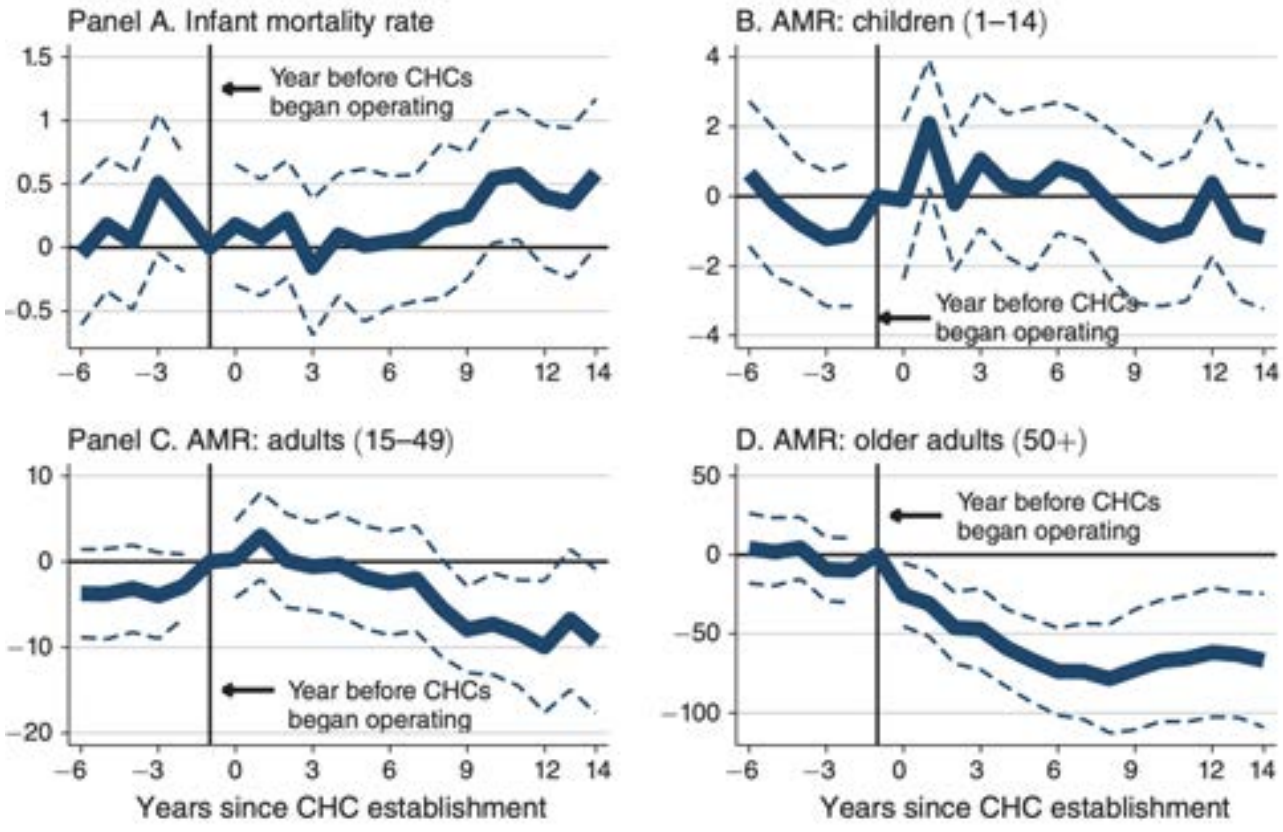


FIGURE 5. THE RELATIONSHIP BETWEEN COMMUNITY HEALTH CENTERS AND MORTALITY RATES

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

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

$$y_{it} = \alpha_i + \alpha_t + \beta^{DD} D_{it} + \epsilon_{it} \tag{3}$$

  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:

$$\tau_{ATT}(g,t) = E(Y_{it}(1) - Y_{it}(0)|D_{it} = 1 \forall t \geq g), \tag{4}$$

  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", $G_i = \infty$
  2. Parallel trends of treatment group with the group of the "not yet treated", $G_i \geq t + 1$

What happened to TWFE?

- It turns out that the logic of the TWFE does not naturally extend to differential timings
  - Regression does a variance weighted approximation (MHE, 3.3.1):

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

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

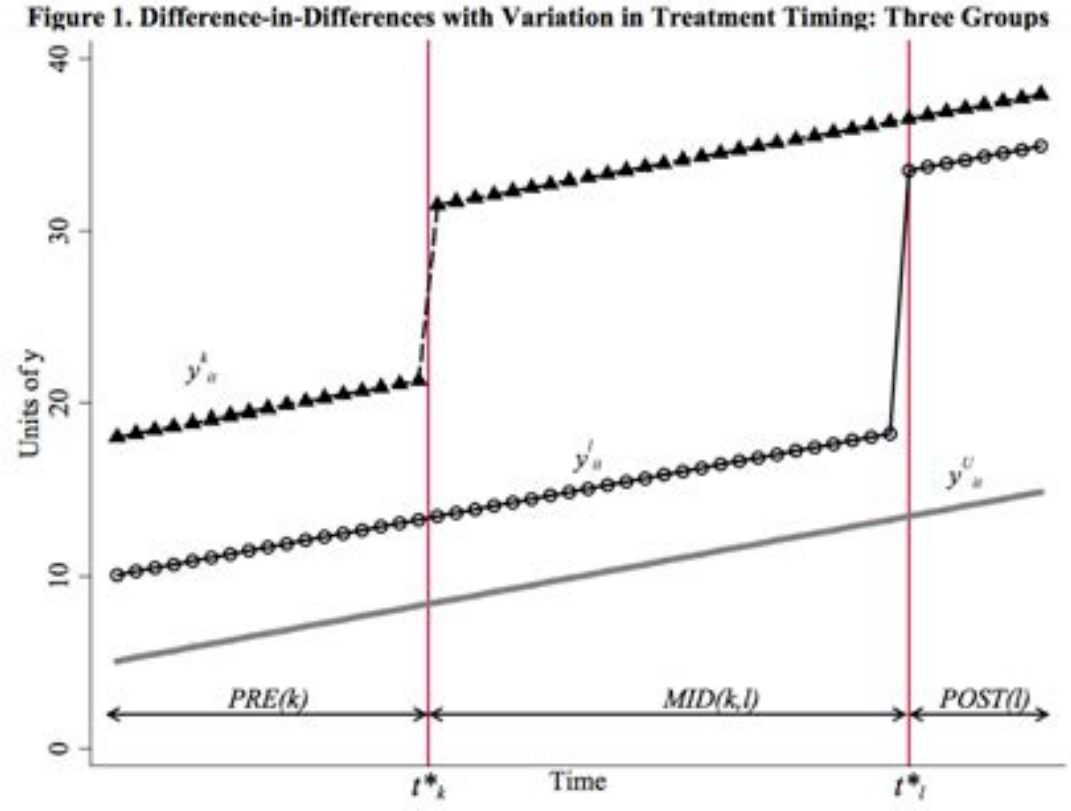
- Key insight from several papers: with staggered timings + heterogeneous effects, the TWFE approach to DinD can put negative weight on certain groups' TE
  - Serious issue for interpretability

- This is *solveable*. Merely a construct of being overly casual with estimator definition

What happened to TWFE?

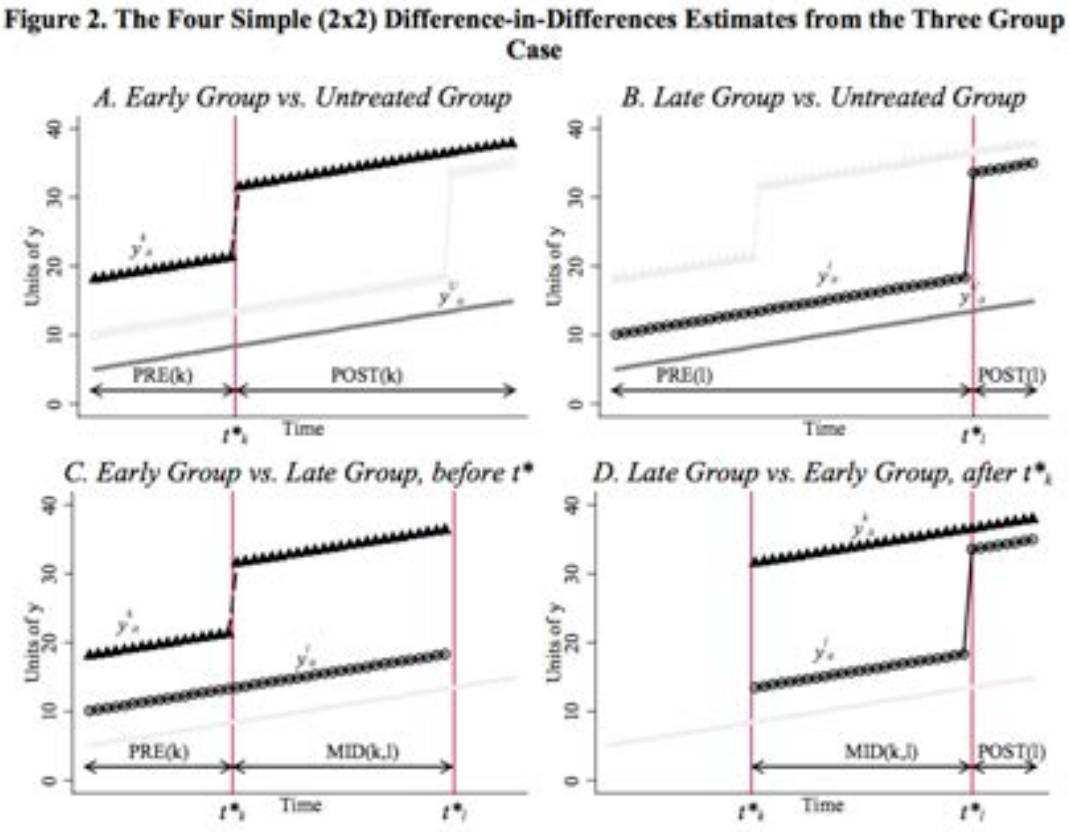
Goodman-Bacon 2x2 comparisons

- Consider two staggered treatments 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!)

