Diff-in-diff

Advanced Applied Econometrics

Felix Weinhardt

Slides based on Scott Cunningham and Paul Goldsmith-Pinkham
Thanks also to Renke Schmacker

Differences-in-differences

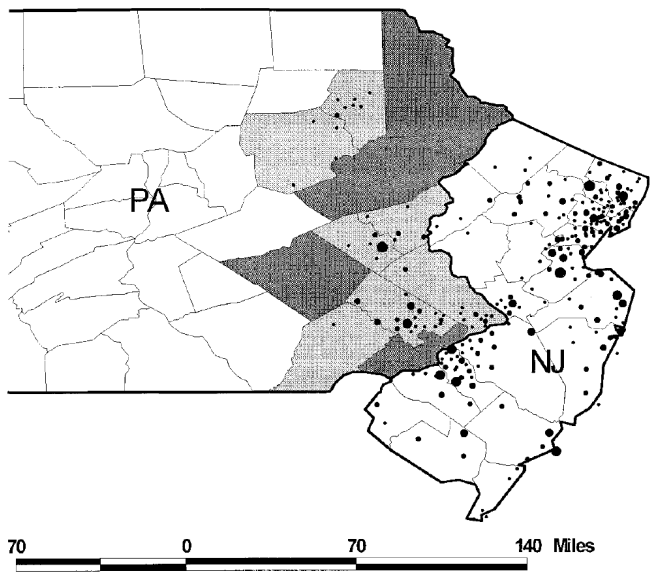
- Goal: estimate effect of a policy
- Naive approach: compare outcomes for treated and untreated groups
- Problem: but policy-adoption is not random $\Rightarrow$ what is the counterfactual?
- Solution: estimate difference-in-difference (DiD)

Differences-in-differences: Card and Krueger (1994)

- Suppose you are interested in the effect of minimum wages on employment (a classic and controversial question in labor economics).
- In a competitive labor market, increases in the minimum wage would move us up a downward sloping labor demand curve $\rightarrow$ employment would fall

DD: Card and Krueger (1994)

- Card and Krueger (1994) analyzed the effect of a minimum wage increase in New Jersey using a differences-in-differences (DD) methodology
- In February 1992, New Jersey increased the state minimum wage from \$4.25 to \$5.05. Pennsylvania's minimum wage stayed at \$4.25.



- They surveyed about 400 fast food stores both in New Jersey and Pennsylvania before and after the minimum wage increase in New Jersey

DD Strategy

- DD is a version of fixed effects estimation. To see this formally:

Y^1_{ist} : employment at restaurant i, state s, time t with a high w^{min}

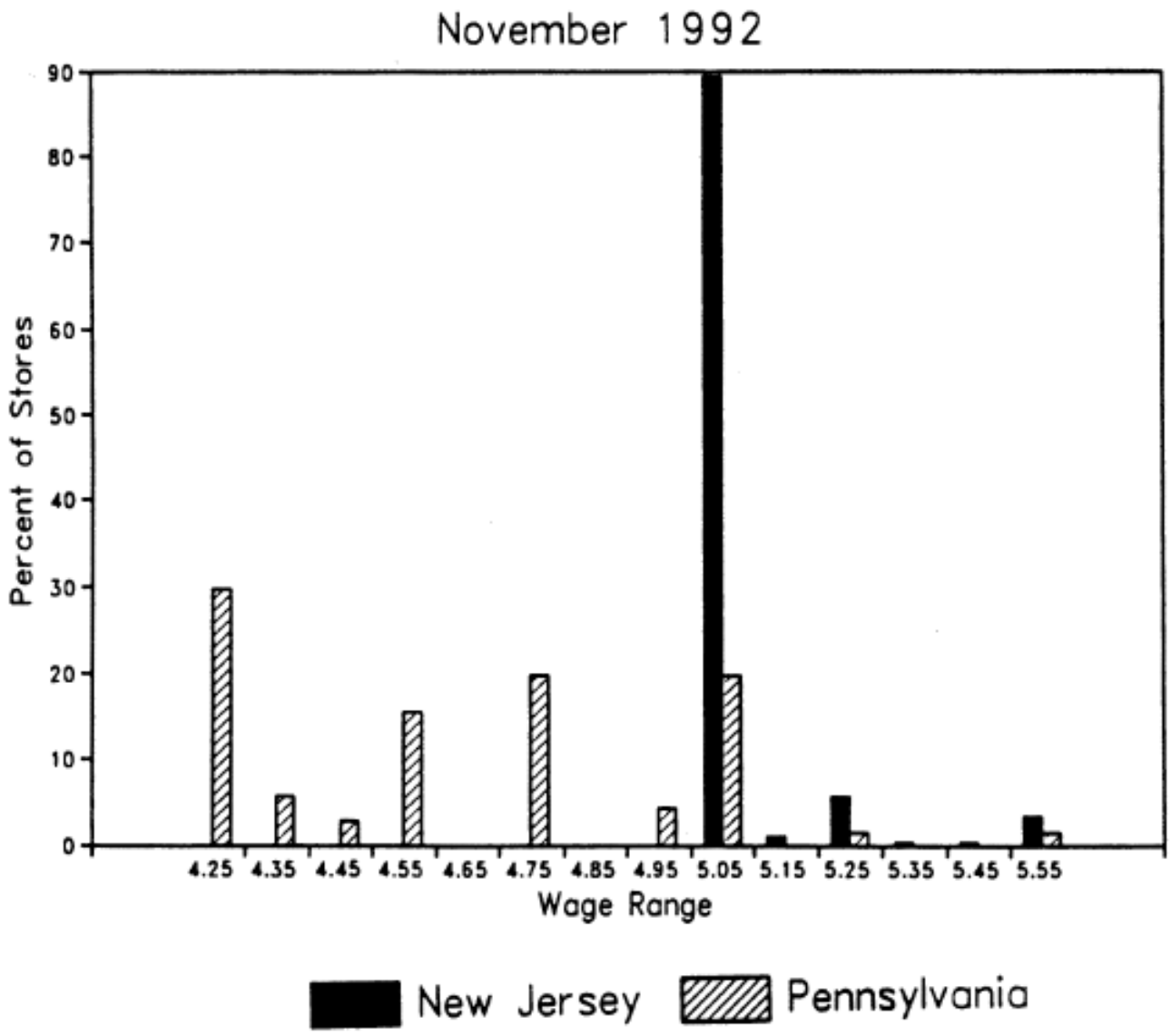
Y^0_{ist} : employment at restaurant i, state s, time t with a low w^{min}

- In practice of course we only see one or the other. We then assume that:

$$E[Y^0_{its}|s,t] = \gamma_s + \lambda_t$$

- In the absence of a minimum wage change, employment is determined by the sum of a time-invariant state effect, γ_s and a year effect, λ_t that is common across states
- Let D_{st} be a dummy for high-minimum wage states and periods
- Assuming $E[Y^1_{its} - Y^0_{its}|s,t] = \delta$ is the treatment effect, observed employment can be written:

$$Y_{its} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{its}$$


November 1992
Percent of Stores
90
80
70
60
50
40
30
20
10
0
4.25
4.35
4.45
4.55
4.65
4.75
4.85
4.95
5.05
5.15
5.25
5.35
5.45
5.55
Wage Range
New Jersey
Pennsylvania

DD Strategy II

- In New Jersey
  - Employment in February is

$$E(Y_{ist}|s = NJ, t = Feb) = \gamma_{NJ} + \lambda_{Feb}$$

  - Employment in November is:

$$E(Y_{ist}|s = NJ, t = Nov) = \gamma_{NJ} + \lambda_{Nov} + \delta$$

  - Difference between November and February

$$E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb) = \lambda_{Nov} - \lambda_{Feb} + \delta$$

- In Pennsylvania
  - Employment in February is

$$E(Y_{ist}|s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb}$$

  - Employment in November is:

$$E(Y_{ist}|s = PA, t = Nov) = \gamma_{PA} + \lambda_{Nov}$$

  - Difference between November and February

$$E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb}$$

DD Strategy II

- In New Jersey
  - Employment in February is

$$E(Y_{ist}|s = NJ, t = Feb) = \gamma_{NJ} + \lambda_{Feb}$$

  - Employment in November is:

$$E(Y_{ist}|s = NJ, t = Nov) = \gamma_{NJ} + \lambda_{Nov} + \delta$$

  - Difference between November and February

$$E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb) = \lambda_{Nov} - \lambda_{Feb} + \delta$$

- In Pennsylvania
  - Employment in February is

$$E(Y_{ist}|s = PA, t = Feb) = \gamma_{PA} + \lambda_{Feb}$$

  - Employment in November is:

$$E(Y_{ist}|s = PA, t = Nov) = \gamma_{PA} + \lambda_{Nov}$$

  - Difference between November and February

$$E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb) = \lambda_{Nov} - \lambda_{Feb}$$

DD Strategy III

- The DD strategy amounts to comparing the change in employment in NJ to the change in employment in PA.
- The population DD are:

$$\Big(E(Y_{ist}|s = NJ, t = Nov) - E(Y_{ist}|s = NJ, t = Feb)\Big)$$
$$-\Big(E(Y_{ist}|s = PA, t = Nov) - E(Y_{ist}|s = PA, t = Feb)\Big) = \delta$$

- This is estimated using the sample analog of the population means

Variable	Stores by state		
	PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Surprisingly, employment *rose* in NJ relative to PA after the minimum wage change

Regression DD

- We can estimate the DD estimator in a regression framework
- Advantages:
  - It's easy to calculate the standard errors
  - We can control for other variables which can make the parallel trend assumption more credible and may reduce the residual variance (lead to smaller standard errors)
  - It's easy to include multiple periods
  - We can study treatments with different treatment intensity. (e.g., varying increases in the minimum wage for different states)
- The typical regression model we estimate is

$$\text{Outcome}_{it} = \beta_1 + \beta_2 \text{Treat}_i + \beta_3 \text{Post}_t + \beta_4 (\text{Treat} \times \text{Post})_{it} + \varepsilon_{it}$$

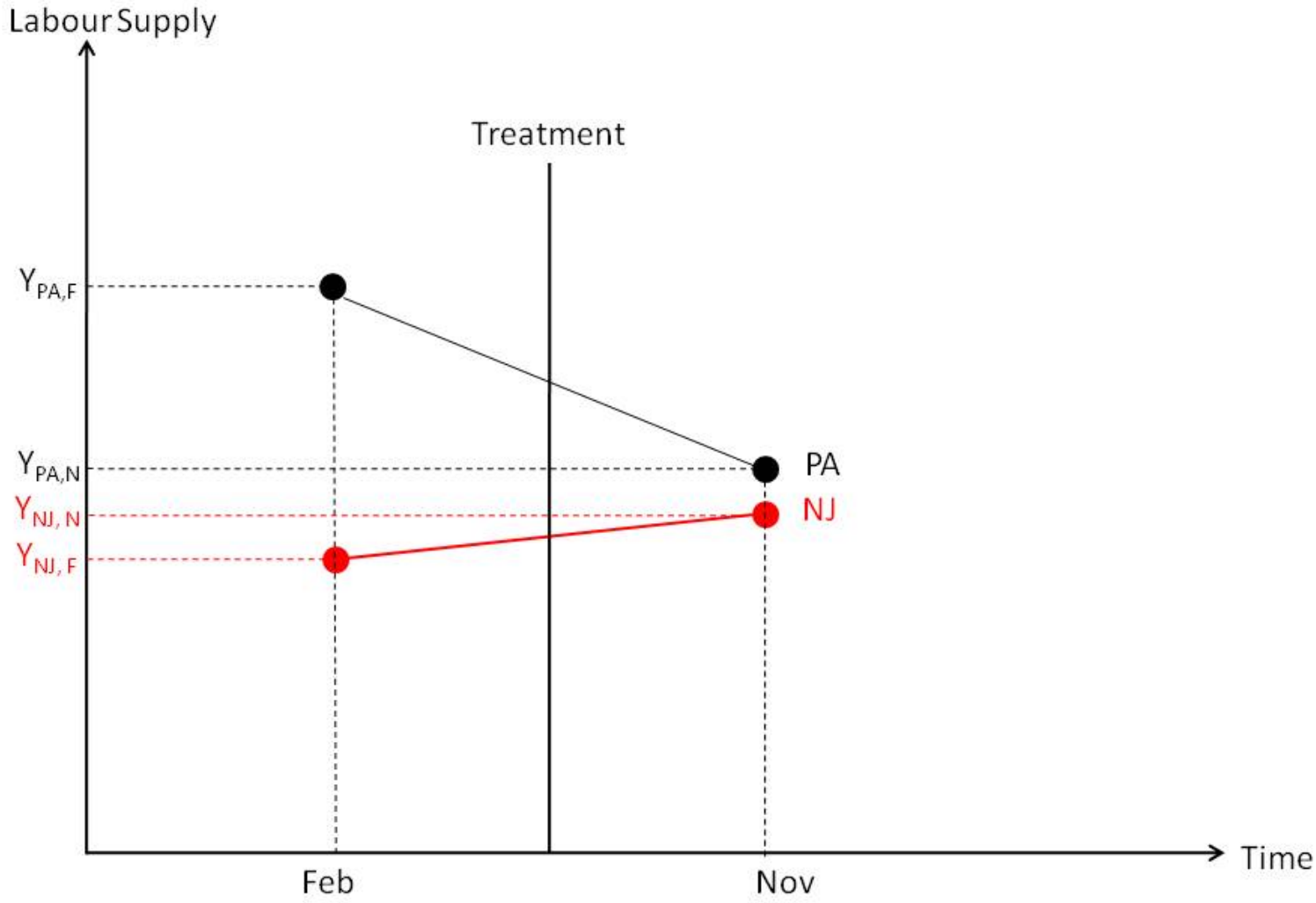
where Treat is a dummy if the observation is in the treatment group and Post is a post treatment dummy

