3. Difference-in-Differences

LPO 8852: Regression II

Sean P. Corcoran

Matching recap

Matching methods seek to construct a comparison group where the conditional independence assumption is satisfied:

$$Y(0), Y(1) \perp\!\!\!\perp D|X$$

That is, conditional on X (or a one-number summary like the propensity score), potential outcomes are independent of treatment status D. If this holds, we can use mean outcomes of the matched comparison group as a stand-in for the treated group counterfactual.

$$\underbrace{E[Y(0)|D=1,X]}_{\text{unobserved}} = \underbrace{E[Y(0)|D=0,X]}_{\text{matched comparison group}}$$

Matching recap

Challenges:

- The conditional independence assumption (selection on observables) is strong! In most settings we have to be concerned about selection on *unobservables*.
- Constructing matched samples is somewhat of an art, and results may be sensitive to specification of the matching model.
- We are typically comparing outcomes at one point in time (e.g., post treatment).

Difference-in-differences

Difference-in-differences (DD) is a research design that most often contrasts *changes over time* for treated and untreated groups. The approach is fruitfully applied to **natural experiments**, settings in which an external force naturally assigns units into treatment and control groups.

Figure: Scott Cunningham's (of *Mixtape* fame) bumper sticker

DD models are often estimated with *panel* data but can also be used with *repeated cross-sections.*

Natural experiments

Examples of natural experiments:

- John Snow's cholera study (1855)
- Natural and other disasters (hurricanes, earthquakes, COVID, 9/11)
- Policy implementation (e.g., graduated drivers license laws, EZ Pass)
- Investments (e.g., school construction)
- Idiosyncratic policy rules (e.g., class size maximum)
- Idiosyncratic differences in location (opposite sides of boundaries)
- Date of birth and eligibility rules

Many natural experiments are analyzed using DD, others are better suited to tools we'll see later.

Chicago high-stakes testing

Do high-stakes accountability policies improve student academic performance?

- A potential "natural experiment": in Chicago, the Illinois State Aptitude Test (ISAT) became "high stakes" in 2002. The test was administered—but was "low stakes"—prior to that year. The test is given in grades 3, 5, and 8.

- This means 5th graders in 2002 were "treated" by the accountability policy while 5th graders in 2001 were not. Neither cohort was treated in 3rd grade.

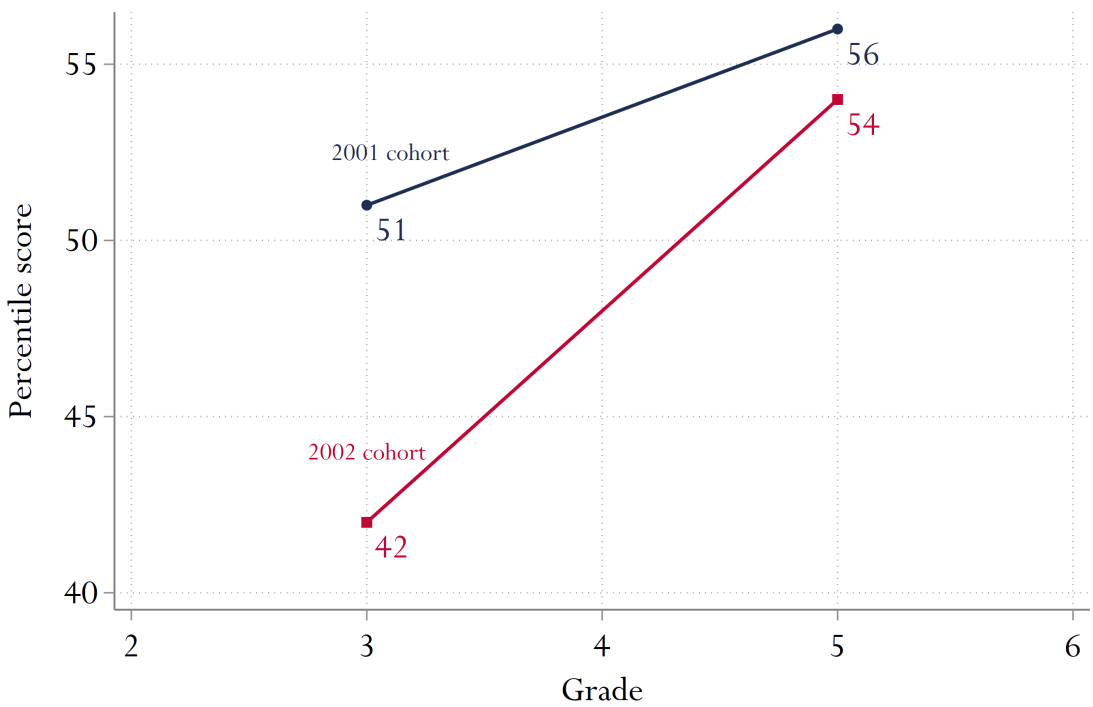
Chicago high-stakes testing

Consider two comparisons:

- "Cross-sectional": the mean scores of 5th graders in 2002 vs. 2001
- First difference or "interrupted time series (ITS)": the change in mean scores of the 2002 5th grade cohort between 3rd and 5th grade

Note a better ITS design would have more data points than two, to establish a trend, but this is just an example!

Chicago high-stakes testing



Chicago high-stakes testing

The cross sectional comparison of 5th graders suggests worse outcomes for the 2002 cohort:

$$Y_{5,2002} - Y_{5,2001} = 54 - 56 = -2$$

The **first difference** for the 2002 cohort suggests a large improvement:

$$Y_{5,2002} - Y_{3,2002} = 54 - 42 = +12$$

Conflicting conclusions!

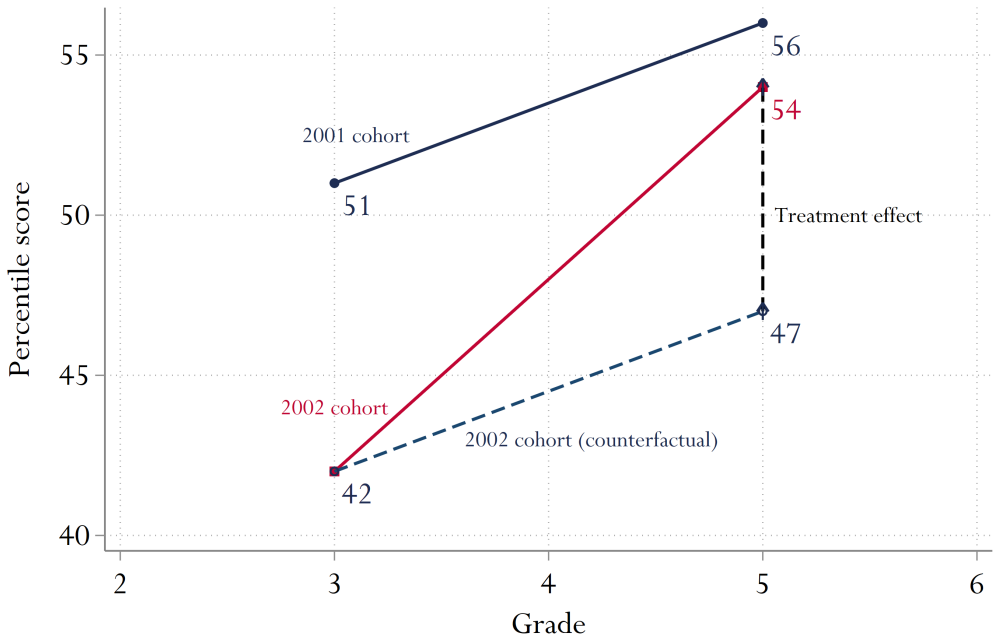
Chicago high-stakes testing

Problems:

- The cross sectional comparison fails to recognize that 5th graders in the 2002 cohort performed worse in 3rd grade than the 2001 cohort did (i.e., *before* treatment).
- The first difference is unable to differentiate between a treatment effect for the 2002 cohort (if any) and improvements between 3rd and 5th grade common to all cohorts.

Chicago high-stakes testing

Under the assumption that improvement in the 2001 cohort represents what *would have happened* to the 2002 cohort in the absence of treatment we can contrast *changes* in the two, or **difference-in-differences**:



Chicago high-stakes testing

The difference-in-differences:

$$\delta_{DD} = \underbrace{(Y_{5,2002} - Y_{3,2002})}_{\text{Change for 2002 cohort}} - \underbrace{(Y_{5,2001} - Y_{3,2001})}_{\text{Change for 2001 cohort}}$$

$$\delta_{DD} = (54 - 42) - (56 - 51) = +7$$

The second term in the above expression is the **second difference**. There was a "counterfactual" gain of 5 implied by the 2001 cohort.

Chicago high-stakes testing

An equivalent way to write δ_{DD}:

$$\delta_{DD} = \underbrace{(Y_{5,2002} - Y_{5,2001})}_{\text{Difference "post"}} - \underbrace{(Y_{3,2002} - Y_{3,2001})}_{\text{Difference "pre"}}$$

Writing δ_{DD} this way makes it clear we are "netting out" pre-existing differences between the two groups.

Note in this example δ_{DD} was calculated using only four numbers (mean scores in the two cohorts, 3rd and 5th grades).

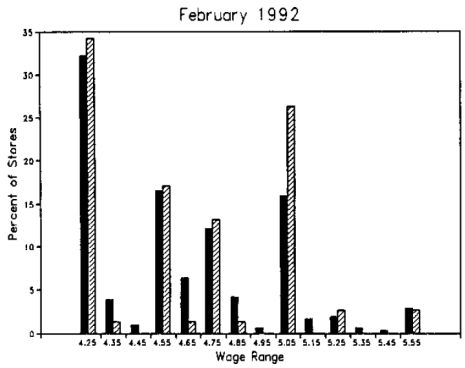
Card & Krueger (1994)

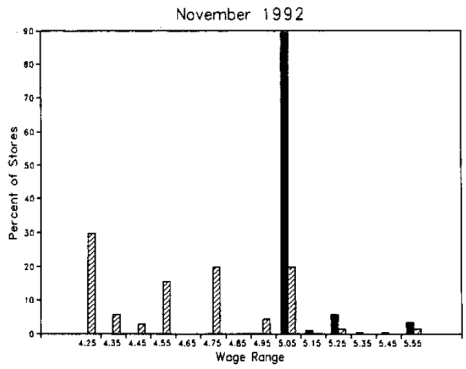
A classic DD study of the impact of the minimum wage on fast food employment (an industry likely to be affected by the minimum wage).