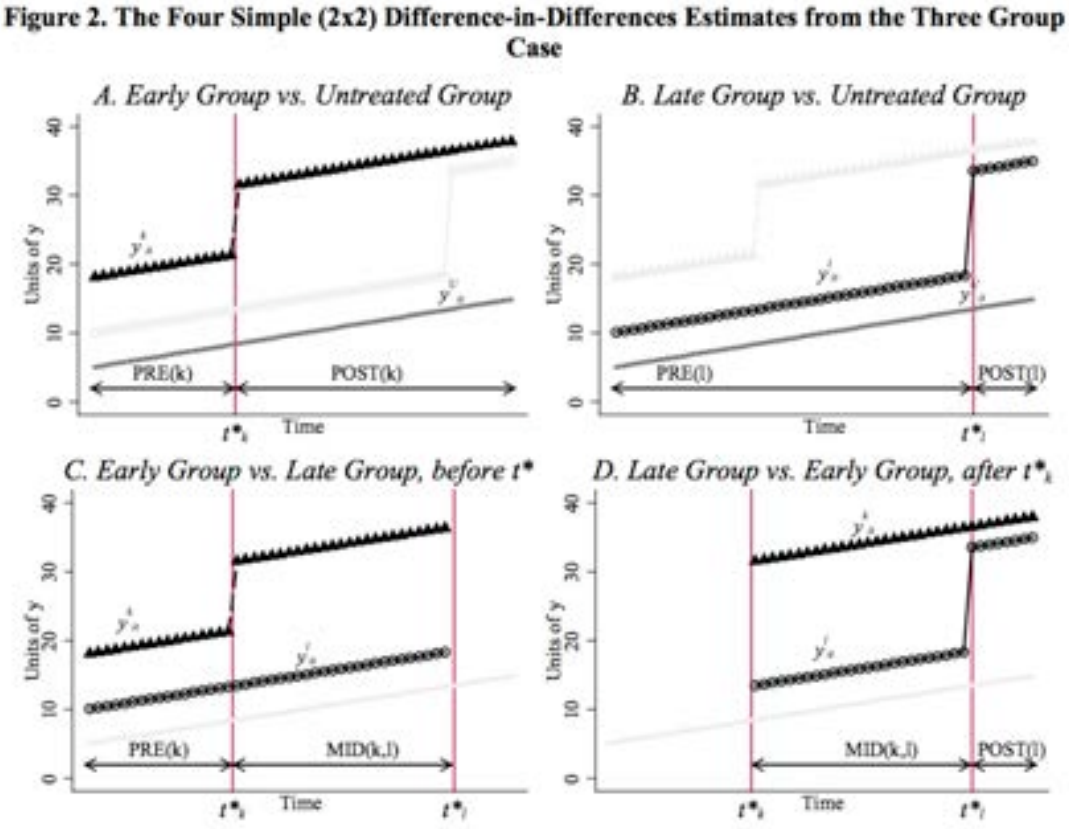


Goodman-Bacon 2x2 comparisons

- 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



Two-Way Fixed Effects Estimator

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

Unit fixed effects

Time fixed effects

Treatment dummy

What is $\widehat{\boldsymbol{\beta}}^{\boldsymbol{DD}}$?

What is $\widehat{\beta}^{DD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

1. Partial out fixed effects (Frisch-Waugh):

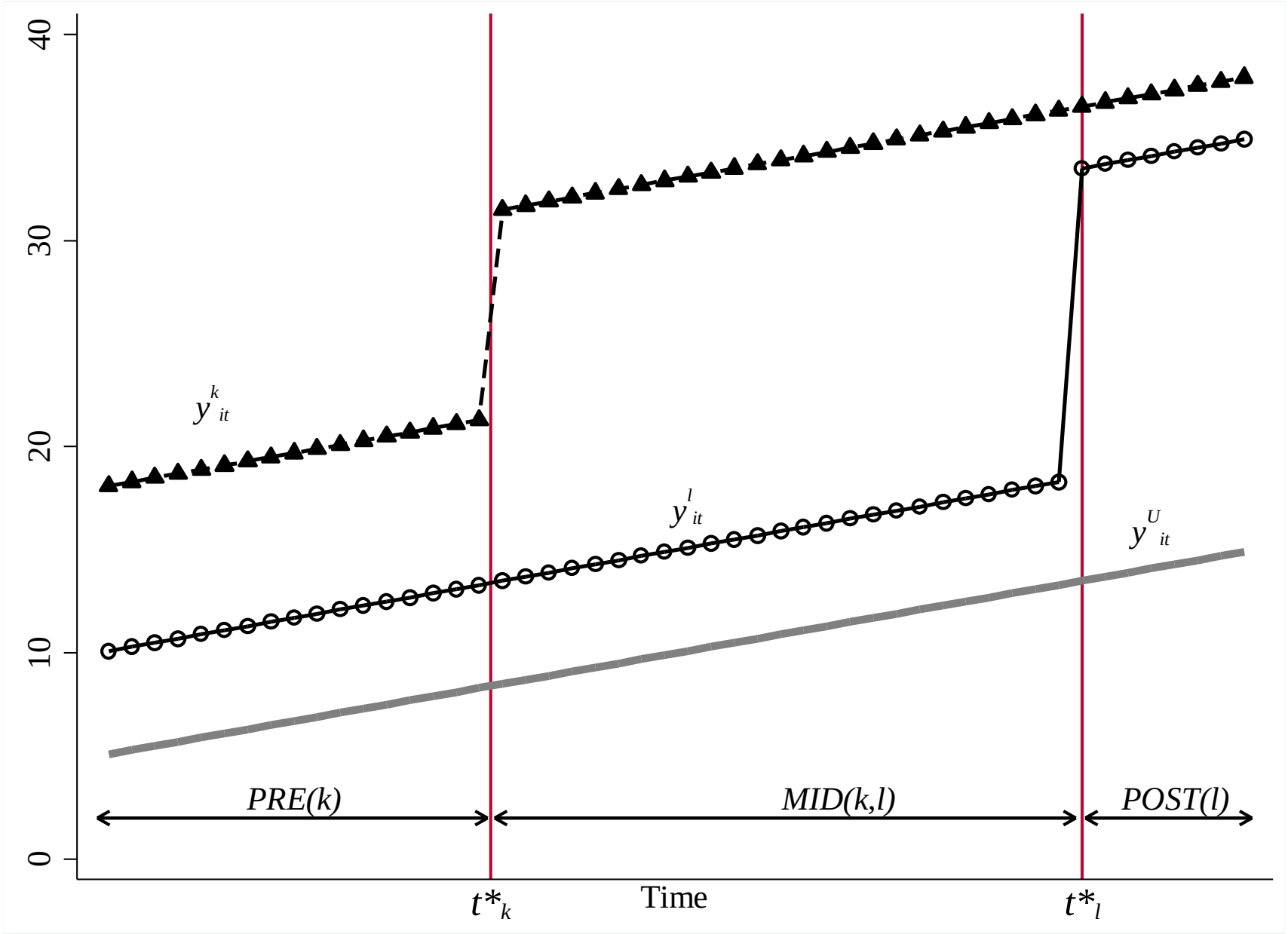
$$\widetilde{D}_{it} = \left(D_{it} - \overline{\overline{D}}\right) - \left(\overline{D}_i - \overline{\overline{D}}\right) - \left(\overline{D}_t - \overline{\overline{D}}\right)$$

$$\tilde{y}_{it} = (y_{it} - \overline{\bar{y}}) - (\bar{y}_i - \bar{y}) - (\bar{y}_t - \overline{\bar{y}})$$