Regression DD - Card and Krueger

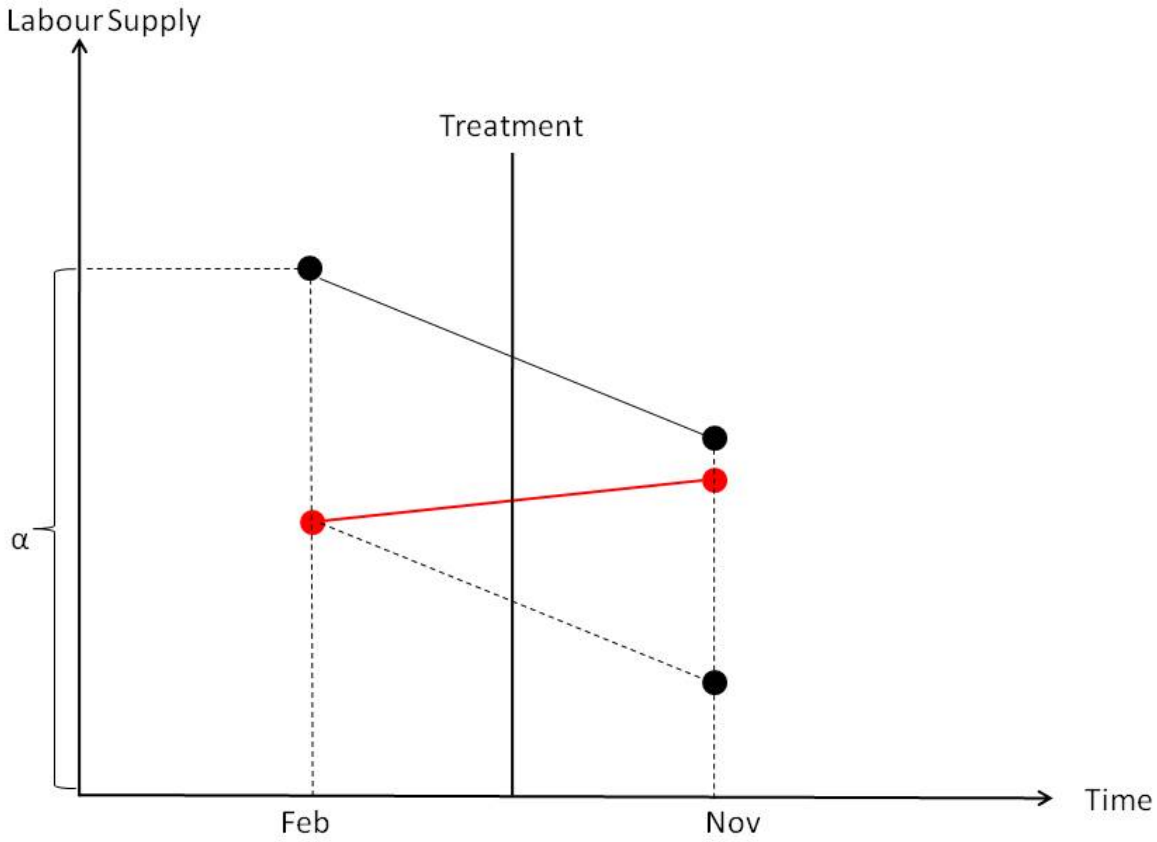
- In the Card and Krueger case, the equivalent regression would be:

$$Y_{its} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{its}$$

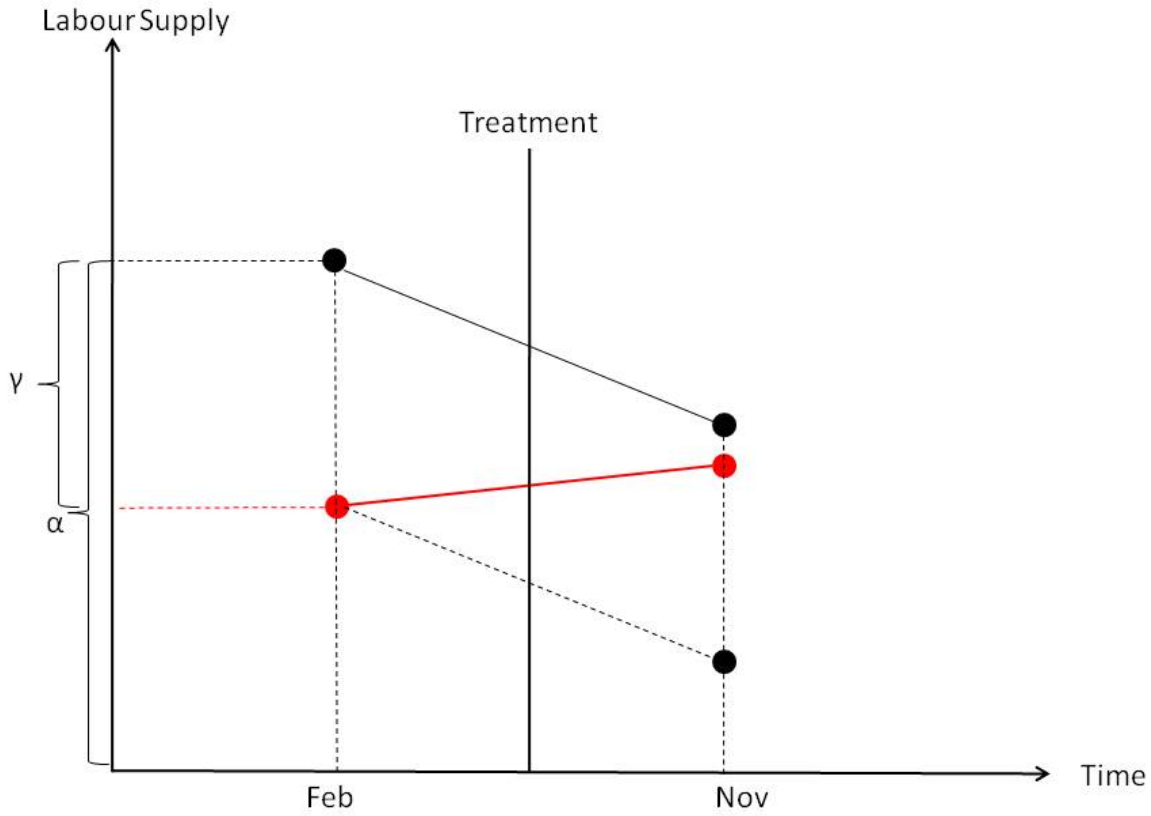
  - NJ is a dummy equal to 1 if the observation is from NJ
  - d is a dummy equal to 1 if the observation is from November (the post period)
- This equation takes the following values
  - PA Pre: α
  - PA Post: $\alpha + \lambda$
  - NJ Pre: $\alpha + \gamma$
  - NJ Post: $\alpha + \gamma + \lambda + \delta$
- DD estimate: (NJ Post - NJ Pre) - (PA Post - PA Pre) $= \delta$
- DD estimates the **ATT**


Labour Supply
Treatment
$Y_{PA,F}$
$Y_{PA,N}$
$Y_{NJ,N}$
$Y_{NJ,F}$
PA
NJ
Time
Feb
Nov

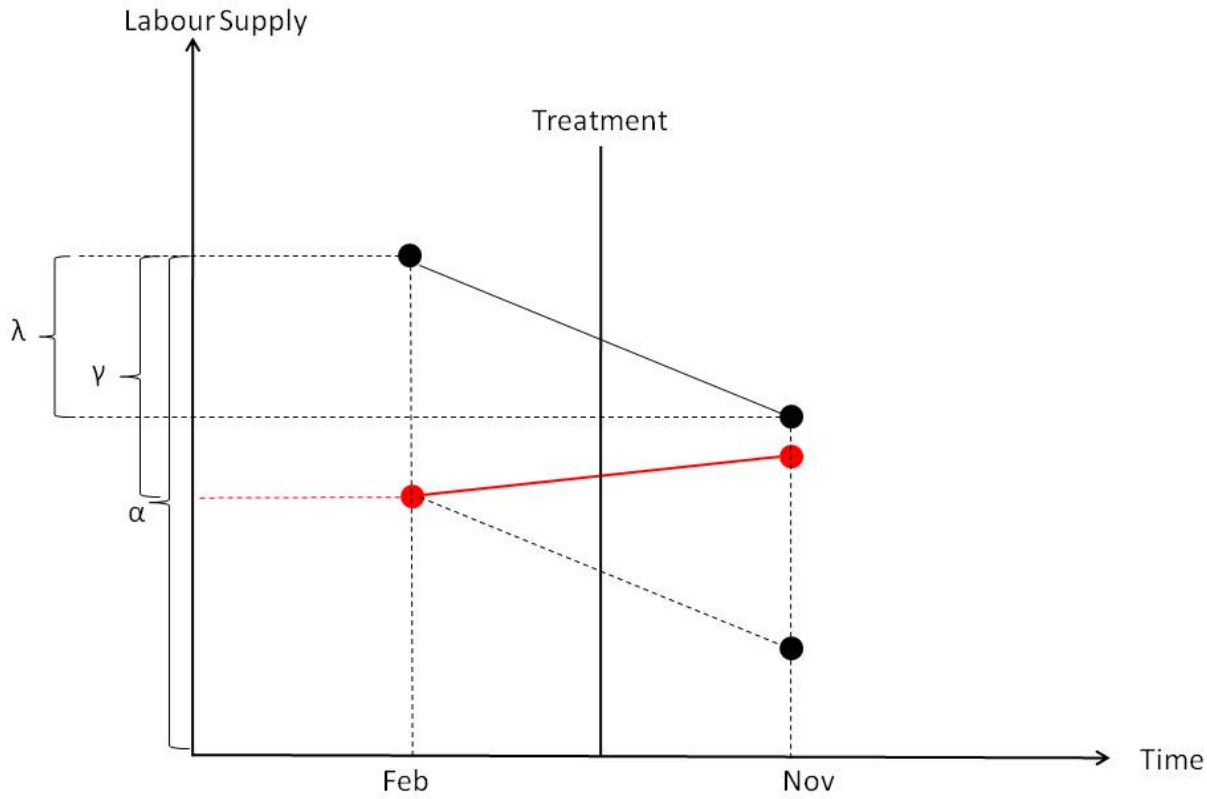
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