- NJ increased its minimum wage in April 1992, PA did not.
- Card & Krueger collected data on employment at fast food restaurants in NJ and Eastern PA before and after the minimum wage hike.

See next figure: the minimum wage increase had a "first stage." That is, it led to higher starting wages in NJ. (This is important—if the minimum wage were not binding, it wouldn't make for a very interesting study).

Card & Krueger (1994)





New Jersey Pennsylvania

FIGURE 1. DISTRIBUTION OF STARTING WAGE RATES

Card & Krueger (1994)

Main result (portion of Table 3 in C&K):

	Stores by State		
	PA	*NJ*	*NJ* − *PA*
FTE before	23.3	20.44	-2.89
	(1.35)	(-0.51)	(1.44)
FTE after	21.15	21.03	-0.14
	(0.94)	(0.52)	(1.07)
Change in mean FTE	-2.16	+0.59	**2.76**
	(1.25)	(0.54)	**(1.36)**

Standard errors in parentheses. FTE=full time equivalent employees.

Mean employment fell in PA and *rose* in NJ, for $\delta_{DD} = 2.76$. A surprising result to many economists who expected to see a reduction in employment following an increase in the minimum wage.

2x2 difference-in-differences

The two examples thus far are the simplest type of a difference-in-difference:

- Two groups: treated and untreated
- Two time periods: pre and post, before and after treatment occurs
- Treated units are all treated at the same time

Difference-in-differences estimation

Under what conditions might the difference-in-differences design estimate a *causal parameter*? What causal parameter is it estimating?

Let's return to the potential outcomes framework, applying it to a 2x2 DD example.

Difference-in-differences estimation

Suppose that—in the absence of treatment—the potential outcome for individual i at time t is given by:

$$Y_{it}(0) = \gamma_i + \lambda_t$$

In the *presence* of treatment, the potential outcome for individual i at time t is:

$$Y_{it}(1) = \gamma_i + \delta + \lambda_t$$

Note: portions of this section were drawn from Jakiela & Ozier's excellent ECON 626 lecture notes from the University of Maryland (2018).

Difference-in-differences estimation

$$Y_{it}(0) = \gamma_i + \lambda_t$$
$$Y_{it}(1) = \gamma_i + \delta + \lambda_t$$

A few things to note:

- There are fixed individual differences represented by γ_i
- The time-specific factor λ_t is the same for all individuals
- The impact of the treatment δ is assumed to be the same for all individuals, and does not vary over time

$$Y_{it}(1) - Y_{it}(0) = \delta \quad \forall i, t$$

Difference-in-differences estimation

In this framework individuals can self-select into treatment, and selection can be related to γ_i.

- Define $D_i = 1$ for those who—*at any point*—are treated
- Define $D_i = 0$ for those who are never treated

Note this indicator is not subscripted with a t. It is important to note that we are grouping i by whether they are *ever* treated, since we observe them in treated/untreated states at different points in time.

Assume for simplicity two time periods, "pre" ($t = 0$) and "post" ($t = 1$), where treatment occurs for the $D_i = 1$ group in $t = 1$.

Difference-in-differences estimation

The causal estimand of interest is:

$$ATT = \underbrace{E[Y_{it}(1)|D_i = 1, t = 1]}_{\text{observed}} - \underbrace{E[Y_{it}(0)|D_i = 1, t = 1]}_{\textit{unobserved}}$$

$$= E[\gamma_i|D_i = 1] + \delta + \lambda_1 - E[\gamma_i|D_i = 1] - \lambda_1$$

$$= \delta$$

That is, the mean difference in outcomes in the treated and untreated state—in the "post" period—*among those who are treated*.

Difference-in-differences estimation

Of course, we can't observe the same i in two different states (0 and 1) in the same period t. Suppose instead we compare the $D_i = 1$ and $D_i = 0$ groups in time period 1 (post):

$$\underbrace{E[Y_{it}(1)|D_i = 1, t = 1]}_{E[\gamma_i|D_i=1]+\delta+\lambda_1} - \underbrace{E[Y_{it}(0)|D_i = 0, t = 1]}_{E[\gamma_i|D_i=0]+\lambda_1}$$

$$= \delta + \underbrace{E[\gamma_i|D_i = 1] - E[\gamma_i|D_i = 0]}_{\text{selection bias}}$$

If treatment were randomly assigned, the $E[\gamma_i]$ would not vary with D_i. However, if there is selection into D related to the fixed characteristics of individuals, then $E[\gamma_i|D_i = 1] \neq E[\gamma_i|D_i = 0]$. The δ is not identified.

Difference-in-differences estimation

Alternatively we might restrict our attention to the $D_i = 1$ group and do a pre-post comparison from time 0 to time 1:

$$\underbrace{E[Y_{it}(1)|D_i = 1, t = 1]}_{E[\gamma_i|D_i=1]+\delta+\lambda_1} - \underbrace{E[Y_{it}(0)|D_i = 1, t = 0]}_{E[\gamma_i|D_i=1]+\lambda_0}$$

$$= \delta + \lambda_1 - \lambda_0$$

This is the first difference or simple interrupted time series (ITS). Unfortunately, δ is still not identified, since this difference reflects both the impact of the program and the time trend.

Difference-in-differences estimation

Consider now the pre-post comparison for the $D_i = 0$ group:

$$\underbrace{E[Y_{it}(0)|D_i = 0, t = 1]}_{E[\gamma_i|D_i=0]+\lambda_1} - \underbrace{E[Y_{it}(0)|D_i = 0, t = 0]}_{E[\gamma_i|D_i=0]+\lambda_0}$$

$$= \lambda_1 - \lambda_0$$

The comparison group allows us to estimate the time trend!

Difference-in-differences estimation

Now subtract the pre-post comparison for the *untreated* group from the pre-post comparison for the *treated* group:

$$\underbrace{E[Y_{it}(1)|D_i = 1, t = 1]}_{E[\gamma_i|D_i=1]+\delta+\lambda_1} - \underbrace{E[Y_{it}(0)|D_i = 1, t = 0]}_{E[\gamma_i|D_i=1]+\lambda_0} -$$

$$(\underbrace{E[Y_{it}(0)|D_i = 0, t = 1]}_{E[\gamma_i|D_i=0]+\lambda_1} - \underbrace{E[Y_{it}(0)|D_i = 0, t = 0]}_{E[\gamma_i|D_i=0]+\lambda_0})$$

$$= (\delta + \lambda_1 - \lambda_0) - (\lambda_1 - \lambda_0)$$

$$= \delta$$

The difference-in-differences estimator recovers the ATT. The **parallel trends assumption** is critical here.

Difference-in-differences estimation

To see this a different way, the ATT again is:

$$ATT = \underbrace{E[Y(1)|D=1,t=1]}_{\text{observed}} - \underbrace{E[Y(0)|D=1,t=1]}_{\textit{unobserved}}$$

The DD estimates:

$$\underbrace{E[Y(1)|D=1,t=1] - E[Y(0)|D=1,t=0]}_{\text{change over time for treated group}}$$

$$-\underbrace{(E[Y(0)|D=0,t=1] - E[Y(0)|D=0,t=0])}_{\text{change over time for untreated group}}$$

From this, subtract and add the *unobserved* term from above right:

Difference-in-differences estimation

$$E[Y(1)|D=1,t=1] - E[Y(0)|D=1,t=0] - \underbrace{E[Y(0)|D_i=1,t=1]}_{\text{unobserved}}$$

$$-(E[Y(0)|D=0,t=1] - E[Y(0)|D=0,t=0]) + \underbrace{E[Y(0)|D_i=1,t=1]}_{\text{unobserved}}$$

Gathering terms, this equals:

$$ATT + \underbrace{(E[Y(0)|D=1,t=1] - E[Y(0)|D=1,t=0])}_{\text{pre to post change in Y(0) for D=1 group}}$$

$$-\underbrace{(E[Y(0)|D=0,t=1] - E[Y(0)|D=0,t=0])}_{\text{pre to post change in Y(0) for D=0 group}}$$

The second term is counterfactual (unobserved). However if parallel trends holds, the second and third term cancel each other out.

Difference-in-differences estimation

To summarize:

- Changes over time in the $D = 0$ group provide the counterfactual
- Selection into treatment related to fixed unobserved differences is OK
- The outcome *levels* are not important, only the *changes*

DD is probably the most commonly used quasi-experimental design in the social sciences and education.

- Its use precedes the RCT (see Snow cholera example, 1855)
- The "comparative interrupted time series" (CITS) design is similar, though not the same. See Section 3 of the MDRC paper by Somers et al. (2013) for a good delineation between the two in the context of an educational intervention.

Regression difference-in-differences (2x2)

With many units, two groups, and two time periods (pre-post):

$$Y_{it} = \alpha + \beta D_i + \lambda Post_t + \delta(D_i \times Post_t) + u_{it}$$

where $D_i = 1$ for units i who are ultimately treated, and $POST_t = 1$ for observations in the "post" period.

Very easy to implement in Stata, especially with factor variable notation:
`reg y i.treated##i.post`

Regression difference-in-differences (2x2)

How does this map onto our earlier notation? There are four expectations estimated in this regression:

$$
\begin{aligned}
E[Y_{it}|D_i = 0, t = 0] &= \alpha \\
E[Y_{it}|D_i = 1, t = 0] &= \alpha + \beta \\
E[Y_{it}|D_i = 0, t = 1] &= \alpha + \lambda \\
E[Y_{it}|D_i = 1, t = 1] &= \alpha + \beta + \lambda + \delta
\end{aligned}
$$

- α is the pre-period mean for the $D_i = 0$ group
- $\alpha + \beta$ is the pre-period mean for the $D_i = 1$ group
- β is the baseline mean difference between the $D_i = 0$ and $D_i = 1$
- $\alpha + \lambda$ is the *post*-period mean for the $D_i = 0$ group
- λ is the change over time for the $D_i = 0$ group
- $\alpha + \beta + \lambda + \delta$ is the *post*-period mean for the $D_i = 1$ group
- $\lambda + \delta$ is the change over time for the $D_i = 1$ group

Regression difference-in-differences (2x2)

The four expectations being estimated in this regression and their differences:

	Pre ($t = 0$)	Post ($t = 1$)	Diff
Untreated ($D = 0$)	α	$\alpha + \lambda$	λ
Treated ($D = 1$)	$\alpha + \beta$	$\alpha + \beta + \lambda + \delta$	$\lambda + \delta$
Diff	β	$\beta + \delta$	δ

Regression (2x2) DD is effectively a comparison of four cell-level means.