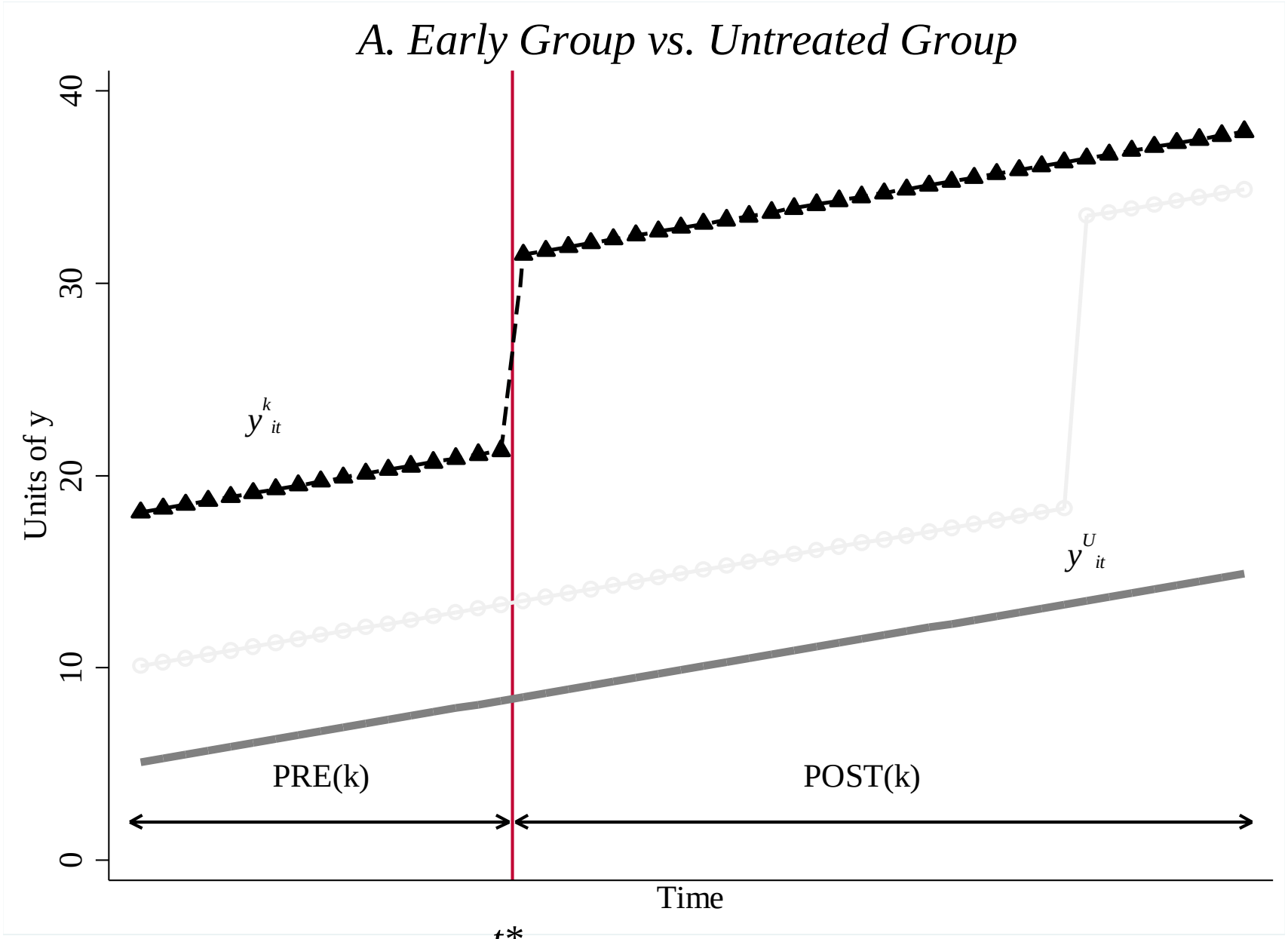
2. Calculate univariate coefficient by brute force:

$$\hat{\beta}^{DD} = \frac{\widehat{cov}(\widetilde{D}_{it}, \tilde{y}_{it})}{\widehat{V}(\widetilde{D}_{it})} = \frac{\frac{1}{NT}\Sigma_i \Sigma_t (y_{it} - \overline{\bar{y}})\left(D_{it} - \overline{\overline{D}}\right)}{\frac{1}{NT}\Sigma_i \Sigma_t \left(D_{it} - \overline{\overline{D}}\right)}$$

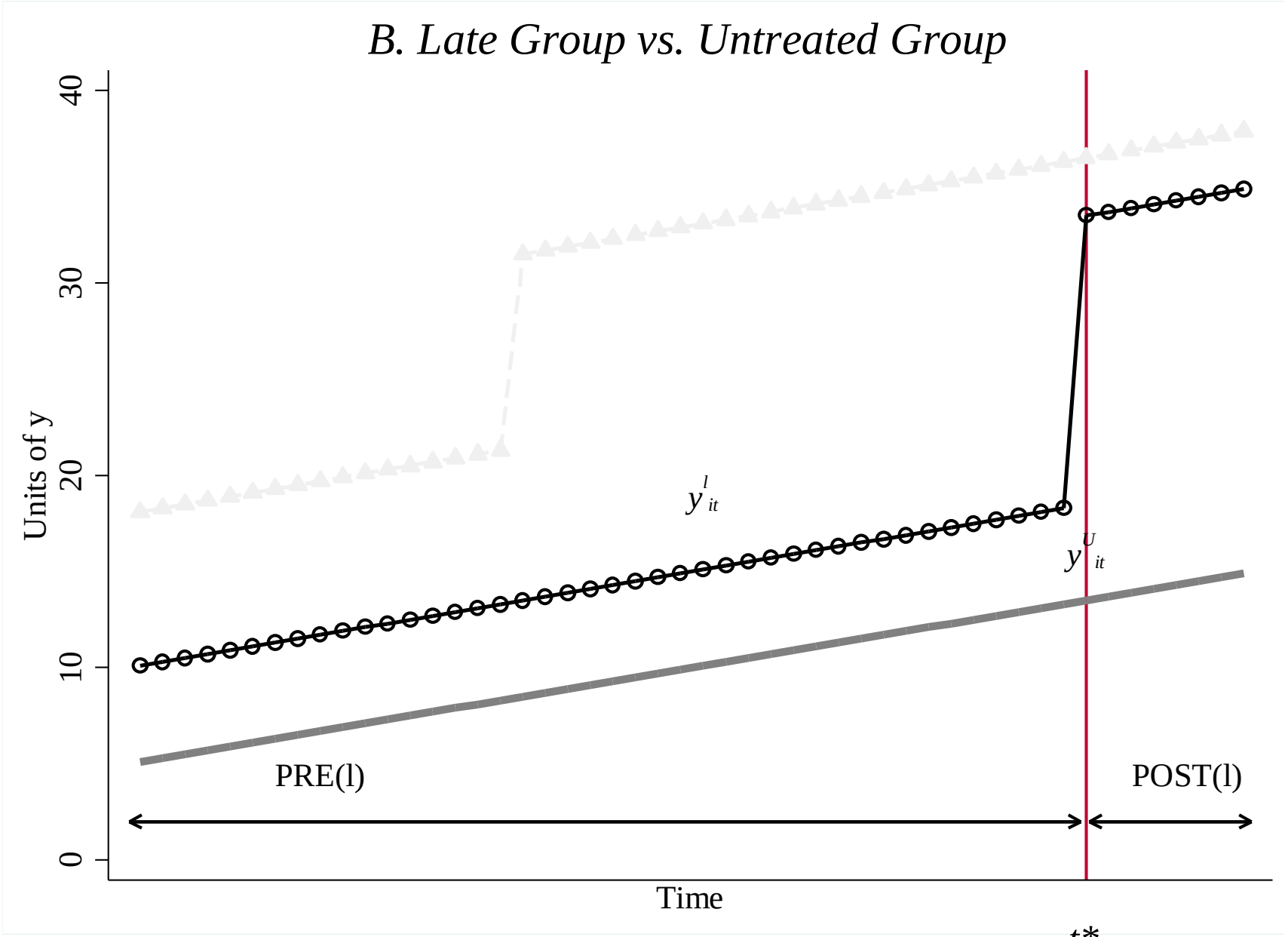
$\widehat{\beta}^{DD}$?