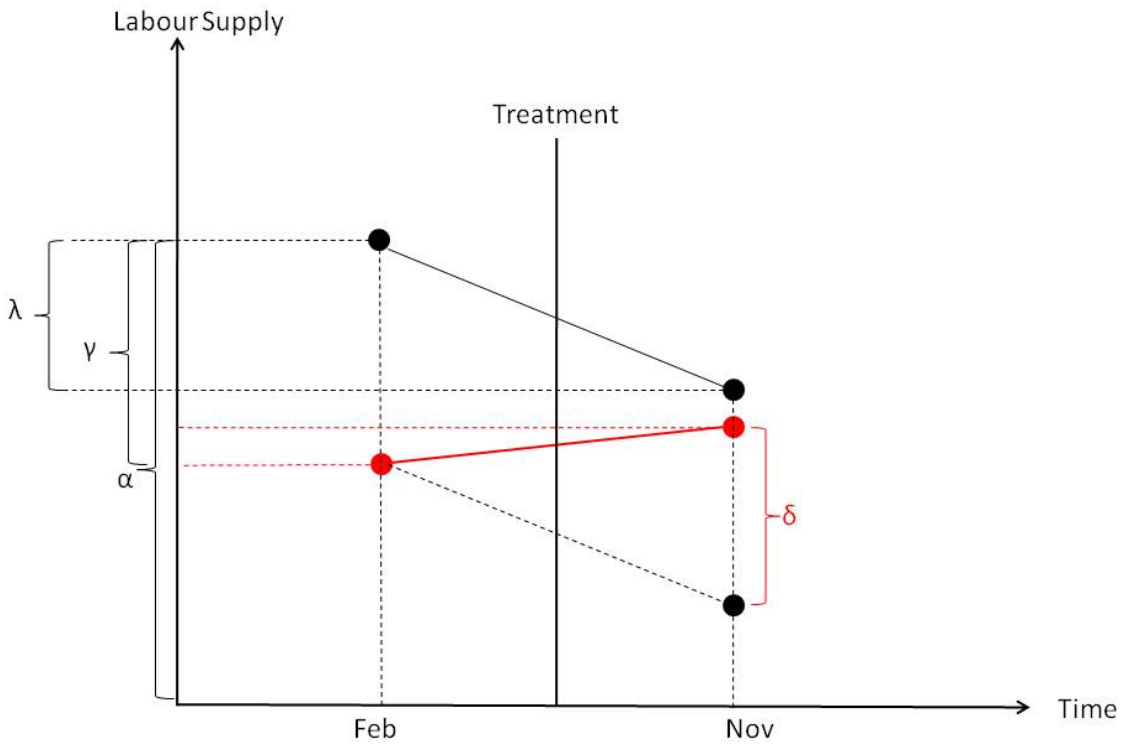
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ \times d)_{st} + \varepsilon_{ist}$$



Assumptions of DD strategy

- Parallel trends:
  - TWFE estimates:

$$\begin{aligned}\beta^{DiD} =& E(Y_{i2}(1) - Y_{i1}(0)|D_{i2} = 1) - E(Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0) \\ =& E(Y_{i2}(1) - Y_{i1}(0)|D_{i2} = 1) - E(Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 1) \\ &+ E(Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 1) - E(Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0)\end{aligned}$$

  - where $E(Y_{i2}(1) - Y_{i2}(0)|D_{i2} = 1) = ATT$ under the parallel trends assumption:

$$E(Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 1) = E(Y_{i2}(0) - Y_{i1}(0)|D_{i2} = 0)$$

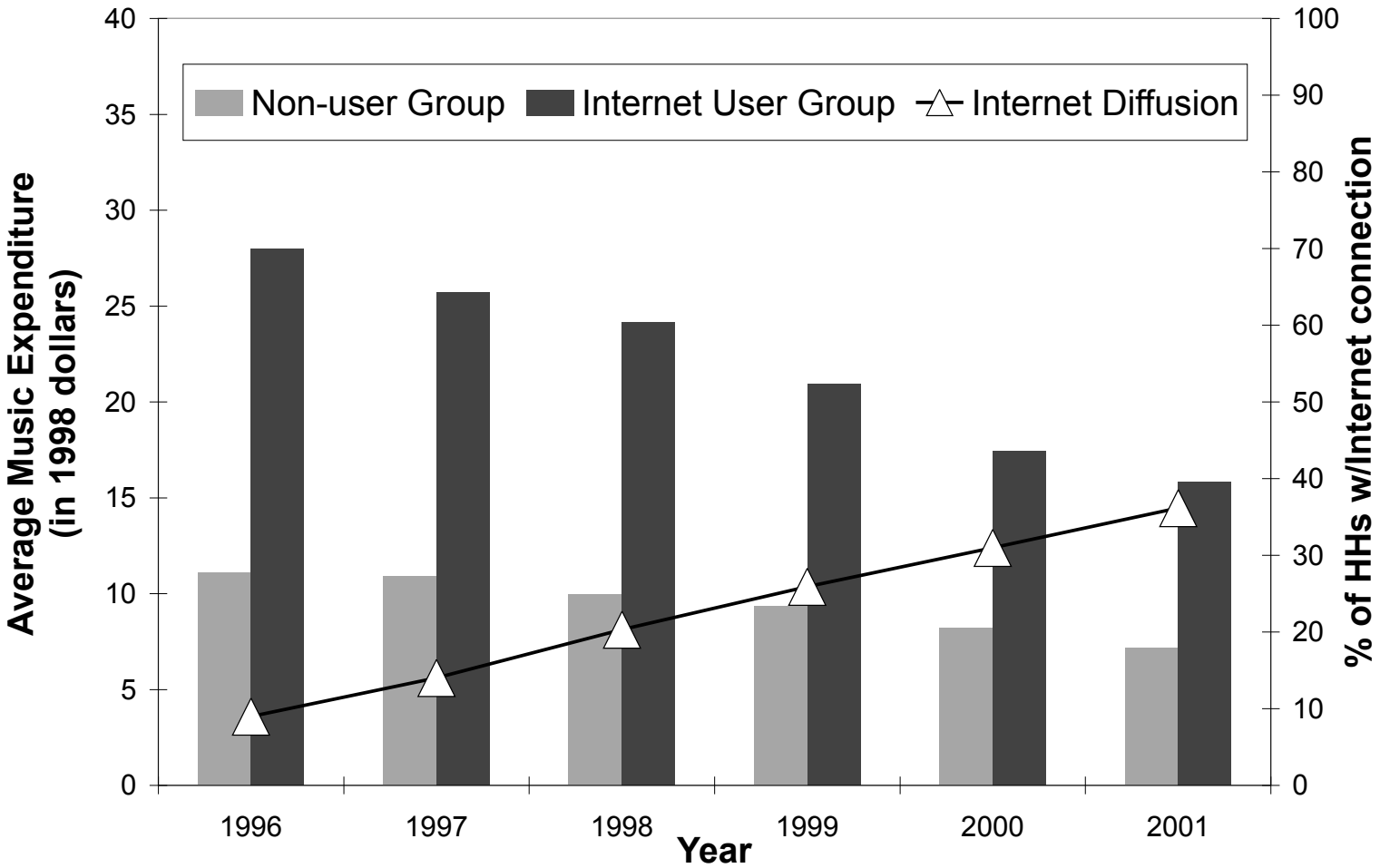
- Stable unit treatment value assumption (SUTVA)
- Fixed sample composition

Threats to validity – compositional differences

- One of the risks of a repeated cross-section (and unbalanced panel) is that the composition of the sample may have changed between the pre and post period
- Hong (2011) uses repeated cross-sectional data from the Consumer Expenditure Survey (CEX) containing music expenditure and internet use for a random sample of households
- Study exploits the emergence of Napter (first file sharing software widely used by Internet users) in June 1999 as a natural experiment
- Study compares internet users and internet non-users before and after emergence of Napster

Compositional differences?

Figure 1: Internet Diffusion and Average Quarterly Music Expenditure in the CEX



Threats to validity – compositional differences

Table 1: Descriptive Statistics for Internet User and Non-user Groups[a]

Year	1997		1998		1999	
	Internet User	Non-user	Internet User	Non-user	Internet User	Non-user
Average Expenditure						
Recorded Music	$25.73	$10.90	$24.18	$9.97	$20.92	$9.37
Entertainment	$195.03	$96.71	$193.38	$84.92	$182.42	$80.19
Zero Expenditure						
Recorded Music	.56	.79	.60	.80	.64	.81
Entertainment	.08	.32	.09	.35	.14	.39
Demographics						
Age	40.2	49.0	42.3	49.0	44.1	49.4
Income	$52,887	$30,459	$51,995	$28,169	$49,970	$26,649
High School Grad.	.18	.31	.17	.32	.21	.32
Some College	.37	.28	.35	.27	.34	.27
College Grad.	.43	.21	.45	.21	.42	.20
Manager	.16	.08	.16	.08	.14	.08

Diffusion of the Internet changes samples (e.g., younger music fans are early adopters)

Threats to validity – SUTVA

- Treatment status of any individual does not affect the potential outcomes of other units (non-interference)
  - Violation: Spillovers and general equilibrium effects
- No change in the composition of the treatment and control group due to the treatment
  - Violation: For example selective migration into area where treatment is introduced

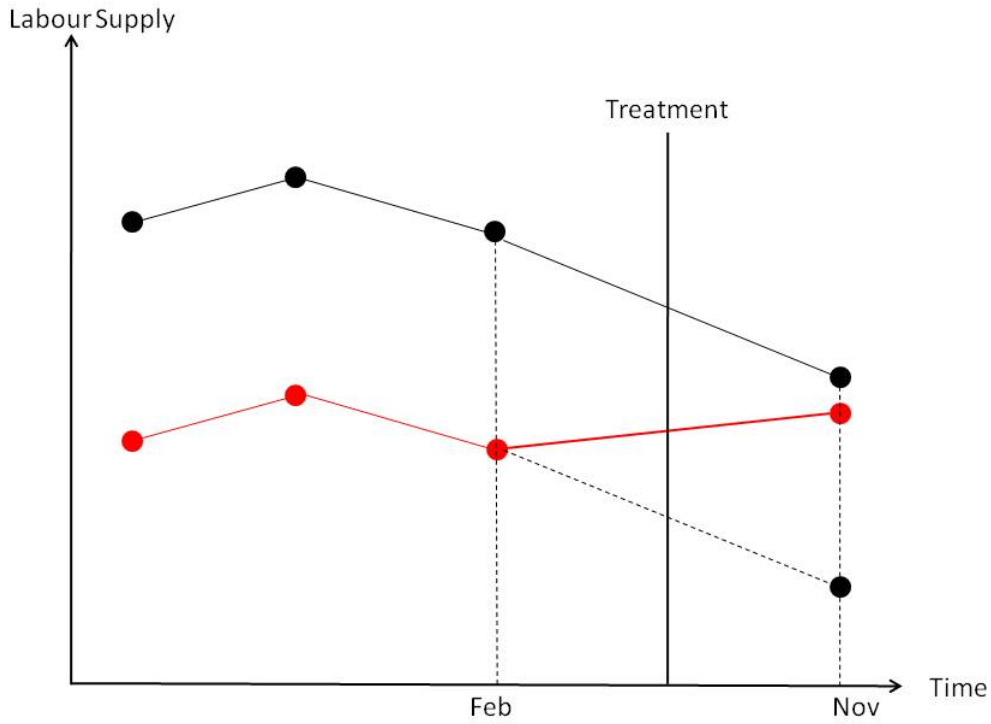
Threats to validity – Non-parallel trends

- Often policymakers will select the treatment and controls based on pre-existing differences in outcomes – practically guaranteeing the parallel trends assumption will be violated.
- "Ashenfelter dip"
  - Named after Orley Ashenfelter, labor economist at Princeton
  - Participants in job trainings program often experience a "dip" in earnings just prior to entering the program
  - Since wages have a natural tendency to mean reversion, comparing wages of participants and non-participants using DD leads to an upward biased estimate of the program effect.
- Regional targeting. NGOs may target villages that appear most promising, or worse off, which is a form of selection bias and violates parallel trends

Key assumption of any DD strategy: Parallel trends

- The key assumption for any DD strategy is that the outcome in treatment and control group would follow the same time trend in the absence of the treatment
- This doesn't mean that they have to have the same mean of the outcome
- Parallel trends are difficult to verify because technically one of the parallel trends is an unobserved counterfactual
- But one often will check using pre-treatment data to show that the trends are the same

Key assumption of any DD strategy: Parallel trends



Even if pre-trends are the same one still has to worry about other policies changing at the same time (omitted variable bias)

Regression DD Including Leads and Lags

- Suppose we have multiple periods before and after the treatment
  - Including leads into the DD model is an easy way to analyze pre-treatment trends
  - Lags can be included to analyze whether the treatment effect changes over time after assignment
- The estimated regression would be:

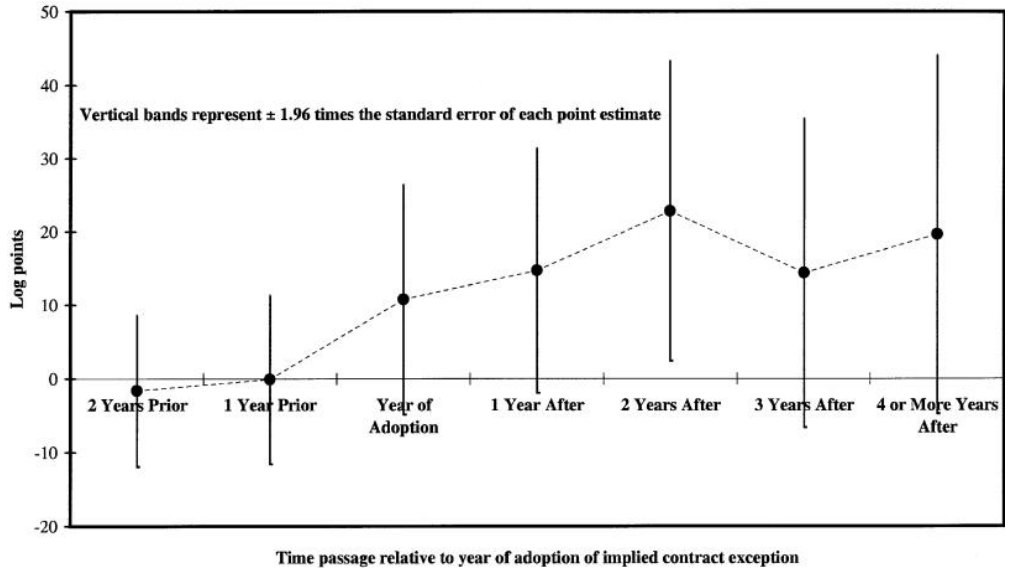
$$Y_{its} = \gamma_s + \lambda_t + \sum_{\tau=-q}^{-1} \gamma_\tau D_{s\tau} + \sum_{\tau=0}^{m} \delta_\tau D_{s\tau} + x_{ist} + \varepsilon_{ist}$$

  - Treatment occurs in year 0
  - Includes q leads or anticipatory effects
  - Includes m lags or post treatment effects
  - One of the coefficients is unidentified (TE relative to this period)

Study including leads and lags – Autor (2003)

- Autor (2003) includes both leads and lags in a DD model analyzing the effect of increased employment protection on the firm's use of temporary help workers
- In the US employers can usually hire and fire workers at will
- Some states courts have made some exceptions to this employment at will rule and have thus increased employment protection
- The standard thing to do is normalize the adoption year to 0
- Autor then analyzes the effect of these exemptions on the use of temporary help workers.

Results



- The leads are very close to 0. $\rightarrow$ no evidence for anticipatory effects (good news for the parallel trends assumption).
- The lags show the effect increases during the first years of the treatment and then remains relatively constant.

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

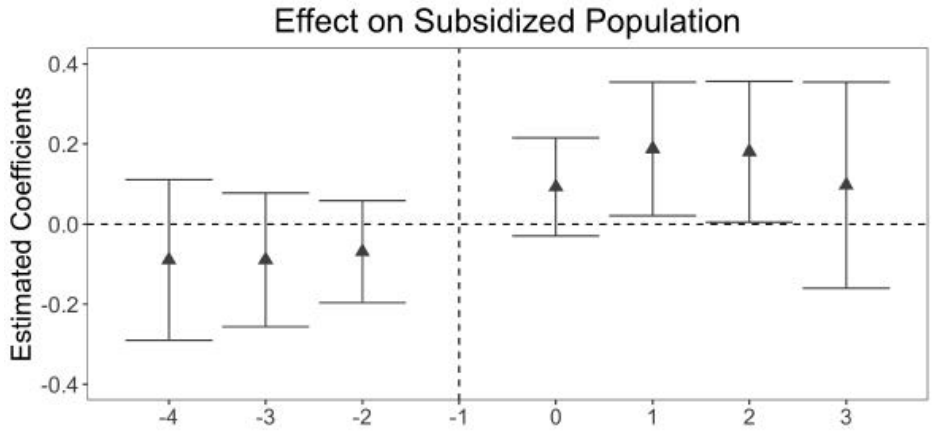
Parallel trends in what?

- 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."

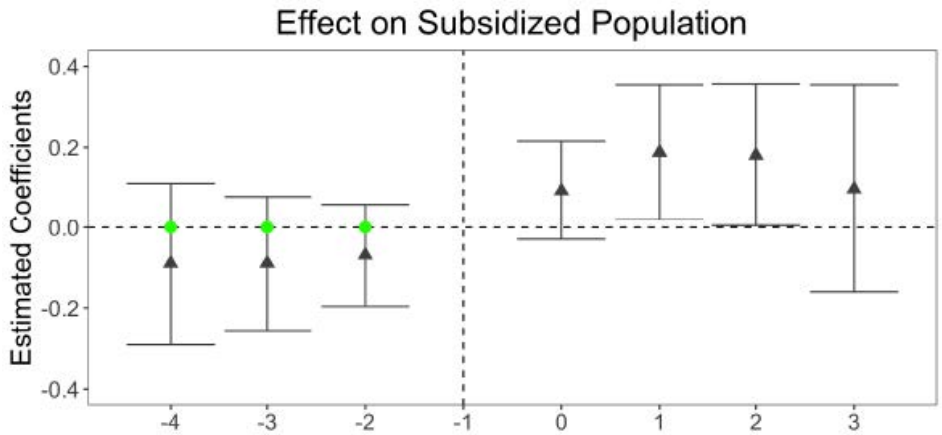
Pre-testing issues (Roth 2020)

- "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



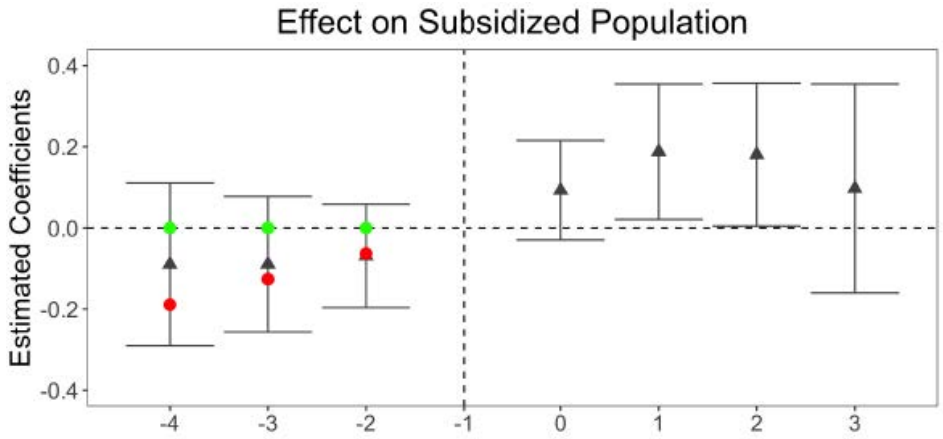
Pre-testing issues (Roth 2020)

- "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



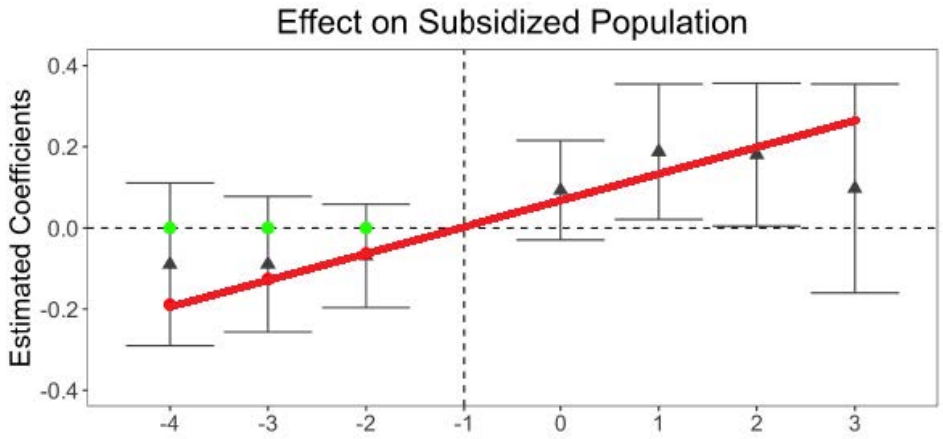
Pre-testing issues (Roth 2020)

- "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



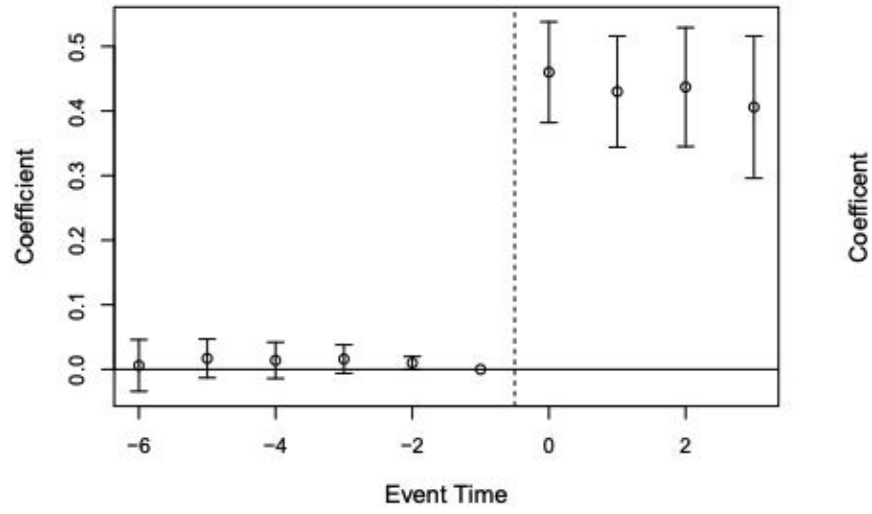
Pre-testing issues (Roth 2020)

- "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)

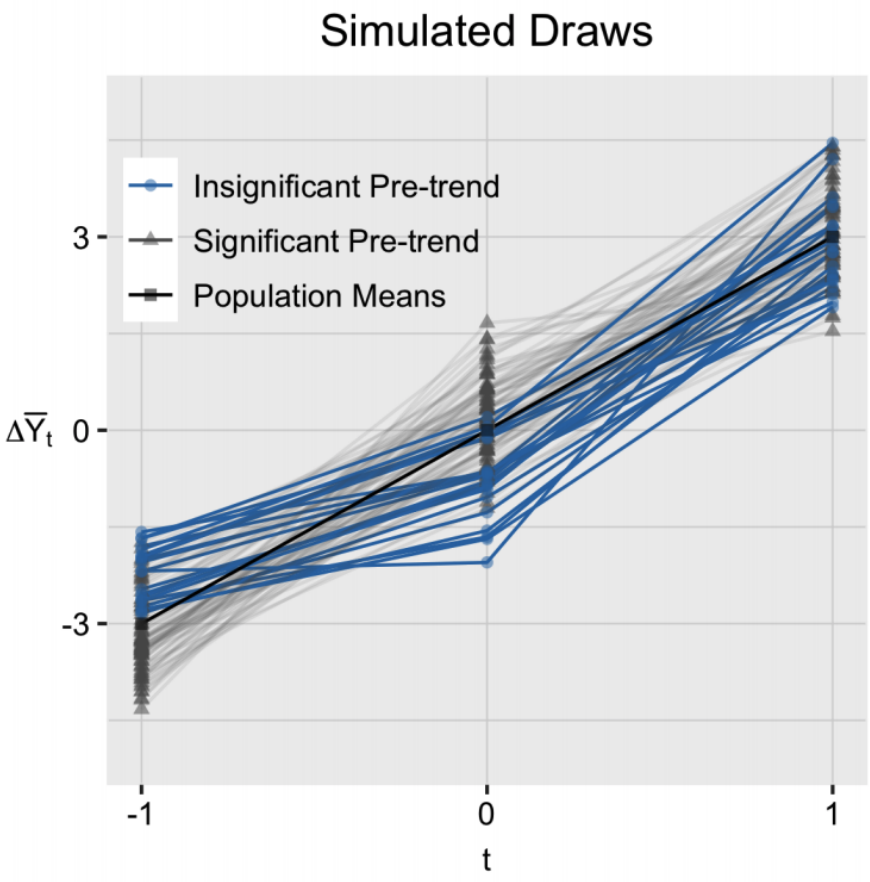
- "Event study" visualization of pre-trends
- Zero is in the 95% CIs, so we can't reject parallel trends
- But a sizable pre-trend is in the 95% CI as well
- With low power, we cannot rule out an important alternative hypothesis
- With highly precise pre-trends, the problem is less relevant