Regression difference-in-differences (2x2)

The 2x2 DD regression:

- Is estimating a CEF since the model is fully saturated.
- The CEF is not necessarily causal (depends on parallel trends).
- OLS will always (mechanically) estimate δ as the differential change in the $D_i = 1$ vs. $D_i = 0$ group.
- Whether that δ can be interpreted as the ATT depends on the parallel trends assumption.

Regression difference-in-differences (2x2)

With panel data we could estimate a regression using first differences for each observation i, subtracting Y_{i0} from Y_{i1} (again assuming 2 periods):

$$Y_{i1} = \alpha + \beta D_i + \lambda + \delta(D_i) + u_{i1}$$
$$Y_{i0} = \alpha + \beta D_i + u_{i0}$$

$$Y_{i1} - Y_{i0} = \lambda + \delta D_i + \epsilon_{it}$$

$$\Delta Y_i = \lambda + \delta D_i + \epsilon_{it}$$

This regression is equivalent to the standard DD regression shown earlier. The intercept here represents the time trend λ, and δ is the DD. The baseline differences wash out in the first difference (Δ)

Regression difference-in-differences (2x2)

The 2x2 regression model can also include covariates:

$$Y_{it} = \alpha + \beta D_i + \lambda Post_t + \delta(D_i \times POST_t) + \mathbf{X_{it}}\eta + u_{it}$$

Thought should be put into the use of covariates (more on this later). Does the parallel trends assumption hold conditional on covariates? Or unconditionally?

Example: Dynarski (2003)

Prior to 1982, 18- to 22-year old children of deceased Social Security beneficiaries were eligible for survivor's benefits that could be applied toward college. This practice ended in 1982. Dynarski (2003) used this policy change to estimate the effect of financial aid on college enrollment.

- Table 8.1 from Murnane & Willett on next page begins with the ITS design, focusing only on survivors (a first difference)
- Data: NLSY high school seniors who would be eligible for benefits just before (N=137) and after (N=54) the policy change.

Note: treatment in this case (benefits) occurs *before* 1982, not after (*offer*=1 for the earlier cohort).

Example: Dynarski (2003)

Table 8.1 "First difference" estimate of the causal impact of an offer of $6,700 in financial aid (in 2000 dollars) on whether high-school seniors whose fathers were deceased attended college by age 23 in the United States

(a) *Direct Estimate*

H.S. Senior Cohort	Number of Students	Was Student's Father Deceased	Did H.S. Seniors Receive an Offer of SSSB Aid?	Avg Value of *COLL* (standard error)	Between-Group Difference in Avg Value of *COLL*	H_0: $\mu_{OFFER} = \mu_{NO\ OFFER}$	
						t-statistic	*p*-value
1979–81	137	Yes	Yes (*Treatment Group*)	0.560 (0.053)			
					0.208*	2.14	0.017†
1982–83	54	Yes	No (*Control Group*)	0.352 (0.081)			

˜p <0.10; * p <0.05; ** p <0.01; *** p <0.001.
†One-tailed test.

(b) *Linear-Probability Model (OLS) Estimate*

Predictor	Estimate	Standard Error	H_0: $\beta = 0$;	
			t-statistic	*p*-value
Intercept	0.352***	0.081	4.32	0.000
OFFER	0.208*	0.094	2.23	0.013†
R^2	0.036			

˜p <0.10; * p <0.05; ** p <0.01; *** p <0.001.
†One-tailed test.

Example: Dynarski (2003)

Table 8.2 from Murnane & Willett reports the DD estimate, incorporating data for high school seniors that were not survivors, before (N=2,745) and after (N=1,050) the policy change—a second difference.

Example: Dynarski (2003)

Table 8.2 Direct "difference-in-differences" estimate of the impact of an offer of $6,700 in financial aid (in 2000 dollars) on whether high-school seniors whose fathers were deceased attended college by age 23, in the United States

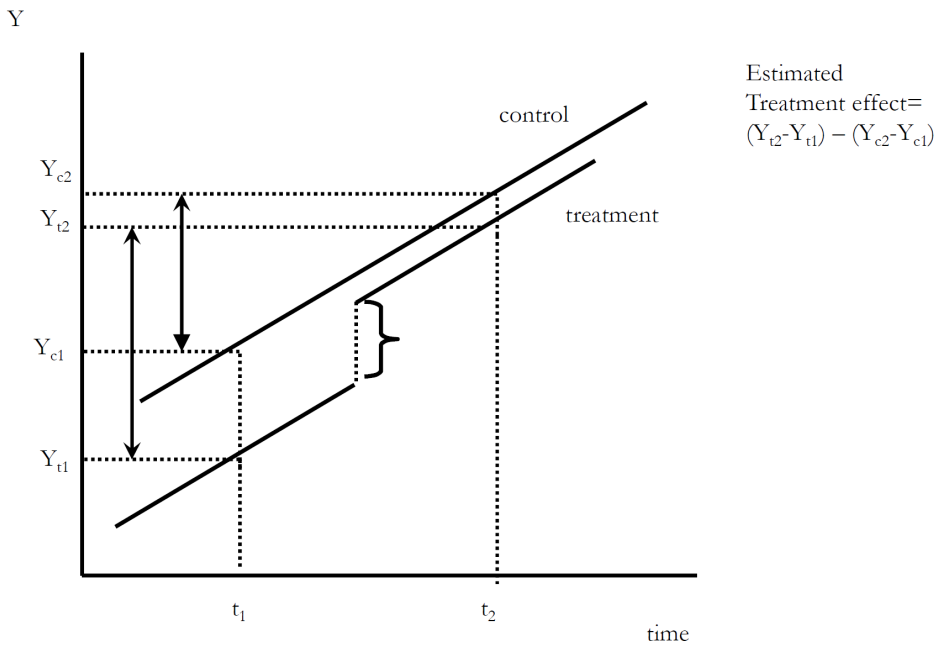
H.S. Senior Cohort	Number of Students	Was Student's Father Deceased?	Did H.S. Seniors Receive an Offer of SSSB Aid?	Avg Value of *COLL* (*standard error*)	Between-Group Difference in Avg Value of *COLL*	"Difference in Differences"	
						Estimate (*standard error*)	*p*-value
1979–81	137	Yes	Yes (*Treatment Group*)	0.560 (0.053)			
					0.208 (*First Diff*)		
1982–83	54	Yes	No (*Control Group*)	0.352 (0.081)			
						0.182* (0.099)	0.033†
1979–81	2,745	No	No	0.502 (0.012)			
					0.026 (*Second Diff*)		
1982–83	1,050	No	No	0.476 (0.019)			

˜ p <0.10; * p <0.05; ** p <0.01; *** p <0.001.
†One-tailed test.

Note: *COLL* went down even for non-survivors.

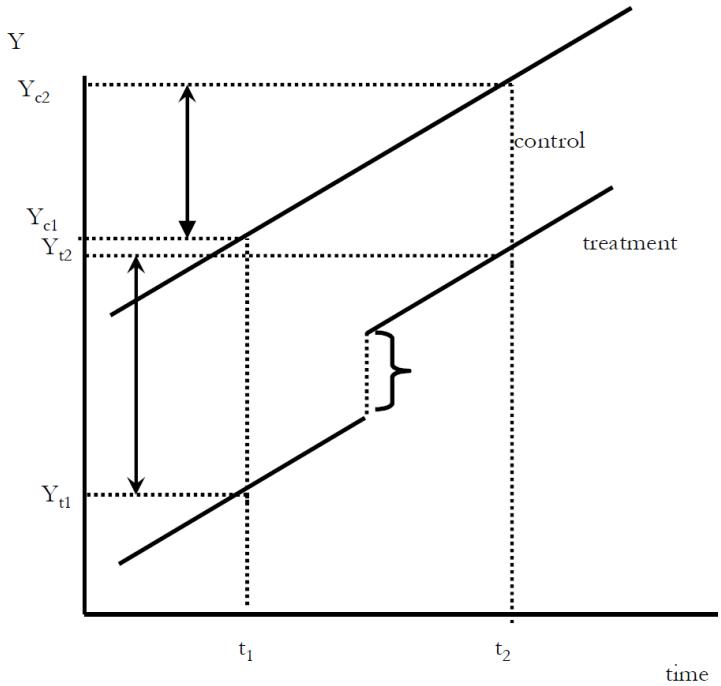
Parallel trends assumption

The key assumption in DD is parallel trends: that the time trend in the absence of treatment would be the same in both groups.



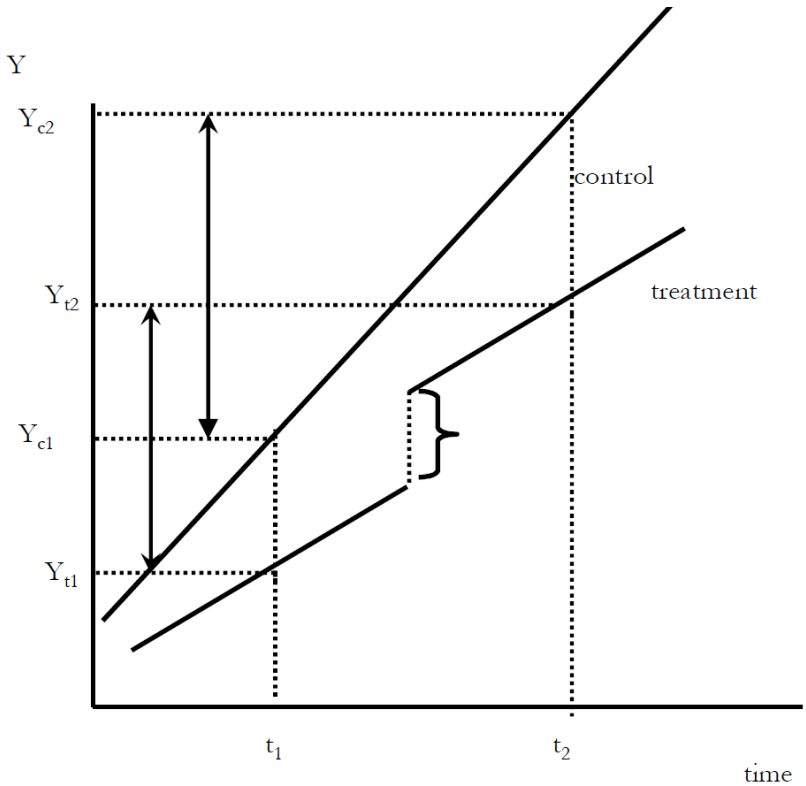
Parallel trends assumption

Size of baseline difference in treated and untreated groups doesn't matter.