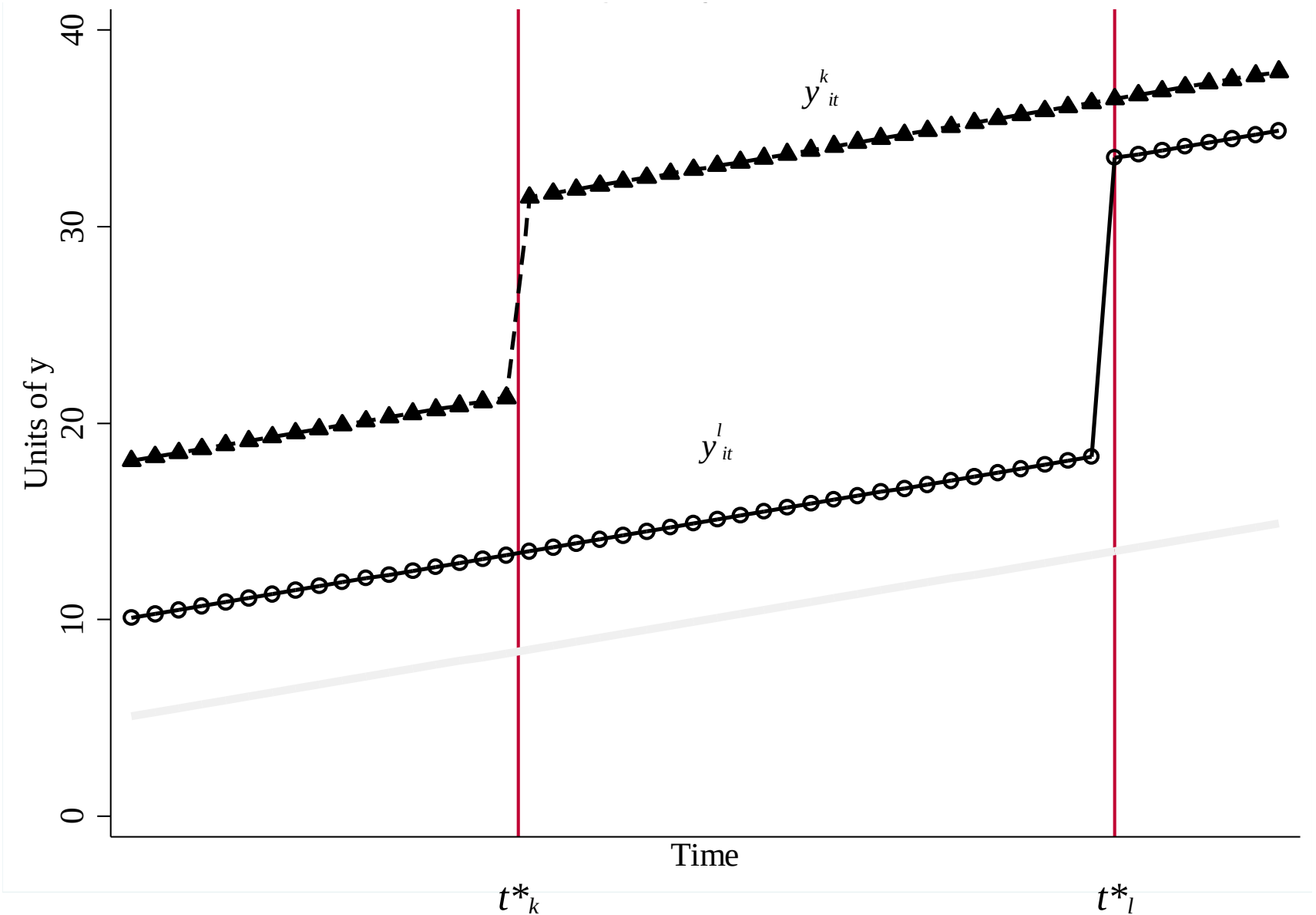
$$\widehat{\beta}_{kU}^{DD}$$



$$\widehat{\beta}^{DD}_{\ell U}$$

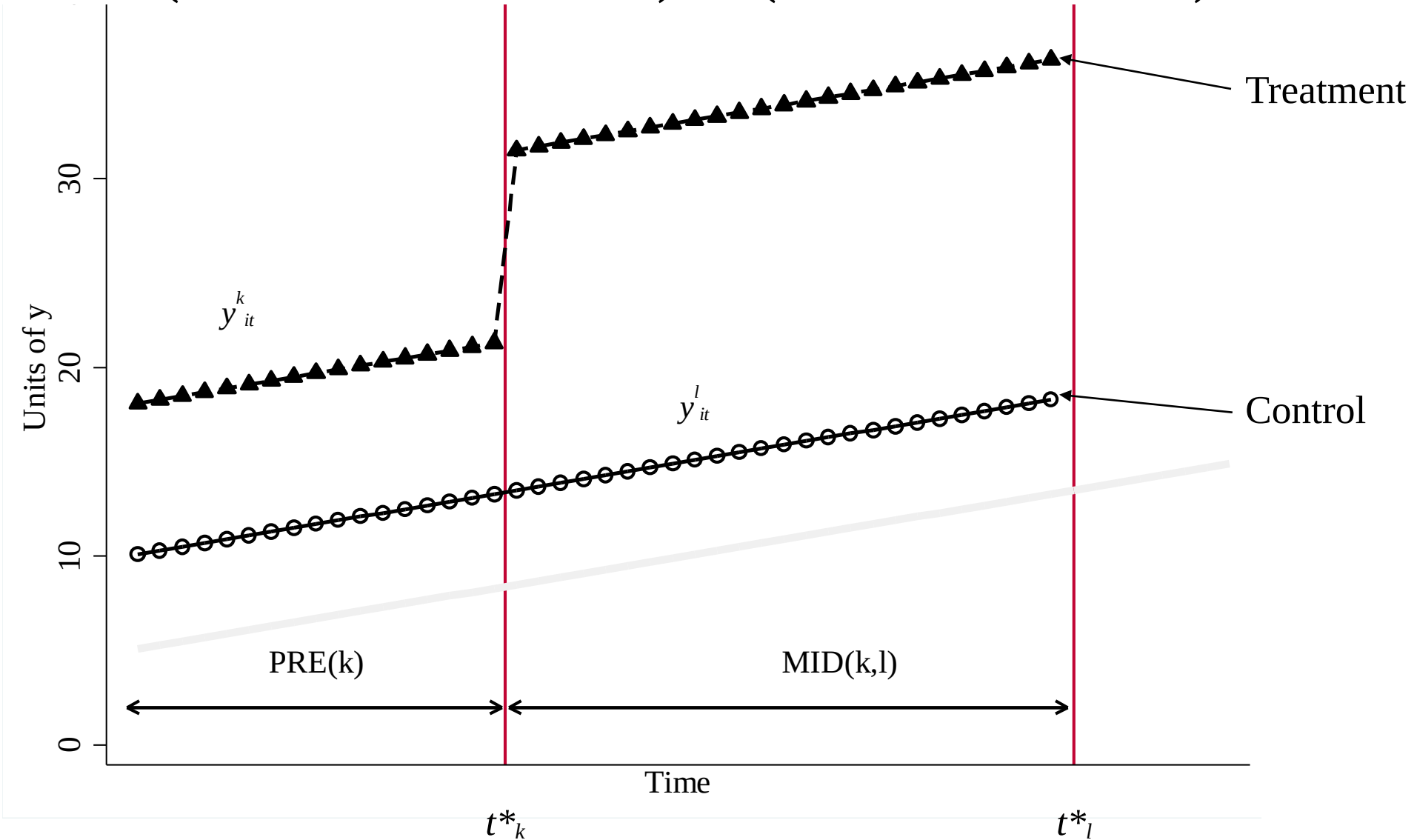


Two-Group Timing-Only Estimator ($\widehat{\beta}^{DD}_{k\ell}$)



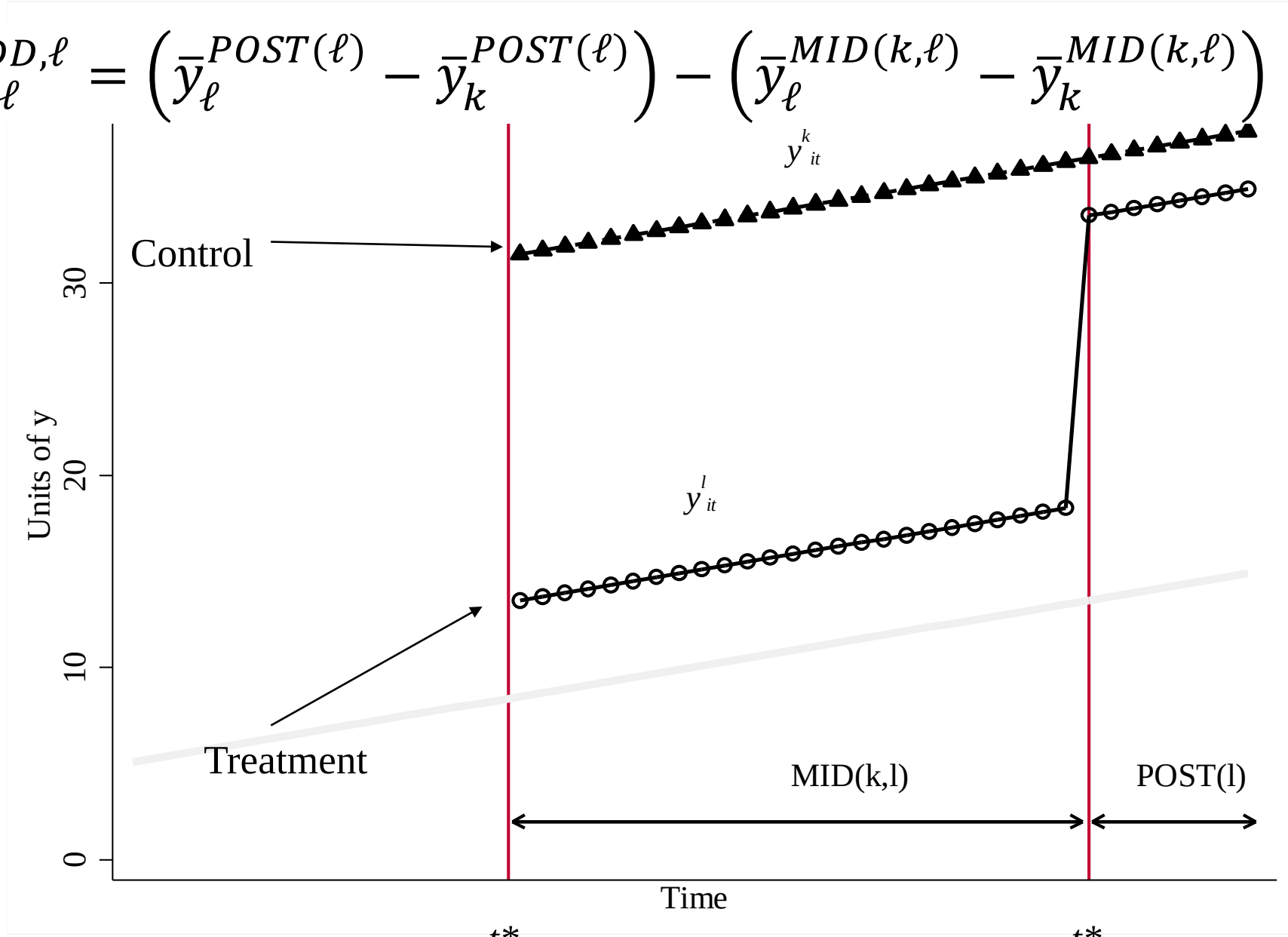
$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{k\ell}}^{\boldsymbol{DD,k}}$$