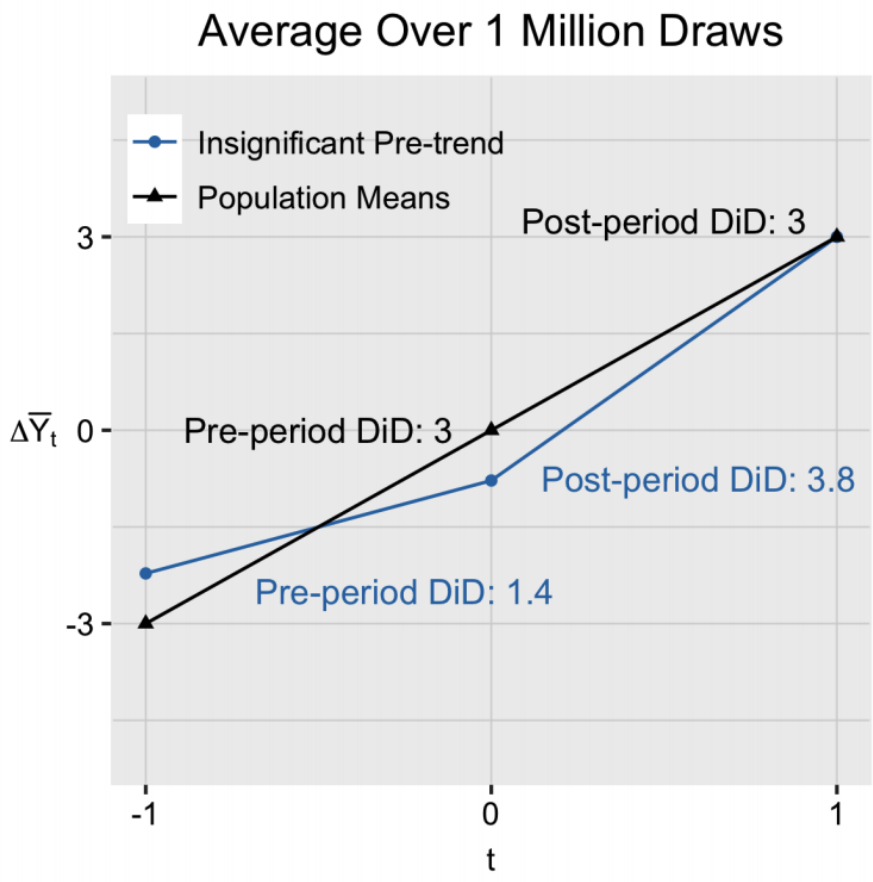


Outcome: Medicaid Eligibility

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

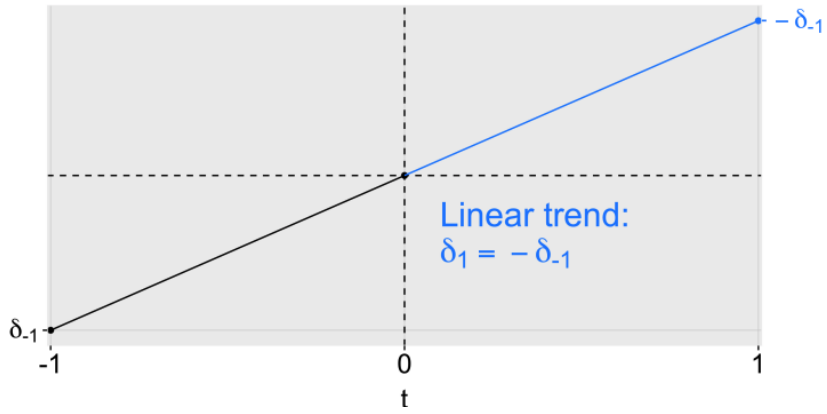
Rambachan and Roth (2020) suggestion

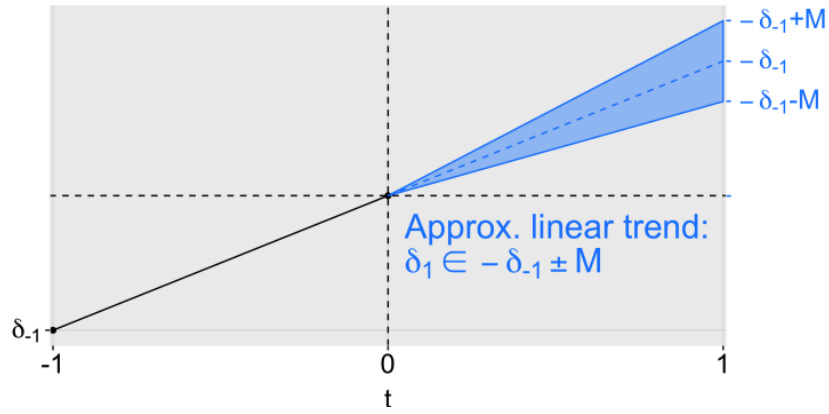
- Intuitive proposed solution for robustness. Note the post and pre

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

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

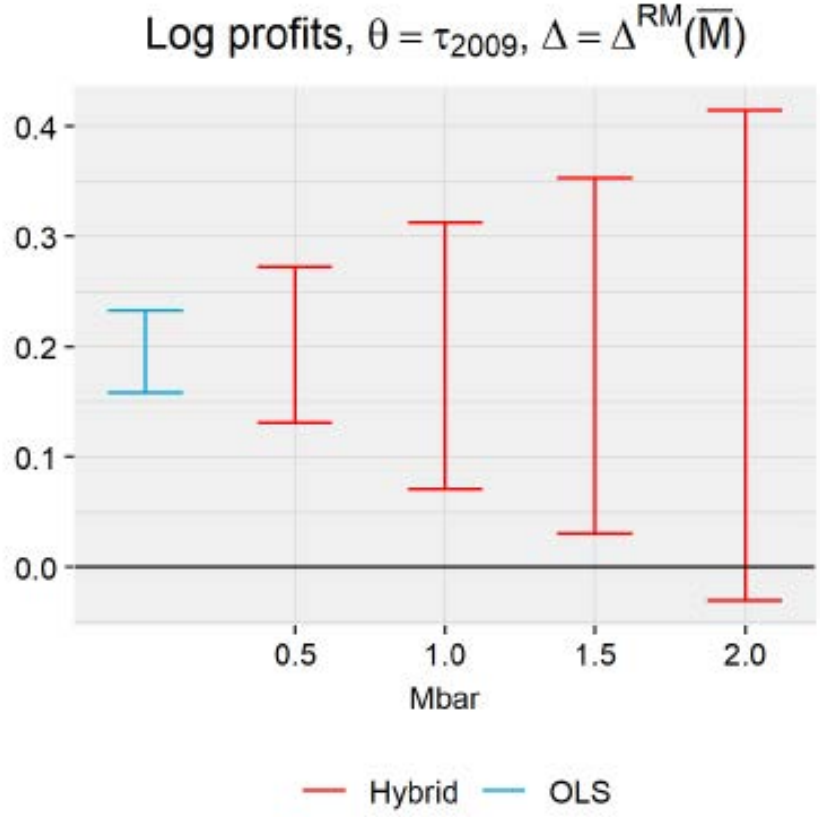
  effects:
- parallel trends assumes these δ are zero. But pre-trends may not be zero.
  - R&R say: we can use the info from our pre-trends to bound post-trend
  - Use a *smoothness* assumption, M, on the second derivative. E.g. simple case:





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

Cases of DinD

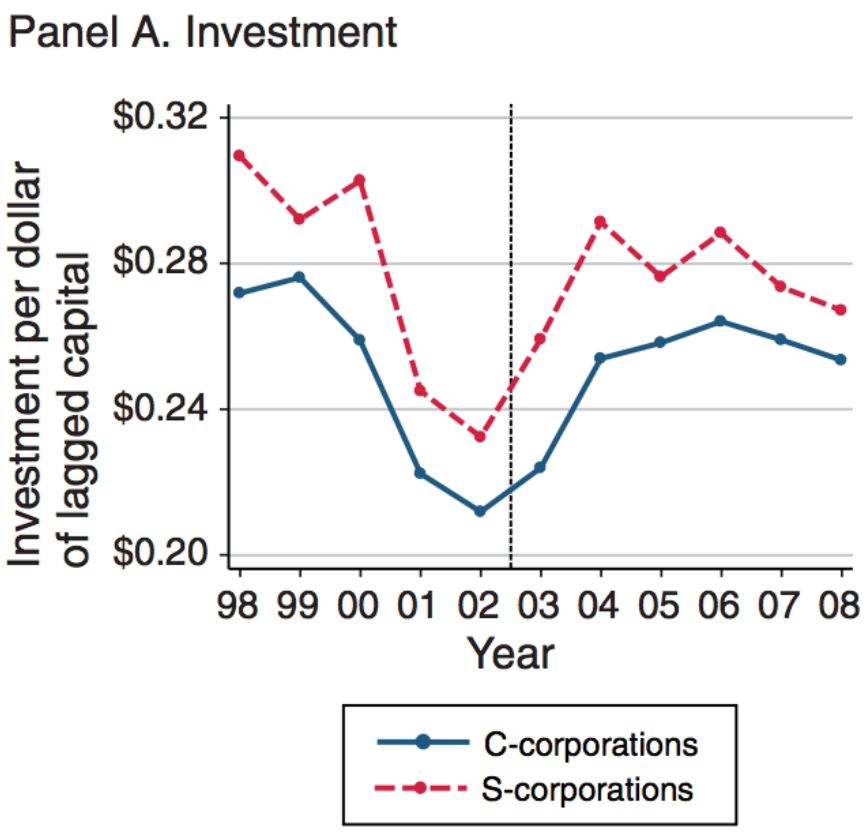
- 1 treatment timing, Binary treatment, 2 periods
  - Card and Krueger (AER, 1994)
- 1 treatment timing, Binary treatment, T periods
  - Yagan (AER, 2015)
- Staggered treatment timing, Binary treatment
  - Bailey and Goodman-Bacon (AER, 2015)

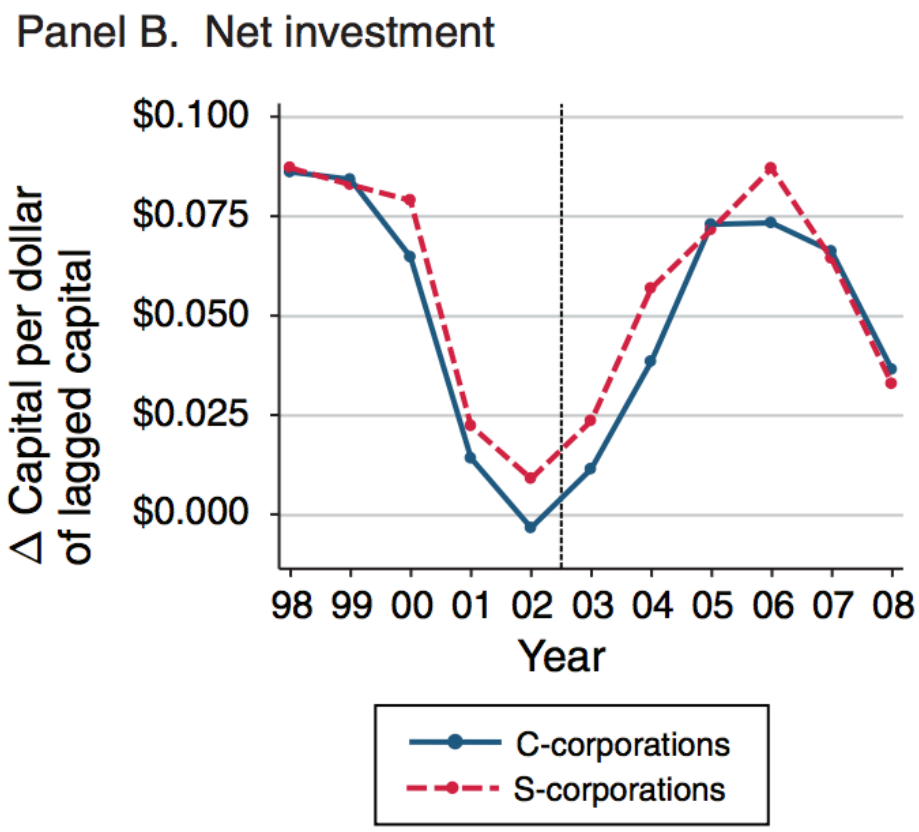
Yagan (2015)

- 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.