Estimated
Treatment effect=
$(Y_{t2}-Y_{t1}) - (Y_{c2}-Y_{c1})$

Violation of parallel trends assumption



Estimated
Treatment effect=
$(Y_{t2}-Y_{t1}) - (Y_{c2}-Y_{c1})$

Parallel trends assumption

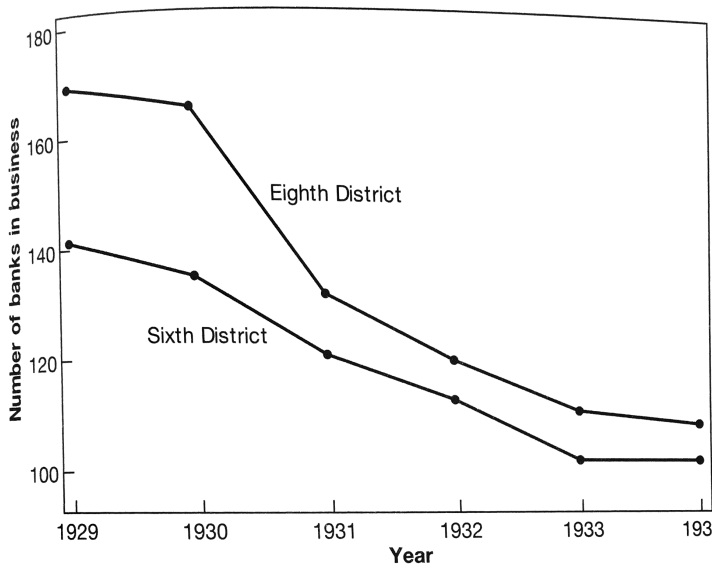
We can't verify the parallel trends assumption directly, but researchers typically defend it in a variety of ways:

- A compelling graph: pointing to similar trends prior to the treatment. Note: common trends prior to treatment are neither necessary nor sufficient for parallel trends assumption!
- Event study regression and graph
- A placebo / falsification test
- Controlling for time trends directly (leans heavily on functional form)
- Triple-difference model
- Probably most important: understanding the context of your study! Ruling out reasons for non parallel-trends.

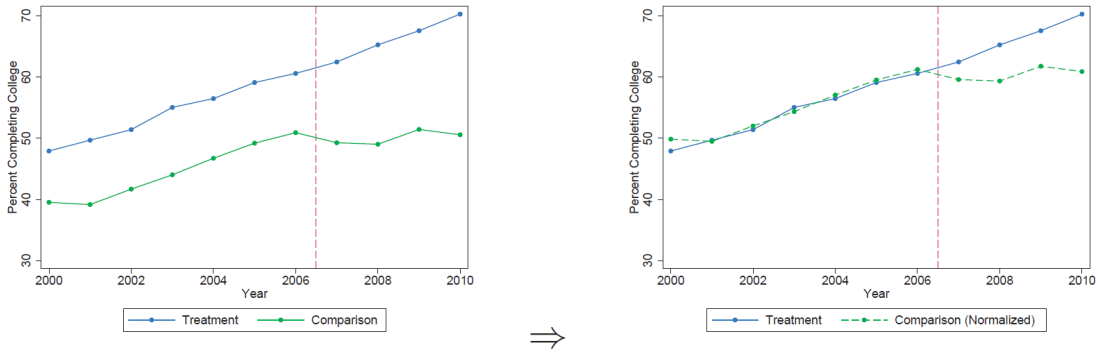
Federal Reserve policy and bank failures

From Mastering 'Metrics chapter 5—treatment in 1930.

FIGURE 5.2
Trends in bank failures in the Sixth and Eighth Federal Reserve Districts

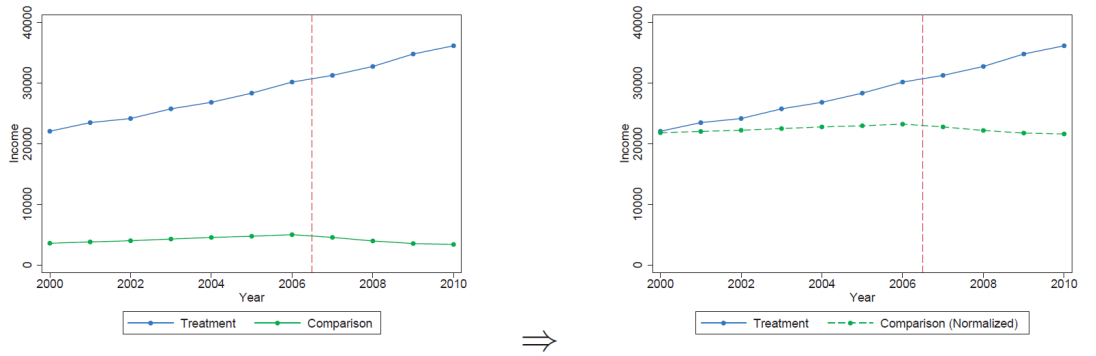


Checking the parallel trends assumption (1)



The graph on the right ("normalized") subtracts baseline difference between Treated and Comparison group, to help see the parallel trend.

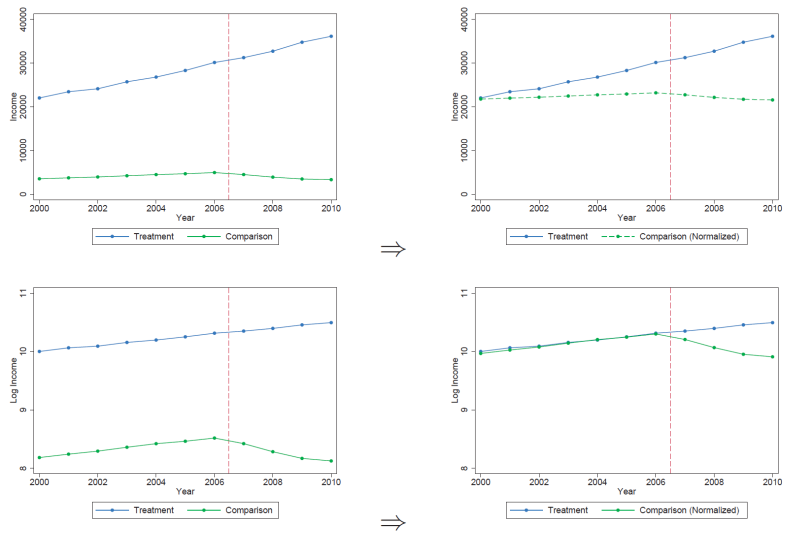
Checking the parallel trends assumption (2)



The graph on the right ("normalized") makes the lack of a parallel trend more visually apparent than the graph on the left.

Checking the parallel trends assumption (3)

A variable transformation may help satisfy the parallel trends assumption. The bottom panels use the *log*:



Note: if trends are parallel in levels they will *not* be parallel in logs, and vice versa!

Event study

An **event study** is like the DD regression shown earlier, except it includes separate time and treated group interactions for all pre and post periods.

With 2 groups and observations J periods before treatment (*lags*) and K periods after treatment (*leads*). Assume treatment occurs at $t = 0$:

$$Y_{it} = \alpha + \beta D_i + \sum_{\tau=-J}^{K} I[t = \tau]\lambda_\tau + \sum_{\tau=-J}^{K} I[t = \tau]D_i\gamma_\tau + u_{it}$$

$I[\ ]$ is the indicator function. It "ticks on" whenever $t = \tau$ and is zero otherwise. One time period needs to be omitted—it is convention to omit $t = -1$, the last period before treatment.

Event study

$$Y_{it} = \alpha + \beta D_i + \sum_{\tau=-J}^{K} I[t=\tau]\lambda_\tau + \sum_{\tau=-J}^{K} I[t=\tau]D_i\gamma_\tau + u_{it}$$

One year before treatment (omitted time period):

- $E(Y|D=0, t=-1) = \alpha$
- $E(Y|D=1, t=-1) = \alpha + \beta$ (the groups differ by β at $t-1$)

Two years before treatment:

- $E(Y|D=0, t=-2) = \alpha + \lambda_{-2}$
- $E(Y|D=1, t=-2) = \alpha + \lambda_{-2} + \beta + \gamma_{-2}$ (γ_{-2} is the *additional* difference between groups at $t-2$)

Event study

$$Y_{it} = \alpha + \beta D_i + \sum_{\tau=-J}^{K} I[t=\tau]\lambda_\tau + \sum_{\tau=-J}^{K} I[t=\tau]D_i\gamma_\tau + u_{it}$$

First year of treatment (time 0):

- $E(Y|D=0, t=0) = \alpha + \lambda_0$
- $E(Y|D=1, t=0) = \alpha + \lambda_0 + \beta + \delta_0$ (δ_0 is the *additional* difference between groups at time 0)

Second year of treatment (time 1):

- $E(Y|D=0, t=1) = \alpha + \lambda_1$
- $E(Y|D=1, t=1) = \alpha + \lambda_1 + \beta + \delta_1$ (δ_1 is the *additional* difference between groups at $t+1$)

And so on!

Event study

The γ_j capture the difference between the treated and untreated groups, compared to their prevailing difference in the omitted base period (the year before treatment, $t = -1$). Note any year can be the base period, but it is common to use $t = -1$ as the reference year.