$$\hat{\beta}_{k\ell}^{DD,k} = \left(\bar{y}_k^{MID(k,\ell)} - \bar{y}_\ell^{MID(k,\ell)}\right) - \left(\bar{y}_k^{PRE(k)} - \bar{y}_\ell^{PRE(k)}\right)$$



$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{k\ell}}^{\boldsymbol{DD,\ell}}$$

$$\hat{\beta}_{k\ell}^{DD,\ell} = \left(\bar{y}_{\ell}^{POST(\ell)} - \bar{y}_{k}^{POST(\ell)}\right) - \left(\bar{y}_{\ell}^{MID(k,\ell)} - \bar{y}_{k}^{MID(k,\ell)}\right)$$



What is $\widehat{\beta}^{DD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

For three groups:

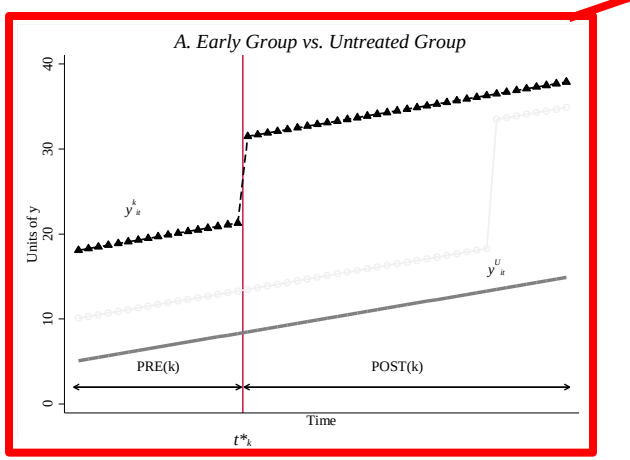
$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}^{DD}_{kU} + s_{\ell U}\hat{\beta}^{DD}_{\ell U} + \left[s^{k}_{k\ell}\hat{\beta}^{DD,k}_{k\ell} + s^{\ell}_{k\ell}\hat{\beta}^{DD,\ell}_{k\ell}\right]$$

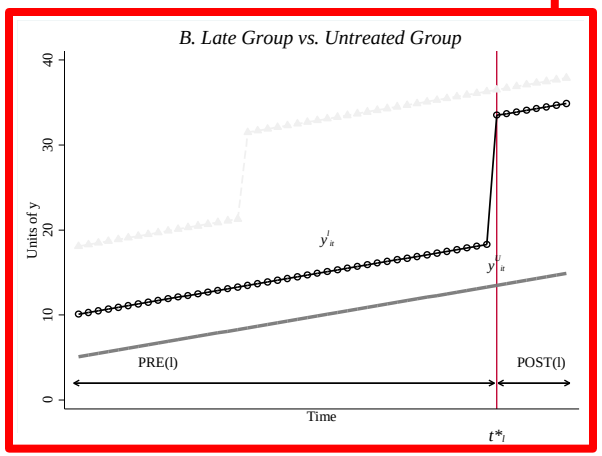
What is $\widehat{\beta}^{DD}$?

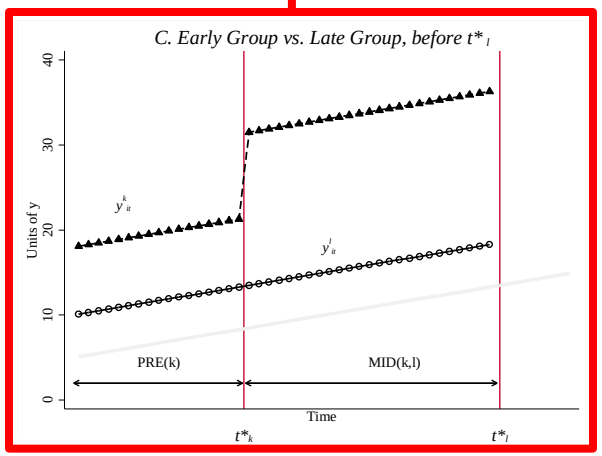
$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

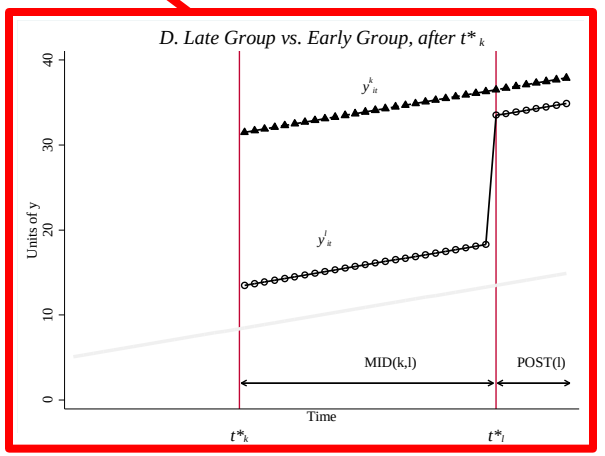
For three groups:

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}^{DD}_{kU} + s_{\ell U}\hat{\beta}^{DD}_{\ell U} + \left[s^{k}_{k\ell}\hat{\beta}^{DD,k}_{k\ell} + s^{\ell}_{k\ell}\hat{\beta}^{DD,\ell}_{k\ell}\right]$$









What is $\hat{\beta}^{DD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

For three groups:

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}^{DD}_{kU} + s_{\ell U}\hat{\beta}^{DD}_{\ell U} + \left[s^{k}_{k\ell}\hat{\beta}^{DD,k}_{k\ell} + s^{\ell}_{k\ell}\hat{\beta}^{DD,\ell}_{k\ell}\right]$$

$$s_{kU} = \frac{(n_k + n_U)^2 n_{kU}(1 - n_{kU})\bar{D}_k(1 - \bar{D}_k)}{V(\tilde{D}_{it})}$$

Sample size^2

$$s^{k}_{k\ell} = \frac{\left((n_k + n_\ell)(1 - \bar{D}_\ell)\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell}\frac{1 - \bar{D}_k}{1 - \bar{D}_\ell}}{V(\tilde{D}_{it})}$$

$$s^{\ell}_{k\ell} = \frac{\left((n_k + n_\ell)\bar{D}_k\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{\bar{D}_k}\frac{\bar{D}_\ell}{\bar{D}_k}}{V(\tilde{D}_{it})}$$

What is $\widehat{\beta}^{DD}$?

$$y_{it} = \alpha_i + \alpha_t + \hat{\beta}^{DD} D_{it} + u_{it}$$

For three groups:

$$\hat{\beta}^{DD} = s_{kU}\hat{\beta}^{DD}_{kU} + s_{\ell U}\hat{\beta}^{DD}_{\ell U} + \left[s^{k}_{k\ell}\hat{\beta}^{DD,k}_{k\ell} + s^{\ell}_{k\ell}\hat{\beta}^{DD,\ell}_{k\ell}\right]$$

$$s_{kU} = \frac{(n_k + n_U)^2 n_{kU}(1 - n_{kU})\bar{D}_k(1 - \bar{D}_k)}{V(D_{it})}$$

$$s^{k}_{k\ell} = \frac{\left((n_k + n_\ell)(1 - \bar{D}_\ell)\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{1 - \bar{D}_\ell}\frac{1 - \bar{D}_k}{1 - \bar{D}_\ell}}{V(\tilde{D}_{it})}$$

$$s^{\ell}_{k\ell} = \frac{\left((n_k + n_\ell)\bar{D}_k\right)^2 n_{k\ell}(1 - n_{k\ell})\frac{\bar{D}_k - \bar{D}_\ell}{\bar{D}_k}\frac{\bar{D}_\ell}{\bar{D}_k}}{V(\tilde{D}_{it})}$$