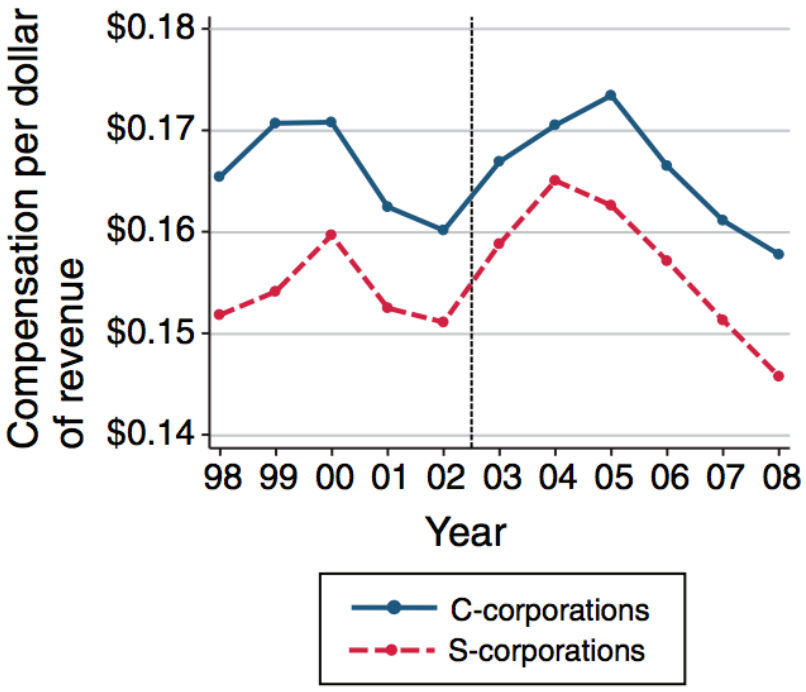
Investment Effects (none)



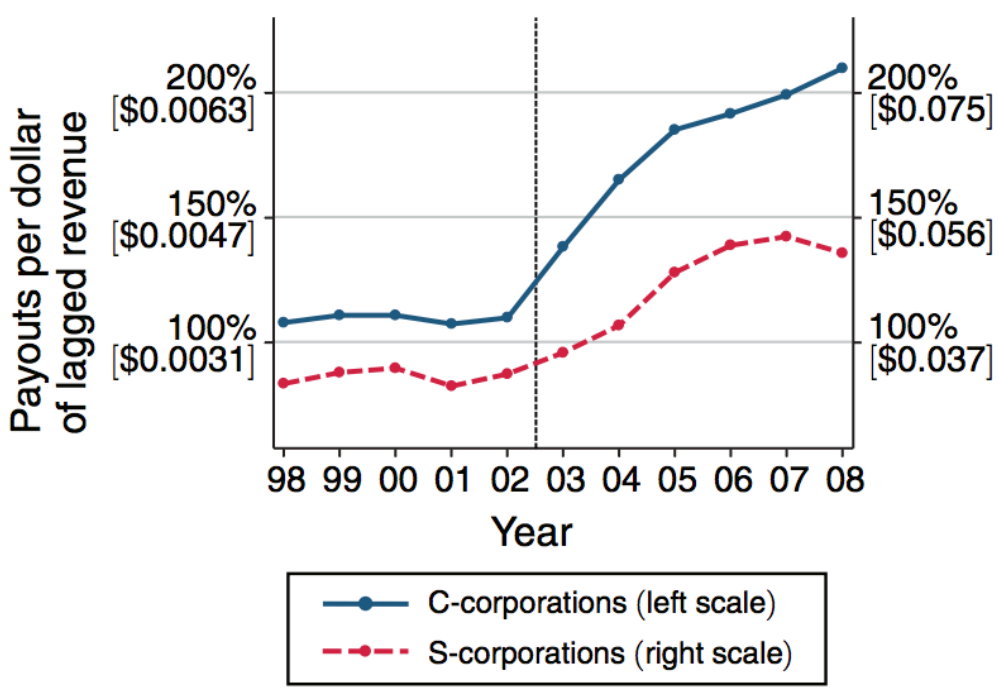


Employee + Shareholder effects (big)

Panel C. Employee compensation



Panel D. Total payouts to shareholders



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)

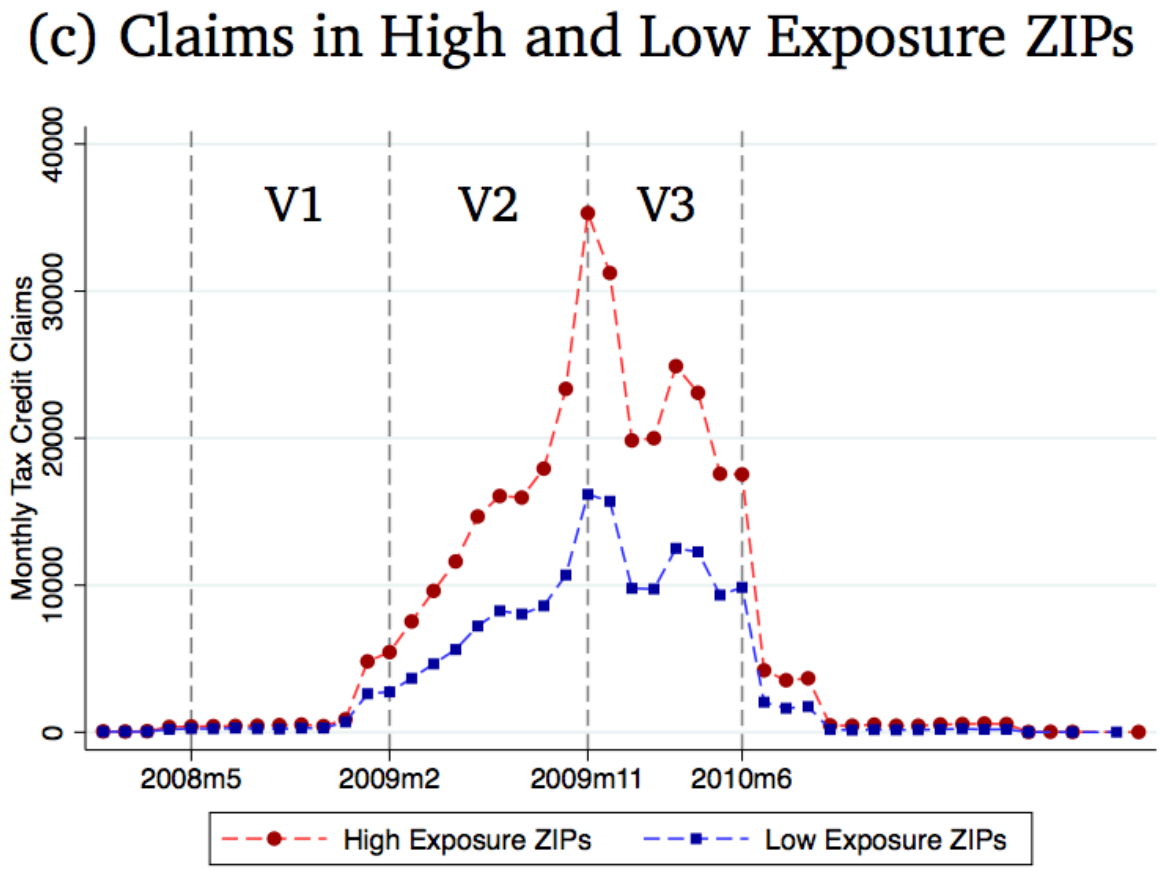
- This paper studies the impact of temporary fiscal stimulus (First-Time Home Buyer tax credit) on housing markets

- 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:

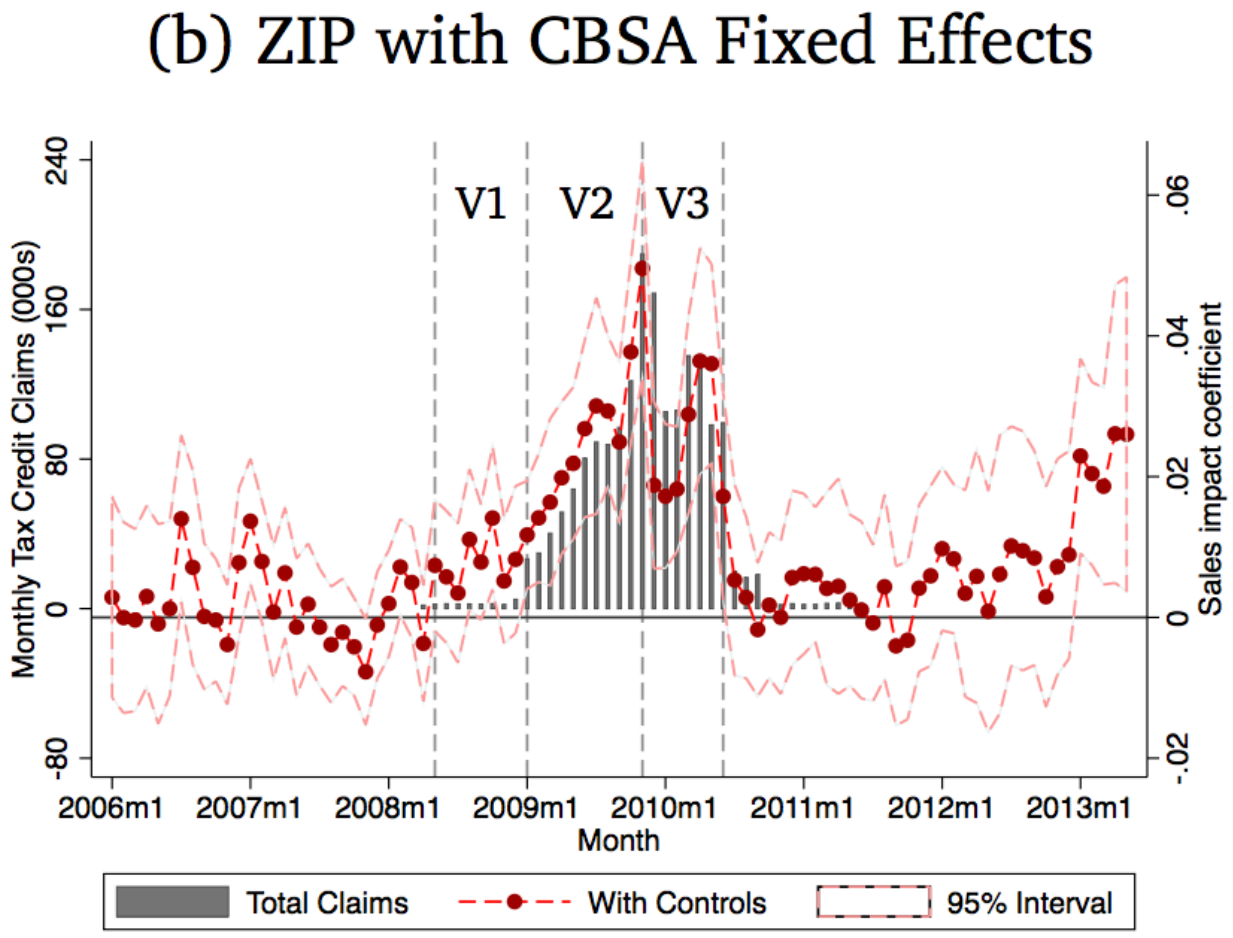
    The key threat to this design is the possibility that time-varying, pla
    specific shocks are correlated with our exposure measure.

- 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.