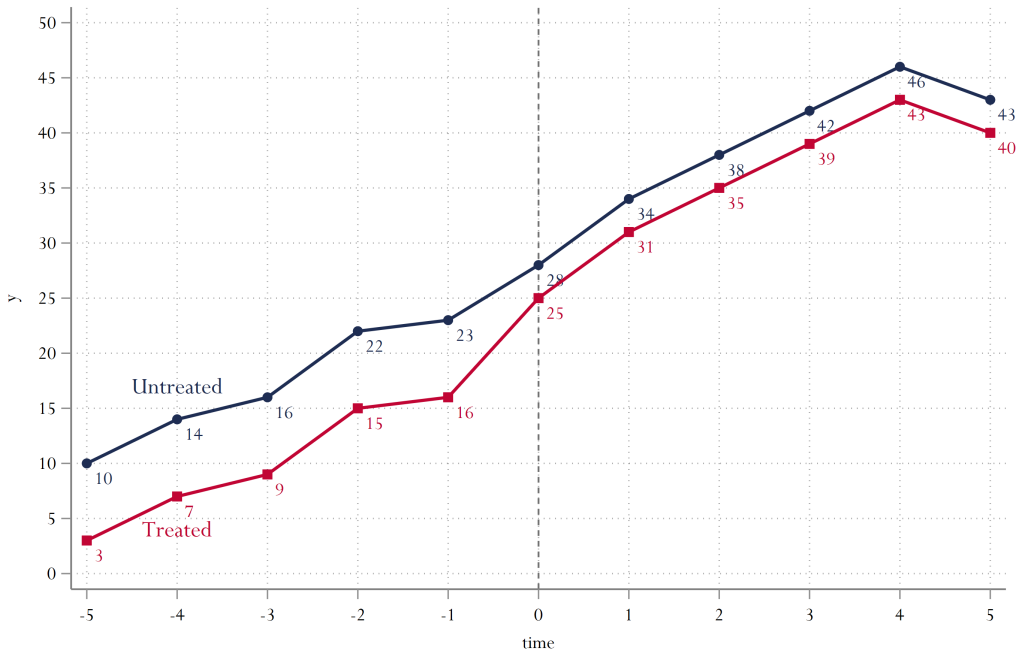
If the coefficients on the pre-treatment (lag) interaction terms are significantly different from zero, this suggests a non-parallel trend before treatment.

The event study *graph* is a plot of these γ_j, along with confidence intervals.

Note: consider *practically* significant differences, not just statistically significant. With large samples, even small differences can be statistically significant.

Event study

Stylized example: suppose these are the mean y in each time period for two groups:



Event study

In this stylized example, pre-treatment:

- $\lambda_{-2} = 22 - 23 = -1$
- $\lambda_{-3} = 16 - 23 = -7$
- $\lambda_{-4} = 14 - 23 = -9$ ... etc.
- $\gamma_{-2} = (15 - 16) - (22 - 23) = 0$
- $\gamma_{-3} = (9 - 16) - (16 - 23) = 0$
- $\gamma_{-4} = (7 - 16) - (14 - 23) = 0$ .. etc.
- Here, the two groups have identical pre-trends.

Event study

In this stylized example, post-treatment:

- $\lambda_0 = 28 - 23 = 5$
- $\lambda_1 = 34 - 23 = 11$
- $\lambda_2 = 38 - 23 = 15$ ... etc.
- $\gamma_0 = (25 - 16) - (28 - 23) = 4$
- $\gamma_1 = (31 - 16) - (34 - 23) = 4$
- $\gamma_2 = (35 - 16) - (38 - 23) = 4$ .. etc.
- The treatment effect appears in time 0 and remains in subsequent periods.

Event study

If there are two groups (*treated* and not) and treatment occurs for everyone in the $D_i = 1$ group in the same time period, this is easy to implement in Stata (assume *year* is the time period):

```
reg y i.treated##i.year
```

This is a full factorial of treatment group status and time period—includes main effects for *treated* and individual years, and their interaction. You will need to specify the omitted time period. For example, if treatment occurred in 2006 and you wish to use 2005 as the reference year:

```
reg y i.treated##ib2005.year
```


Event study example

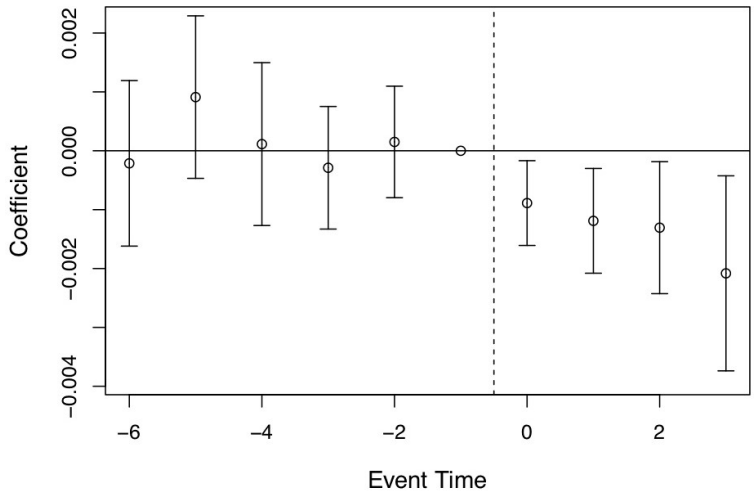
The following figures are from Miller et al. (QJE 2021), via the *Mixtape*. The authors estimate the impact of state expansion of Medicaid under ACA on the annual mortality rates of older persons under 65 in the U.S.

A causal interpretation of DD assumes changes over time in states that did *not* expand Medicaid provide the counterfactual for those that did.

They find a 0.13 percentage-point decline in annual mortality, a 9.3% reduction over the sample mean, as a result of Medicaid expansion.

Event study example

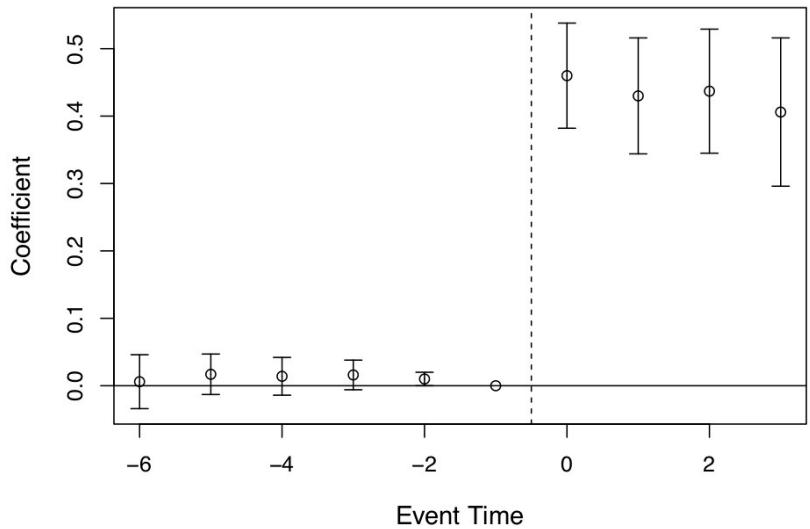
Plotted points are event study coefficients, shown with 95% confidence intervals. (Time zero is the first year of expansion). Outcome: mortality rate



There is no evidence these states were on different trajectories prior to Medicaid expansion.

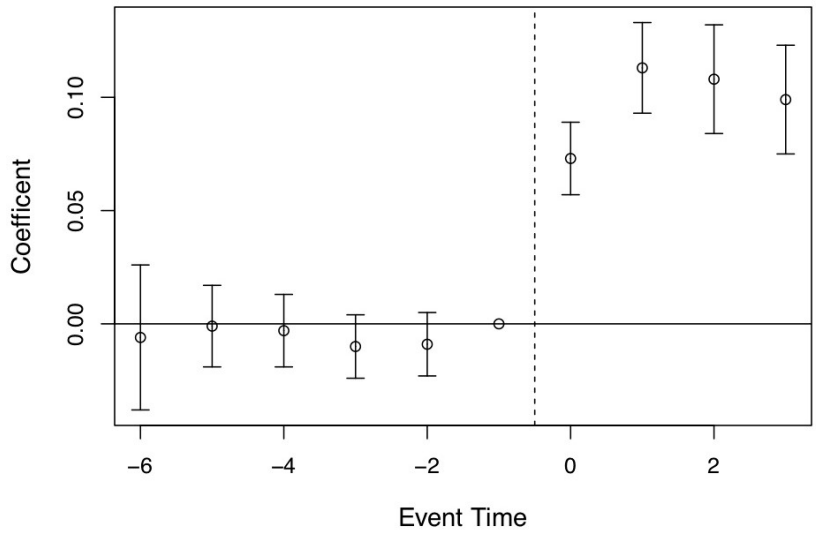
Event study example

The authors first look for a "first stage": did the expansion of Medicaid actually increase rates of eligibility for Medicaid? Did it increase Medicaid coverage? Did it lower the uninsured rate? Here: eligibility



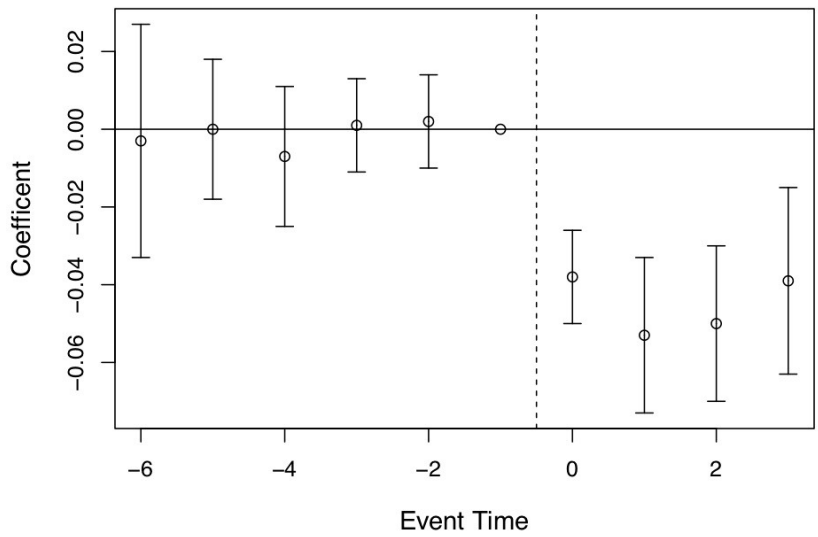
Event study example

Here: Medicaid coverage rates



Event study example

Here: Medicaid uninsured rates



Taken together, these graphs are compelling: Medicaid expansion increased eligibility and coverage, and reduced the uninsured. One would hope to see these first stage effects before expecting an effect on health outcomes.