Subsample variance of treatment

Difference-in-Differences Decomposition Theorem

Assume that there are $k = 1, \ldots, K$ groups of treated units ordered by treatment time t_k^* and one control group, U, which does not receive treatment in the data. The share of units in group k is n_k, and the share of periods that group k spends under treatment is $\overline{D}_k$. The DD estimate from a two-way fixed effects model is a weighted average all two-group DD estimators:

$$\hat{\beta}^{DD} = \sum_{k \neq U} s_{kU}\, \hat{\beta}_{kU}^{DD} + \sum_{k \neq U} \sum_{\ell > k} \left[s_{k\ell}^{k} \hat{\beta}_{k\ell}^{DD,k} + s_{k\ell}^{\ell} \hat{\beta}_{k\ell}^{DD,\ell} \right]$$

With weights equal to:

$$s_{kU} = \frac{(n_k + n_U)^2\, \hat{V}_{kU}^{D}}{V\left(\widetilde{D}_{it}\right)}$$

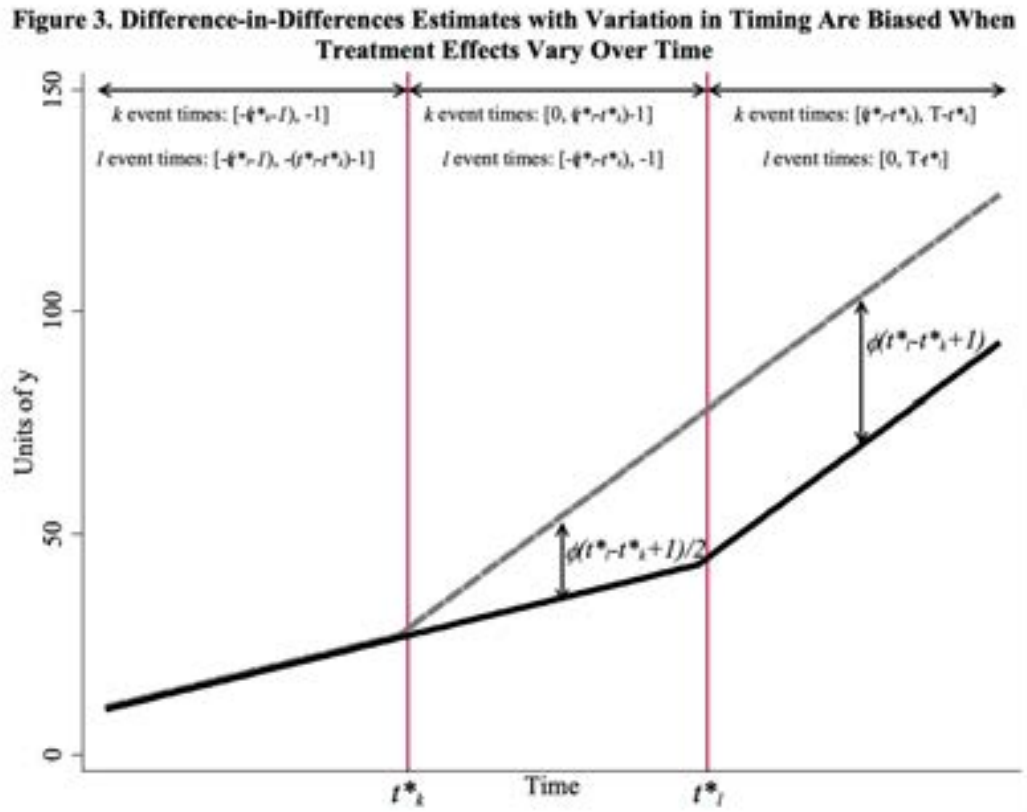
$$s_{k\ell}^{k} = \frac{\left((n_k + n_\ell)(1 - \overline{D}_\ell)\right)^2 \hat{V}_{k\ell}^{D,k}}{V\left(\widetilde{D}_{it}\right)}$$

$$s_{k\ell}^{\ell} = \frac{\left((n_k + n_\ell)\overline{D}_k\right)^2 \hat{V}_{k\ell}^{D,\ell}}{V\left(\widetilde{D}_{it}\right)}$$

$\sum_{k \neq U} s_{kU} + \sum_{k \neq U} \sum_{\ell > k} \left[s_{k\ell}^{k} + s_{k\ell}^{\ell} \right] = 1.$

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



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 different approaches proposed in the literature that are just as good!
  - E.g. de Chaisemartin and d'Haultfoeuille (2020), Callaway and Sant'anna (2020)
- These all are robust to this issue. If treatment is not absorbing, de Chaisemartin and d'Haultfoeuille (2020) preferable.
- 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

Synthetic controls

- In diff-and-diff, we estimate the counterfactual by asserting the linear model:

$$Y_{it} = \alpha_i + \gamma_t + D_{it}\tau + \epsilon_{it}$$

- Instead of imposing the parallel trends assumption directly through the linear model, we could construct a combination of units to approximate $Y_{it}(0)$
  - This is what one does in the cross-sectional setting with matching methods! E.g. consider the ATT:

$$\tau_{ATT} = \underbrace{Y(1)}_{\text{Fully observed}} - \underbrace{\hat{Y}(0)}_{\text{Constructed}}$$

- How would one pick? Recall that with p-score methods or regression, weights effectively reweight based on comparability to treated group
  - With panel data, can use pre-treatment data to construct these weights
  - This method is known as synthetic control (and its various descendents)

Synthetic Control example - (Abadie et al. 2010))

- Consider following problem: California bans smoking in 1989. What does that do to smoking?
  - Define estimand: $\tau_{ban,CA} = Y_{california,post}(1) - Y_{california,post}(0)$
  - This is the effect of the *California* smoking ban
  - How can we get at it?
- We need a "synthetic California" as our control
- In an ideal world, the average of the other states would work – however, not clear empirically that they are a good counterfactual

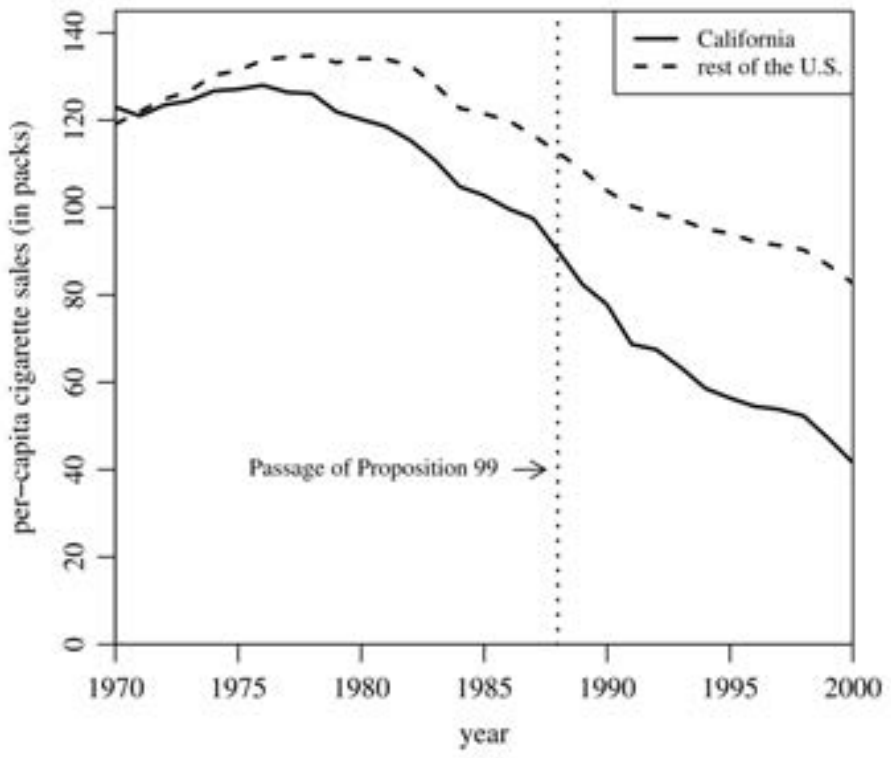


Figure 1. Trends in per-capita cigarette sales: California vs. the rest of the United States.