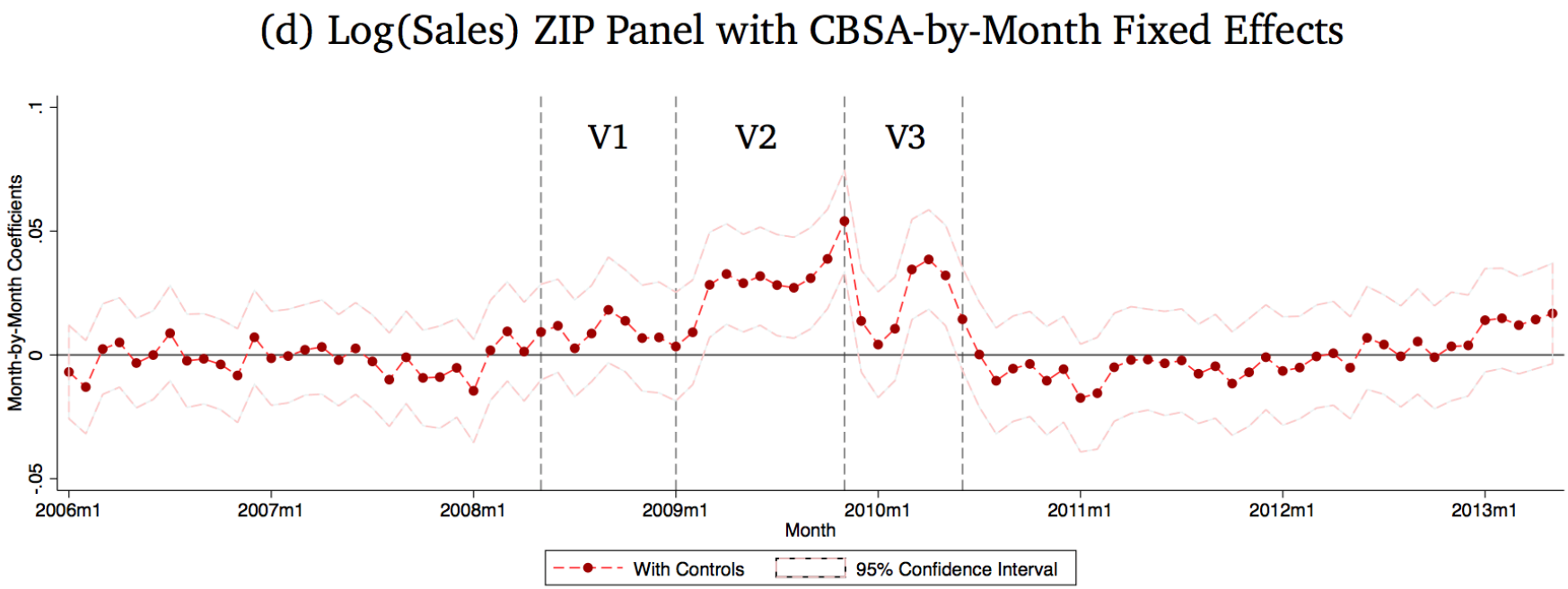
First stage: Binary approximation



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.

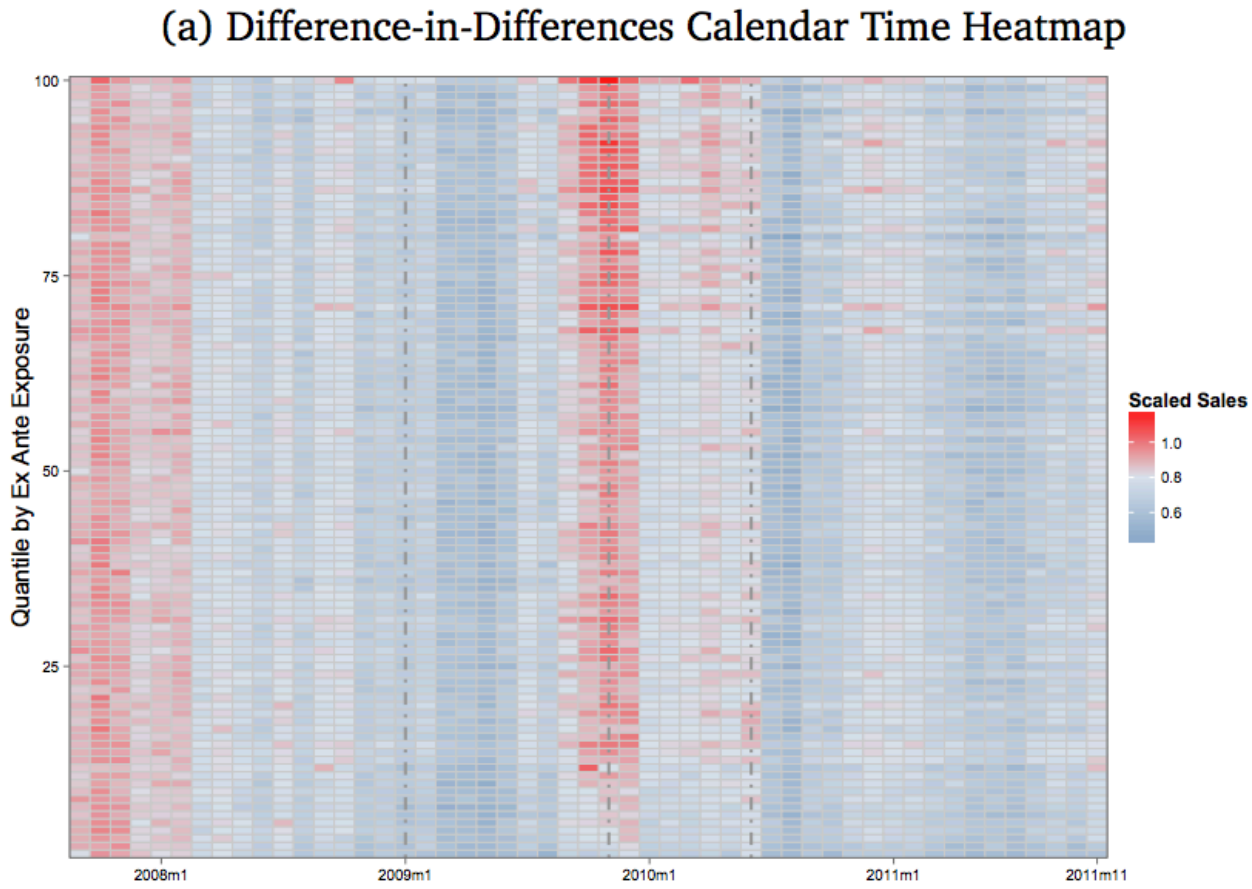
$$Y_{it} = \alpha_i + \gamma_t + \sum_{t=1, t \neq t_0}^{T} \delta_t D_{it} + \epsilon_{it}, \tag{1}$$

- 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

- 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

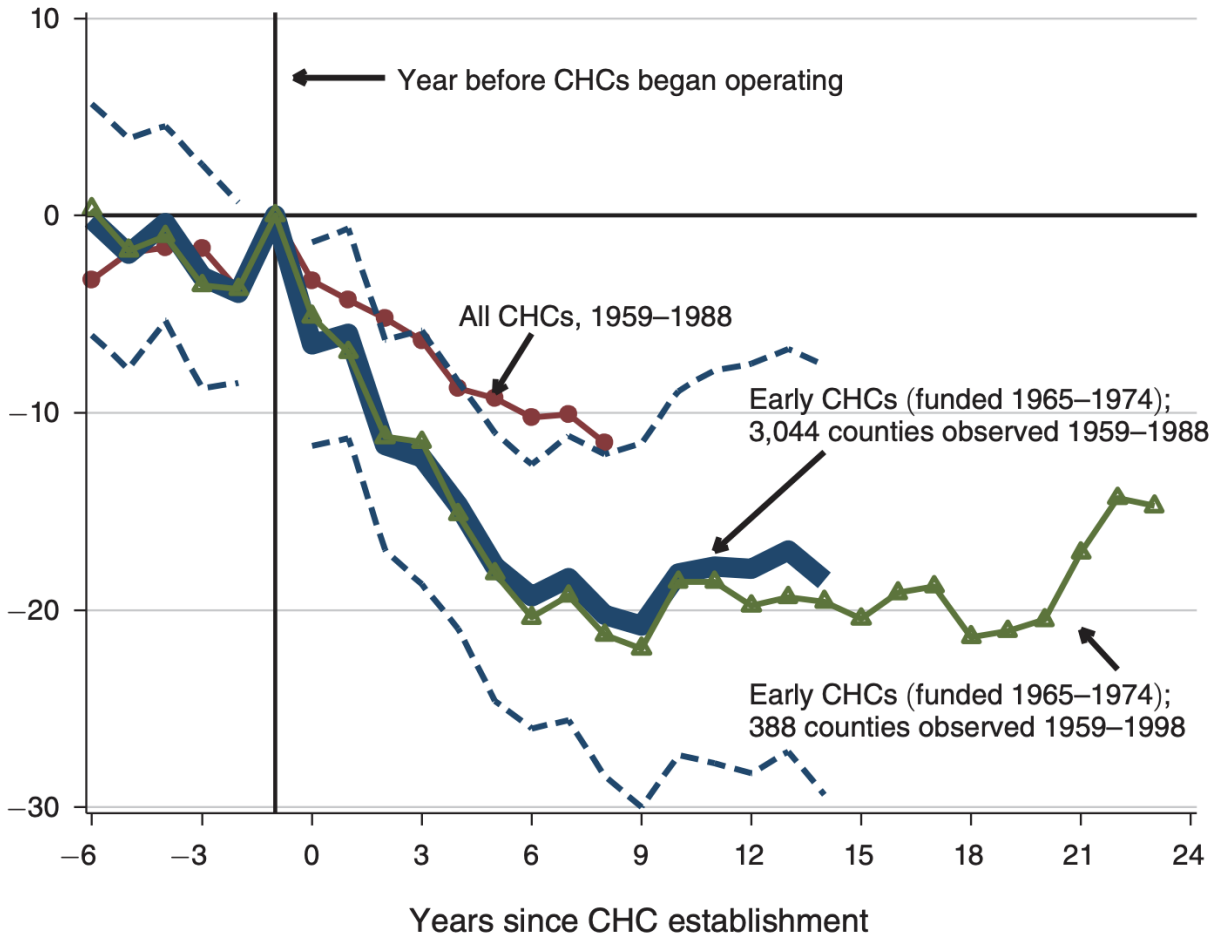
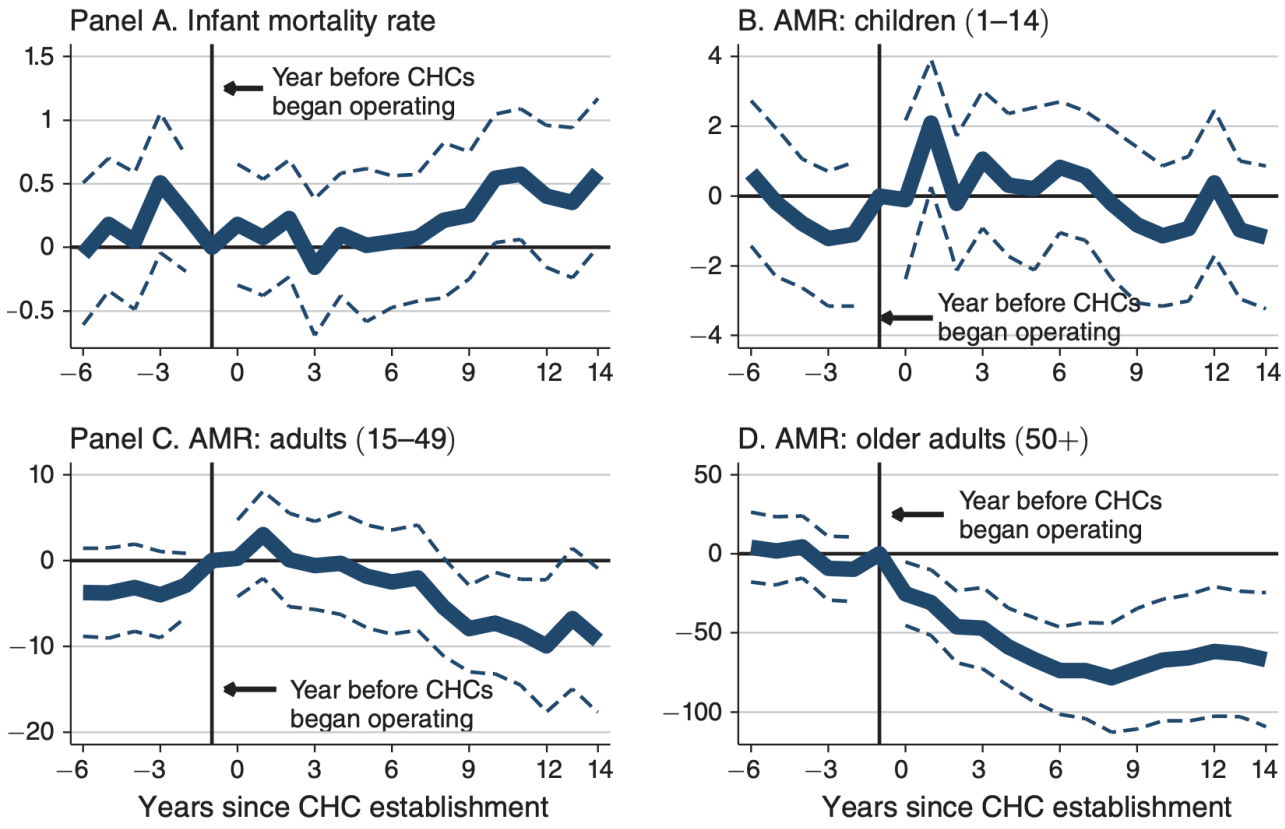