Event study

Most event studies are not as simple as two groups and one common treatment period. The user-written Stata package `eventdd` is a flexible solution that automatically generates the needed variables, estimates the regression, and produces a graph. Example syntax:

`eventdd` *`y x1 x2 i.year i.state`*`, timevar(`*`timetoevent`*`)`

This syntax estimates an OLS model, and the group (*state*) and time (*year*) variables are included in the covariates. The key variable here is *timetoevent* (a name you provide), defined as the relative time to treatment. 0 corresponds to the first year of treatment, -1 refers to the first lag, and so on. This variable should be **missing for groups that are never treated**. See Clarke and Schythe (2020).

Parallel trends assumption

When covariates are included in the model, the parallel trends assumption is *conditional* on the covariates. It is possible that the unconditional outcomes do not follow a parallel trend, but the conditional outcomes do.

Put another way, controlling for covariates allows you to account for factors that might produce different time trends.

Common violations of parallel trends assumption

Two common scenarios that would violate the parallel trends assumption:

- Targeted treatments: often programs are targeted at subjects who are most likely to benefit from it. In many cases, the fact that a subject was on a different trajectory is what made them a good candidate for the program (e.g., a struggling student).
- Ashenfelter's dip: treated cases may experience a "dip" just prior to treatment that results in a reversion to the mean after treatment (e.g, job training).
- Anticipation: behavior (and outcomes) change prior to treatment due to anticipation effects.

Placebo/falsification tests

The DD design assumes that any change over time beyond that predicted by the untreated group is the ATT, and not some other time-varying factor specific to the treated group.

If there is indeed an unobserved time-varying factor specific to the treated group, one might see its effects show up on *other* outcomes that shouldn't have been affected by the treatment.

- Card & Krueger: employment in higher-wage firms
- Miller et al.: mortality of populations not eligible for Medicaid
- Cheng & Hoekstra (2013): effects of Stand Your Ground laws on other non-homicide crimes (see *Mixtape*)

Estimate the same DD model for these outcomes. If there is an "effect", this may indicate an unobserved, time-varying confounder specific to the treated group.

Placebo/falsification tests

Another approach is to apply the same treatment assignment to an earlier period, well before the treatment actually occurred, and re-estimate the DD model on this earlier data. If there is an apparent treatment "effect" in these untreated years, there may well be unobserved, group-specific trends driving the result.

There are lots of ways to do this, including picking your own period for the "fake" treatment, or trying lots of alternatives.

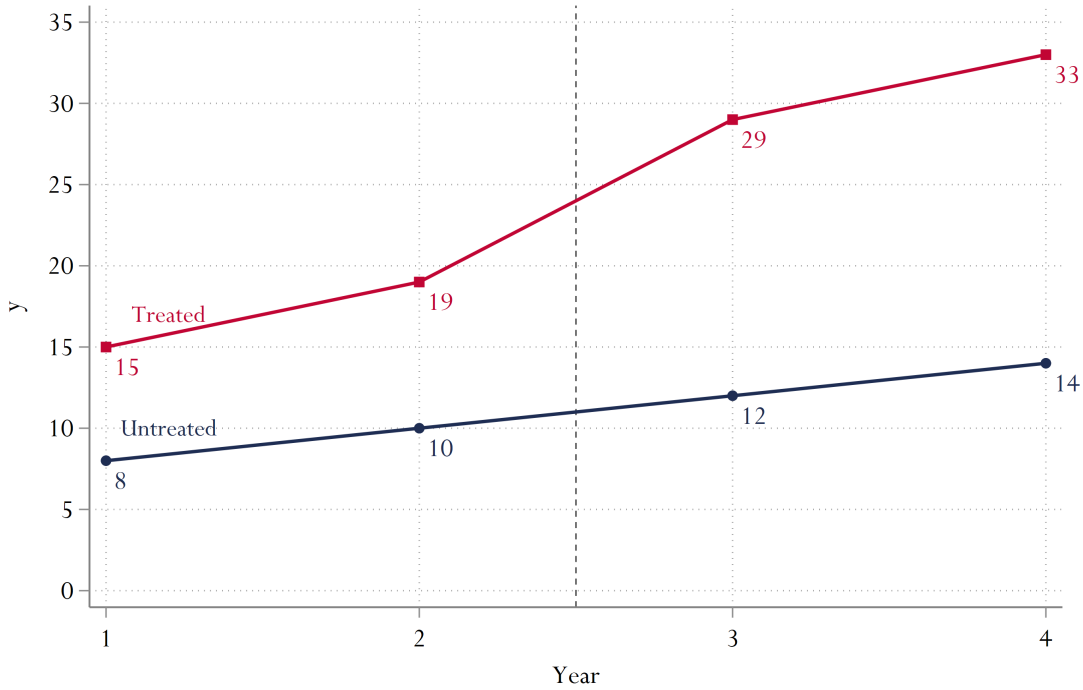
Triple difference

The **triple difference** uses an additional untreated group to difference out time trends unique to the treatment group that are also experienced by the added untreated group. For example:

- In C&K, suppose we were concerned that the (treated) state of NJ was on a different time trend from the (untreated) state of PA.
- The lack of parallel trends could make DD invalid.
- The minimum wage treatment should only affect *low-wage* workers.
- We might be able to contrast *higher-wage* workers in NJ and PA to identify any differential time trend in NJ.
- The treatment effect of the minimum wage on low wage workers would be any *additional* change over time experienced by low-wage workers in NJ.

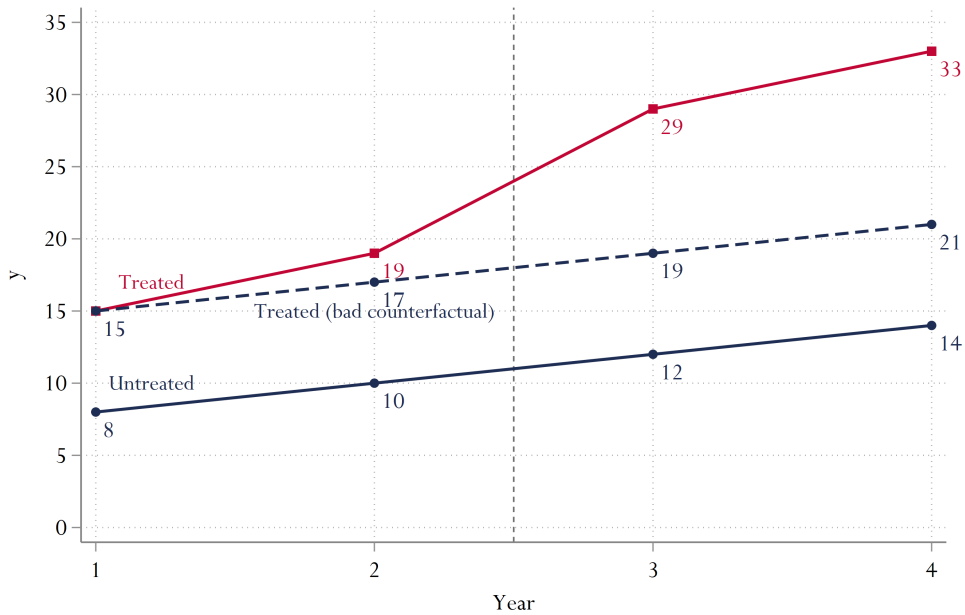
Triple difference

Stylized example: non-parallel trends



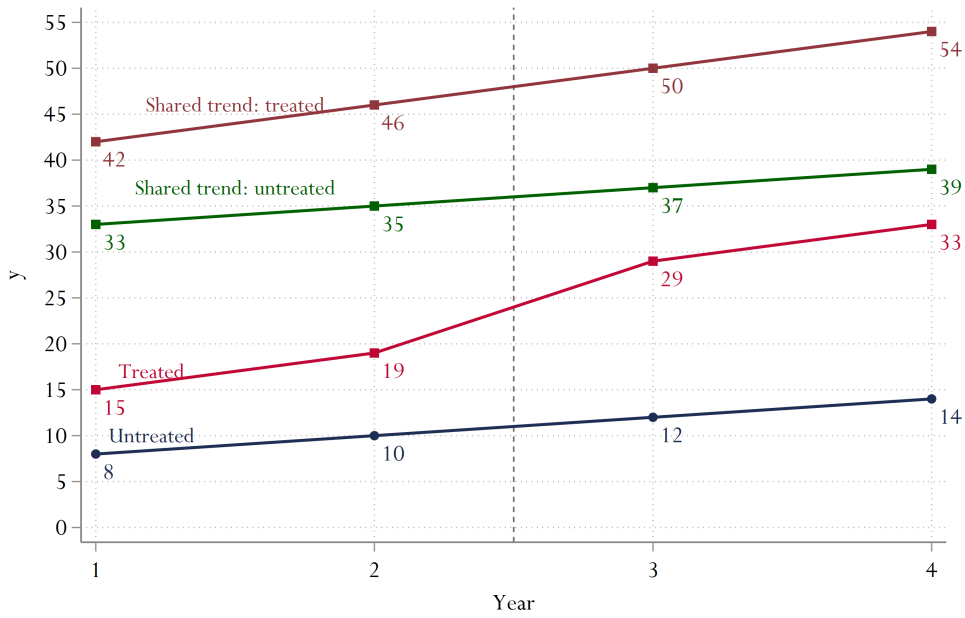
Triple difference

The untreated group is a bad counterfactual for the DD. Focusing only on time points 4 and 1, the DD estimate $(33 - 15) - (14 - 8) = 12$ overstates the treatment effect due to the differential time trend.