The synthetic control method (Abadie et al. 2010)

- Following Doudchenko and Imbens (2018), estimators of the following form nest DiD and synthetic control:

$$\hat{Y}_{t,post}(0) = \mu + \sum_{i \in c} \omega_i Y_{i,T}$$

  - A constant μ allows for very different averages (common in diff-in-diff)
  - Weights are allowed to vary across i – a simple average would be diff-in-diff

- In ADH, they impose
  1. $\mu = 0$
  2. $\sum_i \omega_i = 1$
  3. $\omega_i \geq 0 \; \forall i$

- These three restrictions create a counterfactual California whose outcomes are within the support of the other states, and is a weighted sum of a subset of states

The synthetic control method (Abadie et al. 2010)

$$\hat{Y}_{t,post}(0) = \mu + \sum_{i \in c} \omega_i Y_{i,T}$$

- Formally, the ω_i need to be estimated, and are constructed by minimizing the distance between covariates in the pre-period:

$$\{\hat{\omega}\}_i = \arg\min_{\boldsymbol{W}} ||\boldsymbol{X}_{\text{treat}} - \boldsymbol{X}_{\text{control}} \boldsymbol{W}||$$

- The crucial piece tying this together: $\boldsymbol{X}$ can include both lagged outcomes, and covariates.

- Note we can now re-envision our panel data:
  - Observed outcomes: $\mathbf{Y}_{t,post}(1)$, $\mathbf{Y}_{c,post}(0)$
  - Observed covariates / predictors: $\mathbf{Y}_{t,pre}(0)$, $\mathbf{Y}_{c,pre}(0)$, $\mathbf{X}_t$, $\mathbf{X}_c$

- In many ways, this is just a matching problem using many characteristics!

The synthetic control method (Abadie et al. 2010)

Table 1. Cigarette sales predictor means

Variables	California		Average of 38 control states
	Real	Synthetic	
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

NOTE: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988). GDP per capita is measured in 1997 dollars, retail prices are measured in cents, beer consumption is measured in gallons, and cigarette sales are measured in packs.

Table 2. State weights in the synthetic California

State	Weight	State	Weight
Alabama	0	Montana	0.199
Alaska	–	Nebraska	0
Arizona	–	Nevada	0.234
Arkansas	0	New Hampshire	0
Colorado	0.164	New Jersey	–
Connecticut	0.069	New Mexico	0
Delaware	0	New York	–
District of Columbia	–	North Carolina	0
Florida	–	North Dakota	0
Georgia	0	Ohio	0
Hawaii	–	Oklahoma	0
Idaho	0	Oregon	–
Illinois	0	Pennsylvania	0
Indiana	0	Rhode Island	0
Iowa	0	South Carolina	0
Kansas	0	South Dakota	0
Kentucky	0	Tennessee	0
Louisiana	0	Texas	0
Maine	0	Utah	0.334
Maryland	–	Vermont	0
Massachusetts	–	Virginia	0
Michigan	–	Washington	–
Minnesota	0	West Virginia	0
Mississippi	0	Wisconsin	0
Missouri	0	Wyoming	0

The synthetic control method (Abadie et al. 2010)

- This approach can be incredibly successful

  By careful construction of a synthetic control, can calculate counterfactual impacts due to policy

  Still subject to same caveats from DinD – not invariant to some transformations (e.g. log and linear)

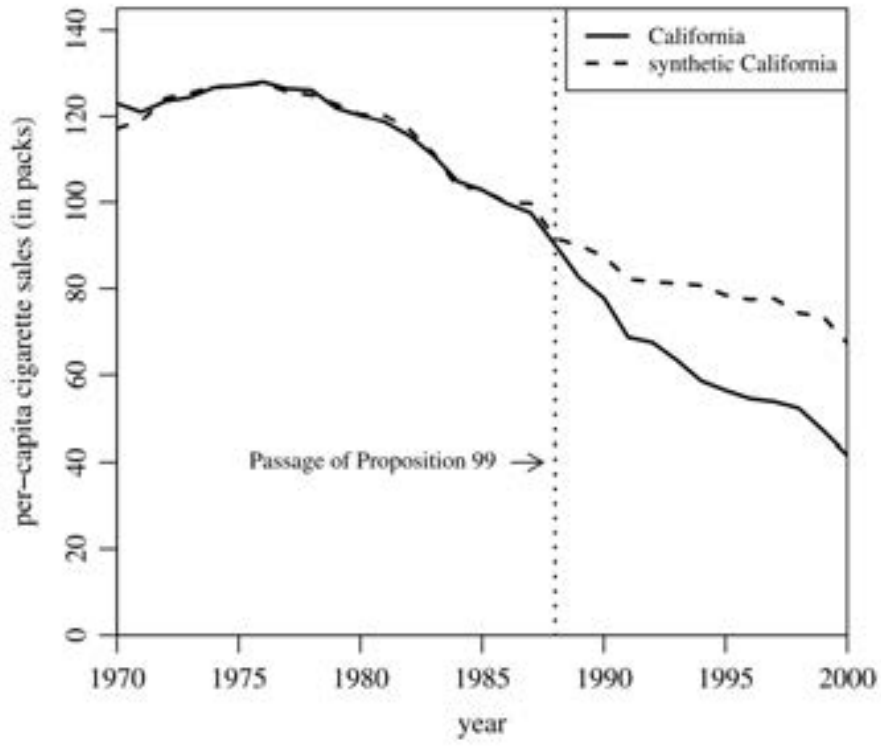


Figure 2. Trends in per-capita cigarette sales: California vs. synthetic California.

The synthetic control method (Abadie et al. 2010)

- This approach can be incredibly successful
- By careful construction of a synthetic control, can calculate counterfactual impacts due to policy
- Still subject to same caveats from DinD – not invariant to some transformations (e.g. log and linear)

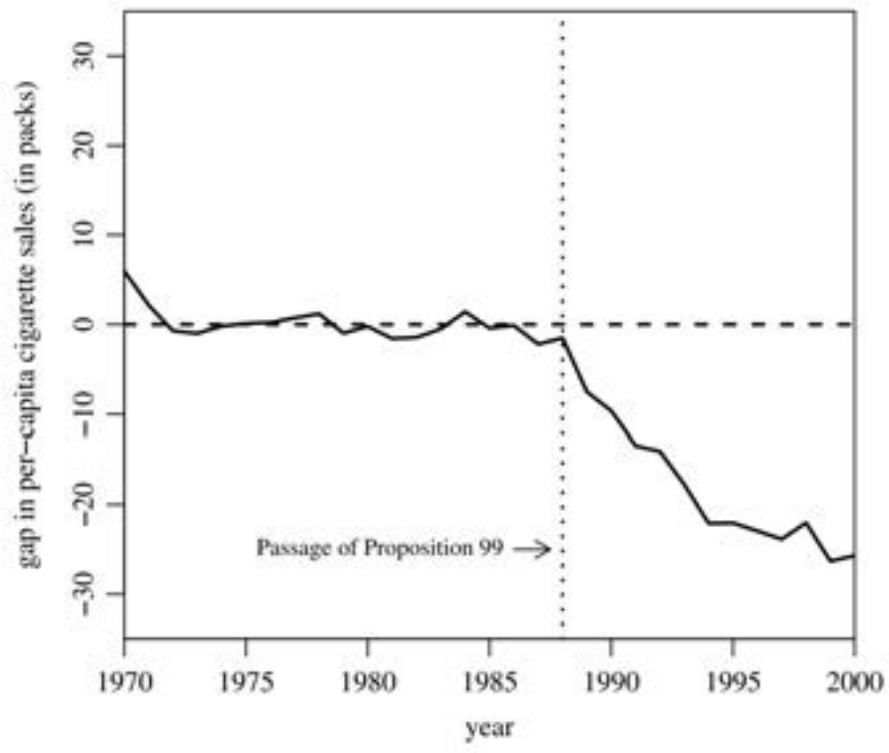


Figure 3. Per-capita cigarette sales gap between California and synthetic California.

Inference in the synthetic control method (Abadie et al. 2010)

- There is only a single treated unit
  - Large sample asymptotics unlikely to work

Placebo approach is standard: apply method to each potential control unit, and report effect in period

- Analogy here is to a randomization inference argument, comparing to a "null" effect

- Best practice: also show placebo treatments

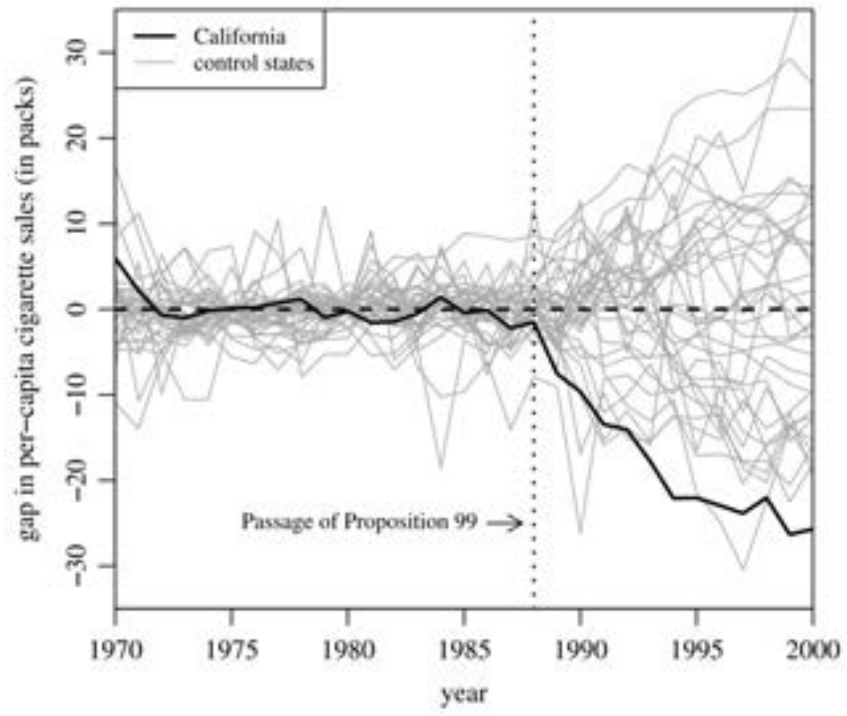


Figure 5. Per-capita cigarette sales gaps in California and placebo gaps in 34 control states (discards states with pre-Proposition 99 MSPE twenty times higher than California's).