FIGURE 5. THE RELATIONSHIP BETWEEN COMMUNITY HEALTH CENTERS AND MORTALITY RATES

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

- 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", $G_i \geq t + 1$

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

- Why? The predicted value of $D_{i,t}$, $\hat{D}_{it}$, may be greater than 1 because $E(D_i|W_i)$ is the model is incorrectly specified by the two-way fixed effects:

$$\hat{\beta}_{post} = \frac{\sum_{i,t}(D_{it} - \hat{D}_{i,t})Y_{i,t}}{\sum_{i,t}(D_{it} - \hat{D}_{i,t})^2}$$

Misspecification in conditional expectation

- An example from GPHK (2022): Three time periods, and three groups of units: treated early in period 2 (n_E), treated late in period 3 (n_L), and never (n_N)

- The weights in the estimands on the treatment effects will be

$$\lambda(j, t = 3) = \frac{n_E + 2n_N}{\kappa}, j \in L \tag{5}$$

$$\lambda(j, t = 2) = \frac{n_N + 2n_L}{\kappa}, j \in E \tag{6}$$

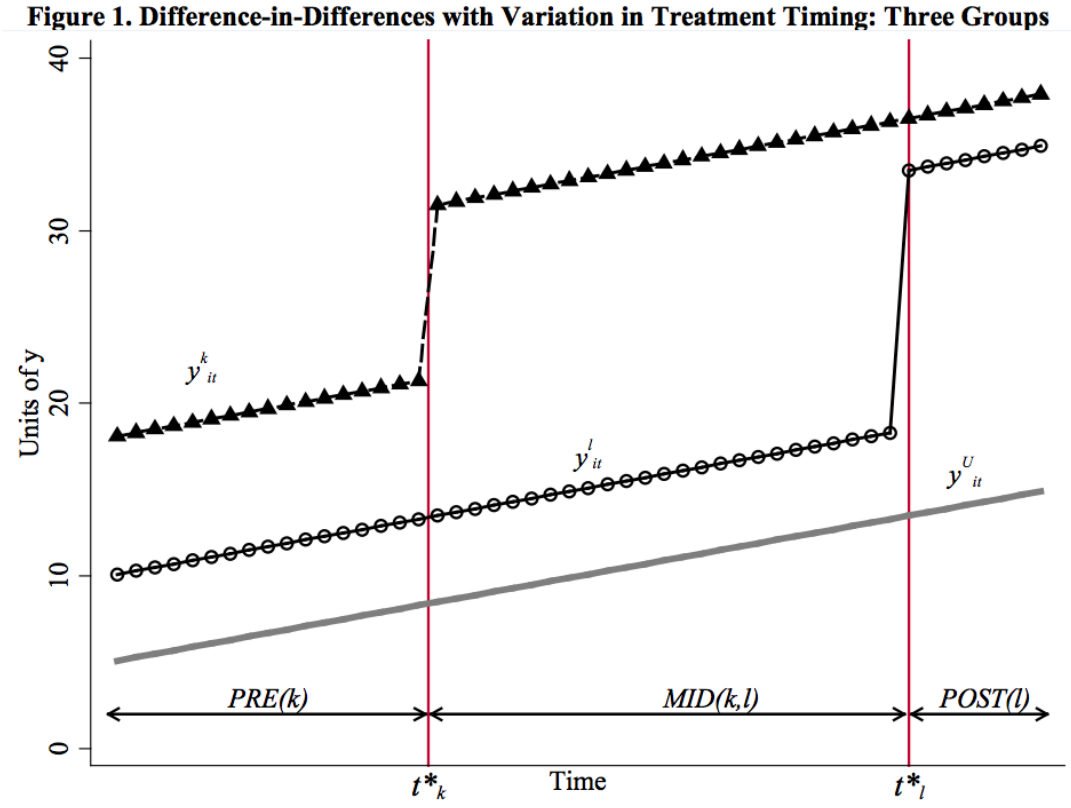
$$\lambda(j, t = 3) = \frac{n_N - n_L}{\kappa}, j \in E, \tag{7}$$

  κ just scaling. Hence, first two always get positive weight, but the ATT from period 3 of the early adopters is negative if the never-adopters is small relative to late adopters

- 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

Goodman-Bacon 2x2 comparisons

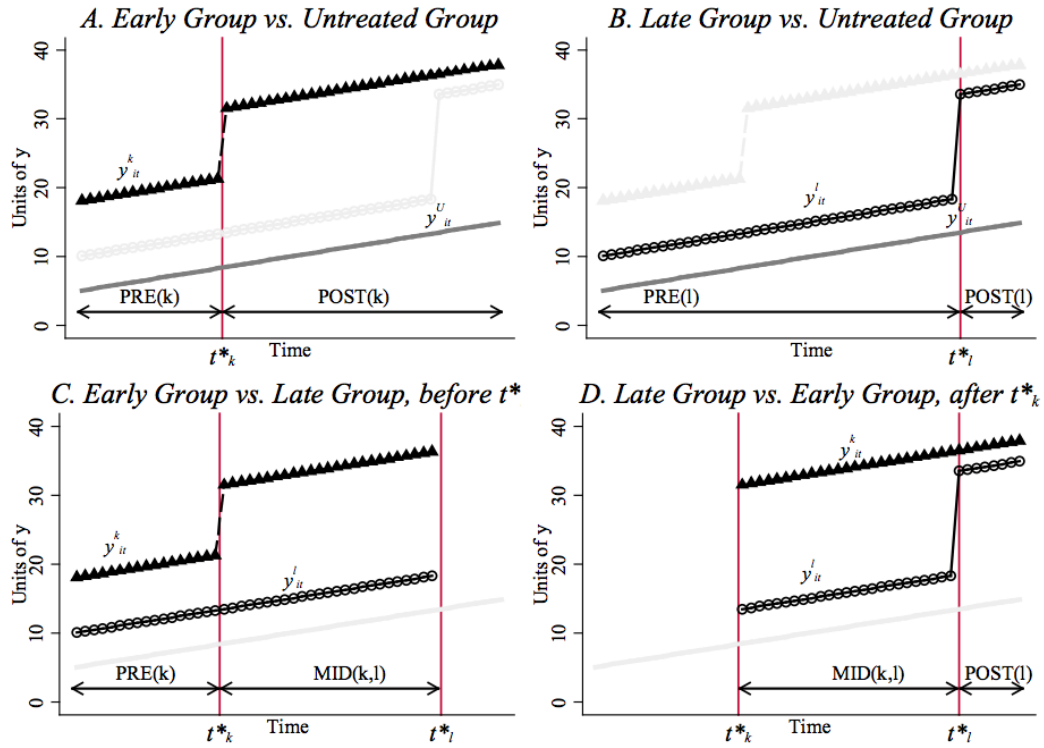
- Consider two staggered treatments and a never-treated group
- What does the TWFE estimator estimate?



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!)

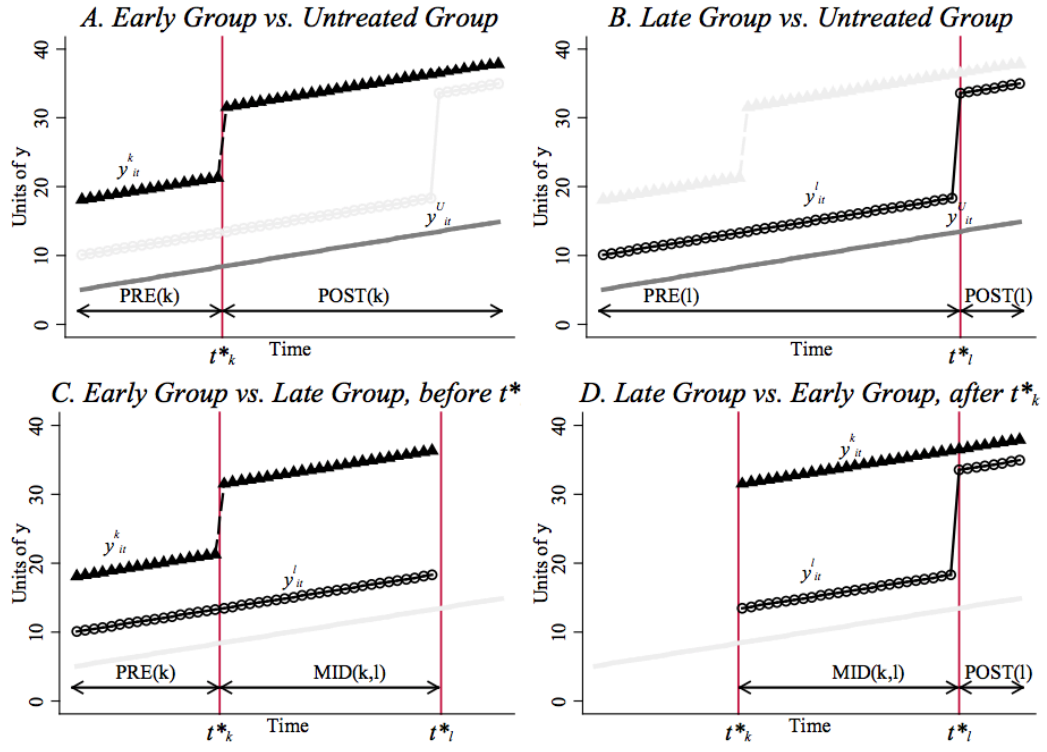
Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



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

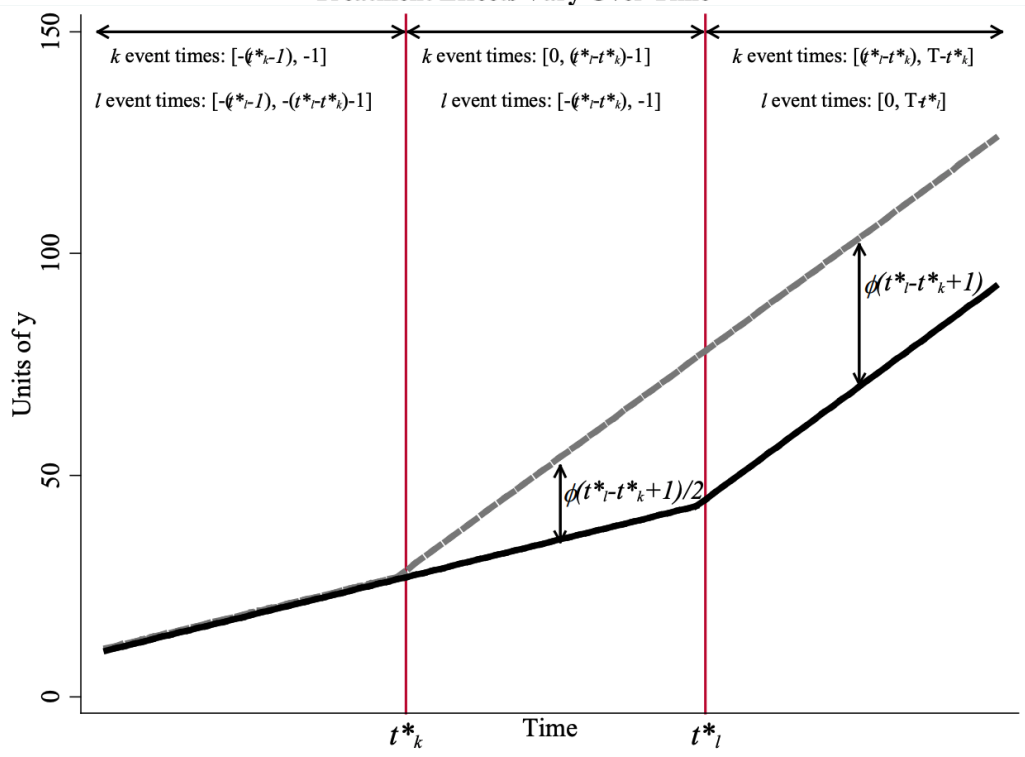
Figure 2. The Four Simple (2x2) Difference-in-Differences Estimates from the Three Group Case



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 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. 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