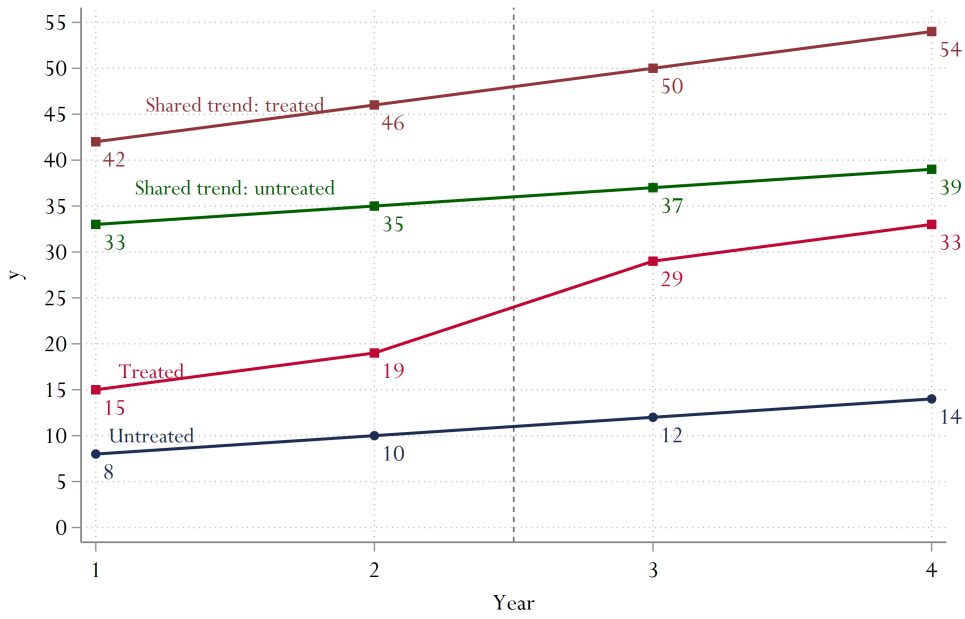
Triple difference

Suppose we have other untreated groups who share the time trends of the original treated and untreated groups.



Triple difference

From these groups we can estimate the differential time trend (again focusing on times 4 and 1): $(54 - 42) - (39 - 33) = 6$. Subtract from the original DD estimate to isolate the treatment effect: $12 - 6 = 6$



Triple difference

Consider the Card and Krueger minimum wage study:

- One might be concerned that the parallel trends assumption fails to hold for NJ and PA employment.
- If that is the case, the DD includes both the treatment effect *and* the differential time trend.
- Looking at a related group of *untreated* workers in both states might help us isolate the differential time trend.

High-wage workers are unlikely to be affected by the minimum wage law, but might be similarly affected by state-specific time trends.

Triple difference

Let $G = 1$ be the focal group (e.g., low-wage workers) and $G = 0$ be the additional untreated group (e.g., higher-wage workers).

The triple difference regression model is:

$$Y_{it} = \alpha + \beta_1 POST_t + \beta_2 G_i + \beta_3 D_i + \beta_4 (G_i \times POST_t)$$
$$+\beta_5 (D_i \times POST_t) + \beta_6 (G_i \times D_i) + \beta_7 (G_i \times D_i \times POST_t) + u_{it}$$

Dummy variables for: original treated group (D), additional untreated group (G), $POST$, three 2-way interactions, and one 3-way interaction.

Triple difference

First, consider the focal group $G = 1$ (e.g., low-wage workers)

$$Y_{it} = \alpha + \beta_1 POST_t + \beta_2 G_i + \beta_3 D_i + \beta_4 (G_i \times POST_t)$$

$$+\beta_5 (D_i \times POST_t) + \beta_6 (G_i \times D_i) + \beta_7 (G_i \times D_i \times POST_t) + u_{it}$$

$D = 0$ and	$E[Y\|D = 0, t = 0, G = 1]$	$= \alpha + \beta_2$
focal group	$E[Y\|D = 0, t = 1, G = 1]$	$= \alpha + \beta_1 + \beta_2 + \beta_4$
$D = 1$ and	$E[Y\|D = 1, t = 0, G = 1]$	$= \alpha + \beta_2 + \beta_3 + \beta_6$
focal group	$E[Y\|D = 1, t = 1, G = 1]$	$= \alpha + \beta_1 + \beta_2 + \beta_3 + \beta_4$ $+\beta_5 + \beta_6 + \beta_7$

The traditional DD here would be $\beta_5 + \beta_7$, reflecting the differential time trend for $D = 1$ (β_5) *and* the ATT (β_7). The latter is not identified.

Triple difference

Now, consider the non-focal group $G = 0$ (e.g., higher-wage workers)

$$Y_{it} = \alpha + \beta_1 POST_t + \beta_2 G_i + \beta_3 D_i + \beta_4 (G_i \times POST_t)$$

$$+\beta_5 (D_i \times POST_t) + \beta_6 (G_i \times D_i) + \beta_7 (G_i \times D_i \times POST_t) + u_{it}$$

$D = 0$,	$E[Y\|D = 0, t = 0, G = 0]$	$= \alpha$
non-focal group	$E[Y\|D = 0, t = 1, G = 0]$	$= \alpha + \beta_1$
$D = 1$,	$E[Y\|D = 1, t = 0, G = 0]$	$= \alpha + \beta_3$
non-focal group	$E[Y\|D = 1, t = 1, G = 0]$	$= \alpha + \beta_1 + \beta_3 + \beta_5$

The DD for the non-focal group is β_5. The difference between this and the focal group DD ($\beta_5 + \beta_7$) is β_7, the ATT we are looking for.

Put more simply, the β_7 coefficient gives you the triple difference.

Triple difference

Two examples:

- Monarrez, Kisida, and Chingos (2022): looks at the effect of charter schools on segregation. Problem: trends in factors affecting school segregation may differ between high- and low-charter growth districts. They use grade levels that were not affected (or were less affected) by charter competition in a triple difference design.

- Bravata et al. (2021): looks at the effect of school re-openings on COVID-19 infection. Problem: counties that re-opened schools may have different underlying trends from those that didn't. They use households with and without school aged children in a triple difference design.

See handout with excerpts from these studies. See also Olden & Møen (2022) for more on the triple difference.

Generalized difference-in-differences

Most examples thus far had a common treatment period. In practice, "treatment" can occur for different groups at different times.

This brings us to the "generalized difference-in-differences" model, or difference-in-difference with variable timing. Usually estimated as a "two-way fixed effects" (TWFE) model with fixed effects for cross-sectional units (i) and time periods (t). Sometimes written:

$$Y_{it} = \beta_i + \gamma_t + \delta(D_i \times POST_{it}) + u_{it}$$

Note the main effect for D_i is not included. Why?

Generalized difference-in-differences

Mastering 'Metrics: effect of a lower Minimum Legal Drinking Age (MLDA), based on Carpenter & Dobkin (2011).

- Following the 26th Amendment (1971), some states lowered the drinking age to 18
- In 1984, federal legislation pressured states to increase MLDA to 21
- Was a lower MLDA associated with more traffic fatalities among 18-20 year olds?

The authors used panel data (state $\times$ year) to address this question.

Note: this will be the in-class exercise, later

Generalized difference-in-differences

$$Y_{st} = \alpha + \delta(TREAT_s \times POST_t) + \sum_{k=2}^{50} \beta_k STATE_{ks} + \sum_{j=2}^{T} \gamma_t YEAR_{jt} + u_{st}$$

- $STATE_{ks} = 1$ if observation is from state k. States indexed from $k = 2...50$ as a reminder that one state dummy must be omitted.
- $YEAR_{jt} = 1$ if observation is from year j. Years indexed from $j = 2...T$ as a reminder that one time dummy must be omitted.
- β_k is a *state effect*.
- γ_t is a *year effect*.
- Covariates X_{st} may be included to control for other time-varying factors associated with Y and treatment (though see later discussion of this).

Generalized difference-in-differences

- Analogous to the 2x2 model, each group (state) has its own intercept ($\alpha + \beta_k$ for $k = 2, ..., 50$) reflecting baseline differences.
- There need not be a single common post-treatment period. The year effects (γ_t) capture trends in the outcome common to all states.
- The coefficient on the interaction (δ) represents how much, on average, outcomes *differ* in treatment states in the post period from that predicted by the state and year effects.
- In other words, we are contrasting *within-state changes over time* in the outcome, for treated and untreated states.
- This is an example of a fixed effects panel model, a topic covered more in Lecture 4.

Generalized difference-in-differences

Implementing in Stata: can be done in multiple ways, including `xtreg`:

```
xtreg y x i.year i.treat##i.post, i(state) fe
```

`xtreg` is a panel data command where *state* is the cross-sectional unit and `fe` implements the fixed-effects (within) estimator—covered in Lecture 4. Essentially equivalent to separate intercepts for every state.

The fixed effect variable (*state* here) should be numeric. If it is not, can use `encode`.

Alternative Stata command: `reghdfe`

Generalized difference-in-differences

Using our notation for *potential outcomes* for a state k:

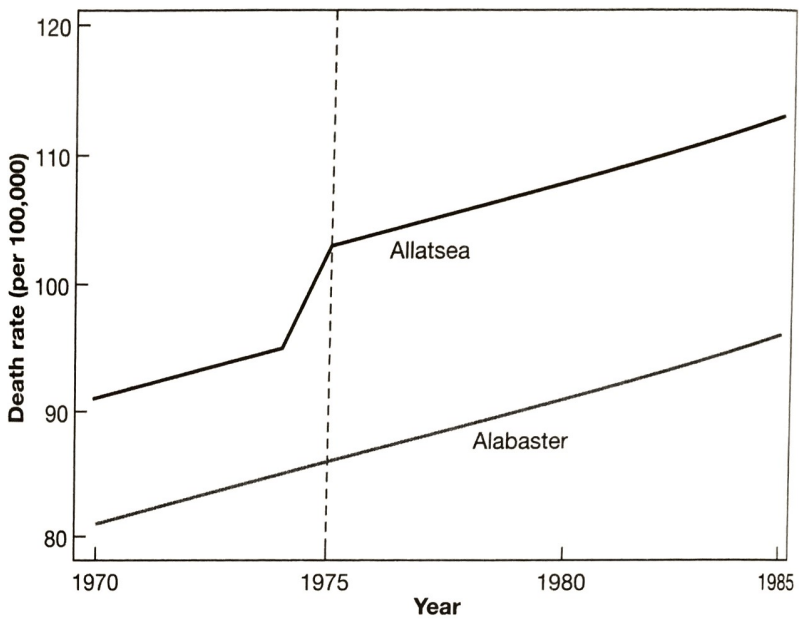
$$Y_{kt}(0) = \alpha + \beta_k + \gamma_t$$
$$Y_{kt}(1) = \alpha + \beta_k + \gamma_t + \delta$$

Potential outcomes are described by a unique intercept for each state $(\alpha + \beta_k)$ and a yearly deviation from this intercept that is common to every state (γ_t). The treatment effect is δ.