Inference in the synthetic control method (Abadie et al. 2010)

- There is only a single treated unit
  - Large sample asymptotics unlikely to work
- Placebo approach is standard: apply method to each potential control unit, and report effect in period
  - Analogy here is to a randomization inference argument, comparing to a "null" effect

Best practice: also show placebo treatments

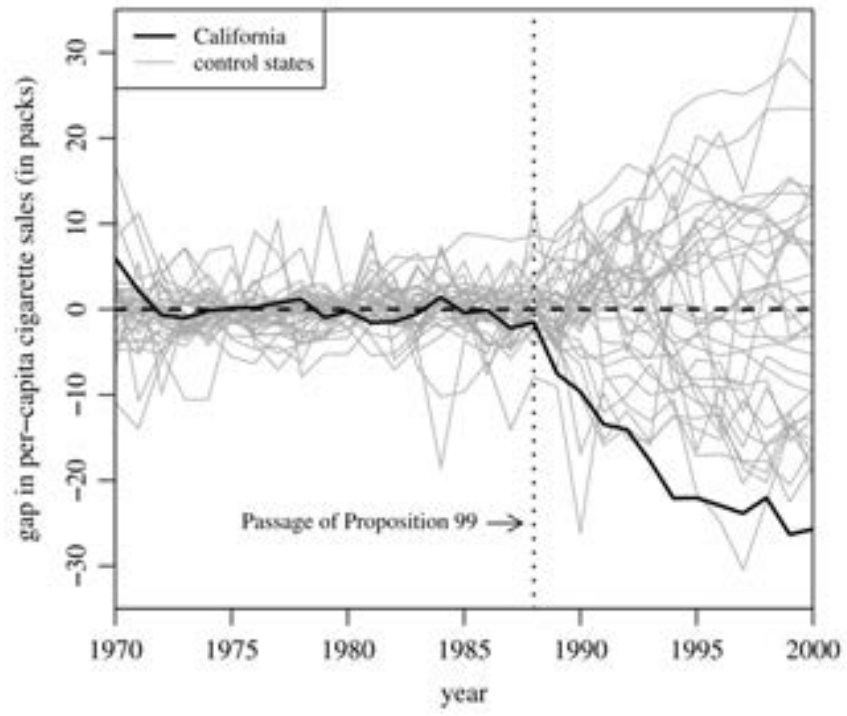
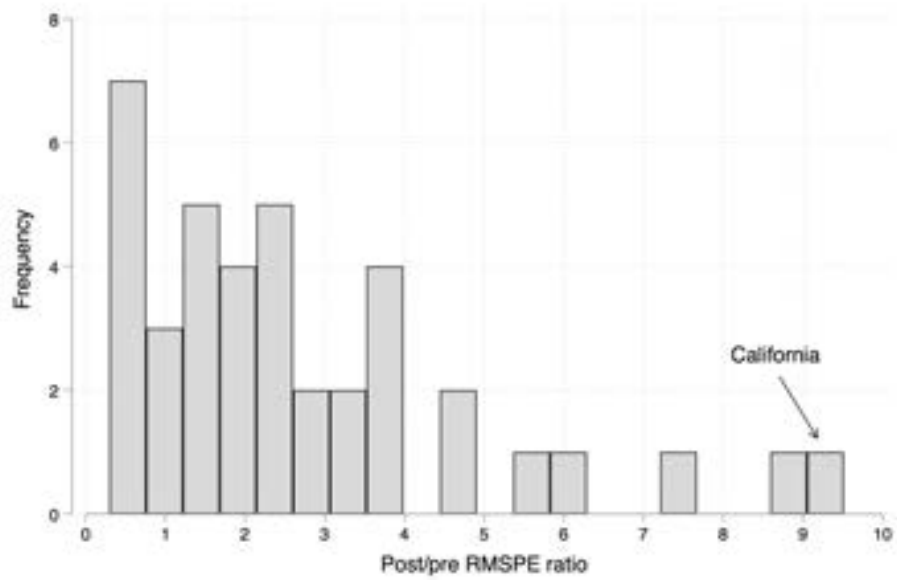


Figure 5. Per-capita cigarette sales gaps in California and placebo gaps in 34 control states (discards states with pre-Proposition 99 MSPE twenty times higher than California's).

Inference in the synthetic control method (Abadie et al. 2010)

- There is only a single treated unit
  - Large sample asymptotics unlikely to work
- Placebo approach is standard: apply method to each potential control unit, and report effect in period
  - Analogy here is to a randomization inference argument, comparing to a "null" effect

Best practice: also show placebo treatments



Inference in the synthetic control method (Abadie et al. 2010)

- There is only a single treated unit
  - Large sample asymptotics unlikely to work
- Placebo approach is standard: apply method to each potential control unit, and report effect in period
  - Analogy here is to a randomization inference argument, comparing to a "null" effect
- Best practice: also show placebo treatments

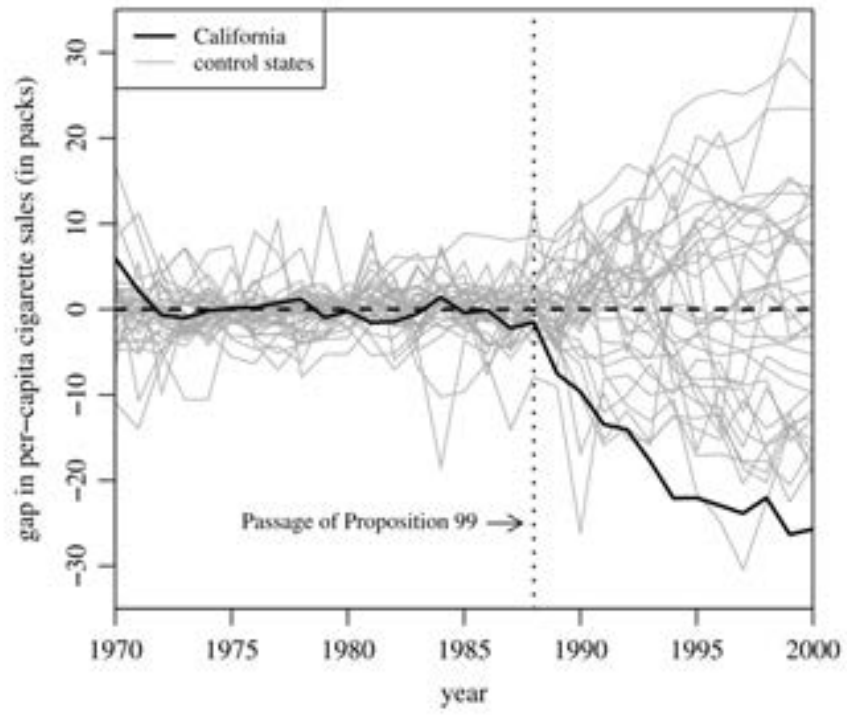


Figure 5. Per-capita cigarette sales gaps in California and placebo gaps in 34 control states (discards states with pre-Proposition 99 MSPE twenty times higher than California's).

So what about synthetic methods?

- Lots of new methodological papers coming out
- Synthetic control is the ideal approach when faced with a single treatment
  - But what about selection on pre-trends (Roth 2020)?
- So far, limited application by researchers. Why?
- Potential reasons:
  - These are strong structural assumptions, and not clear we have good tests yet
  - Despite concerns re: pre-trends in dind, the assumptions felt testable
- Researcher degrees of freedom seem multifold. True in DinD too, but perhaps more transparent?

Further reading if you are interested

- See chapter in Scott Cunningham's Mixtape for a practical introduction
- See Abadie (2021, Journal of Economic Literature) for a slightly more formal introduction
- See Arkhangelsky et al. (2023, AER) for a cutting-edge method to merge diff-in-diff and synthetic controls ("synthetic diff-in-diffs")
- Stata Package: synth or allsynth; many packages only in R (e.g., synthdid)