Intuitively, under the parallel trends assumption that changes within states over time would be the same in the absence of treatment, we can estimate δ as the *differential* change over time associated with treatment.

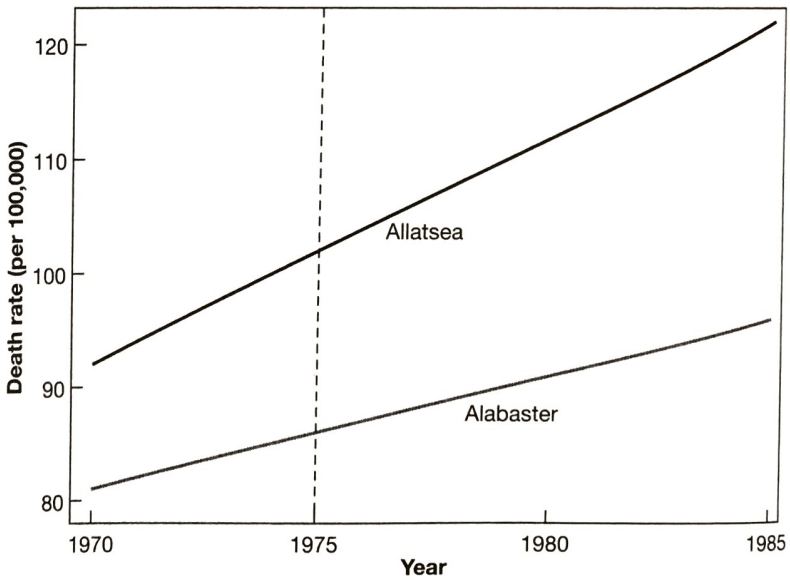
Mastering 'Metrics Fig 5.4

FIGURE 5.4
An MLDA effect in states with parallel trends



Mastering 'Metrics Fig 5.5

FIGURE 5.5
A spurious MLDA effect in states where trends are not parallel



Treatment as "intensity"

Treatment can alternatively be measured as a continuous "intensity" measure. Examples:

$$Y_{it} = \alpha + \delta \text{Intensity}_i + \lambda Post_t + \delta(\text{Intensity}_i \times Post_t) + u_{it}$$

Or:

$$Y_{st} = \alpha + \delta \text{Intensity}_{st} + \sum_{k=2}^{50} \beta_k STATE_{ks} + \sum_{j=2}^{T} \gamma_t YEAR_{jt} + u_{st}$$

Example: Duflo (2001) examined the impact of school construction on educational attainment in Indonesia. The "treatment"—number of area schools per school-aged child—varied from one place to the next.

In-class exercise

Replicate the findings from the MLDA study reported in *Mastering 'Metrics*.

- Generalized DD using two-way fixed effects
- Placebo test using other outcomes, age groups

The treatment in the MLDA study is an "intensity" measure: what proportion of young adults aged 18-20 could legally drink in state s and year t? The measure also accounts for part-year legalization.

Group-specific time trends

With many states and years, we can relax the common trends assumption and allow non-parallel evolution in outcomes (state-specific time trends):

$$Y_{st} = \alpha + \delta(TREAT_s \times POST_t) + \sum_{k=2}^{50} \beta_k STATE_{ks} + \sum_{j=2}^{T} \gamma_t YEAR_{jt} + \sum_{k=2}^{50} \theta_k (STATE_{ks} \times t) + u_{st}$$

Group-specific time trends

How does this work? Consider again the evolution of potential outcomes in two conditions:

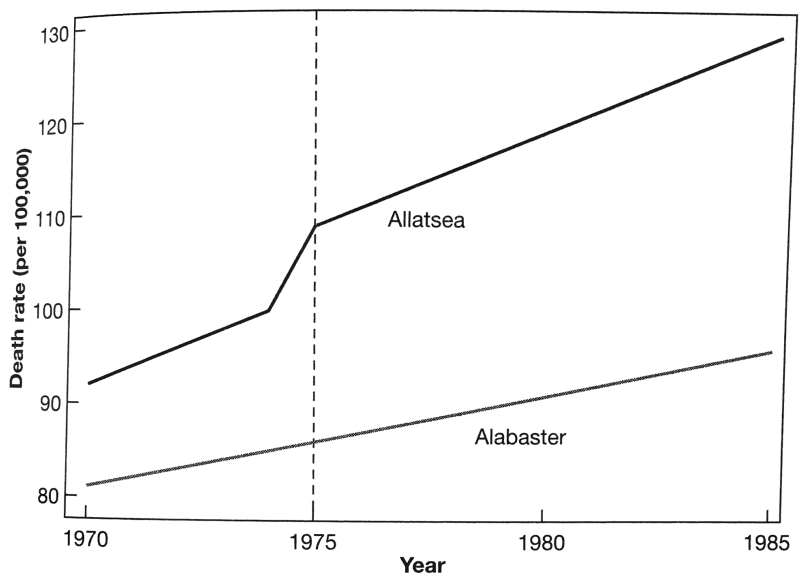
$$Y_{0kt} = \alpha + \beta_k + \gamma_t + \theta_k t$$
$$Y_{1kt} = \alpha + \beta_k + \gamma_t + \theta_k t + \delta$$

Potential outcomes are described by a unique intercept for each state $(\alpha + \beta_k)$ and a yearly deviation from this intercept common to every state (γ_t). Moreover, Y_{kt} deviates from the common year effect according to its own linear trend captured by θ_k.

Intuitively, under the assumption that changes within states over time are accurately described by the common year effect and state-specific time trend, we can estimate δ as the *difference* associated with treatment.

Group-specific time trends

FIGURE 5.6
A real MLDA effect, visible even though trends are not parallel



Group-specific time trends

Here, treatment effects are estimated from sharp deviations from trend, even when not common to other states.

Downside: treatment effect estimates using group-specific time trends are likely to be less precise

Difference-in-differences in other contexts

The DD need not be limited to groups observed in different time periods. The two factors can be anything that define a "treatment" group and are useful for "netting out" unobserved differences that may exist between the treated and untreated:

For example: In a rural poverty reduction program there are program and non-program villages (treated and untreated), and then within these villages, targeted and non-targeted groups. Only targeted groups in program villages are treated. Differencing the outcomes of the non-targeted groups across program and non-program villages can be useful in accounting for unobserved differences between villages.

Difference-in-differences in other contexts

Stylized example: high-poverty households were targeted for the program

	High-Poverty	Low-Poverty
Program village	400	
Non-program village	300	

The cross-sectional comparison is $\bar{Y}_P - \bar{Y}_{NP} = 400 - 300 = 100$. Selection bias is possible if villages were not randomly assigned.

Difference-in-differences in other contexts

Use low-poverty households for the second difference:
$(\bar{Y}_{Ph} - \bar{Y}_{NPh}) - (\bar{Y}_{Pl} - \bar{Y}_{NPl}) = (400 - 300) - (750 - 700) = 50$

	High-Poverty	Low-Poverty
Program village	400	750
Non-program village	300	700

There is a parallel "trends" assumption here too! The difference in the outcome between program and non-program villages for *low-poverty* households represents what would have existed for high-poverty households in the absence of treatment.

Tyler, Murnane, and Willett (2000)

Tyler, Murnane, and Willett (2000): what is the impact of the GED on labor market earnings for high school dropouts?

- "Treated" individuals earned the GED by passing the required exam; "untreated" individuals took the GED but did not pass the exam.
- A cross-sectional comparison of earnings would likely suffer from omitted variables bias.
- TM&W noted that the threshold passing score varied by state.

Note: GED is technically the "General Educational Development Test" but sometimes referred to as a "general equivalency diploma." In Tennessee, the HiSET is used as a high school equivalency test.

Tyler, Murnane, and Willett (2000)

Differences in the passing threshold offer a natural experiment! Consider comparing earnings of individuals with low GED scores who passed—or didn't—depending on the state they lived in.

- The "treatment" is having a low score but living in a state with a low passing threshold
- Concern: there may be systematic, baseline differences in populations and labor market outcomes across states.
- A second difference: compare earnings of *high*-scoring GED test takers who passed in both states, to "net out" state differences

Tyler, Murnane, and Willett (2000)

Cells A-D give mean income in each group:

	States where low scores do earn a GED	States where low scores do not earn a GED	Difference (states)
People with low scores	A = 9,628	B = 7,849	A-B = 1,779
People with high scores	C = 9,981	D = 9,676	C-D = 305
Difference (score groups)	A-C = -353	B-D = -1,827	(A-B)-(C-D) = 1,473

Tyler, Murnane, and Willett (2000)

- Earnings differences in row (1): the effect of the GED, if any, and any unobserved differences between states
- Earnings differences in row (2): no GED effect (all passed), only the effect of unobserved differences between states
- Under the assumption that the second differences is the gap one would observe in column (2) in the absence of treatment, we can interpret the DD as the causal effect of the GED.

Lessons from recent difference-in-differences research

See Roth et al. (2022) for a review of recent developments in difference-in-differences research designs. These new studies can be characterized by which canonical DD assumptions they relax, and how.

- Differential treatment timing: problems with TWFE when treatment effects are heterogeneous across time or units.
- Violation of parallel trends: problems with using pre-treatment trends as a test for parallel trends.
- Inference: alternatives to "sampling based" inferences that assumes a large number of clusters. (Some designs cannot justify the large cluster assumption, e.g., 50 states)

"Overall, the growing DiD econometrics literature emphasizes the importance of clarity and precision in a researcher's discussion of his or her assumptions, comparison group, and time frame selection, causal estimands, estimation methods, and robustness checks. When used in combination with context-specific information...."

Differences-in-differences with variable timing

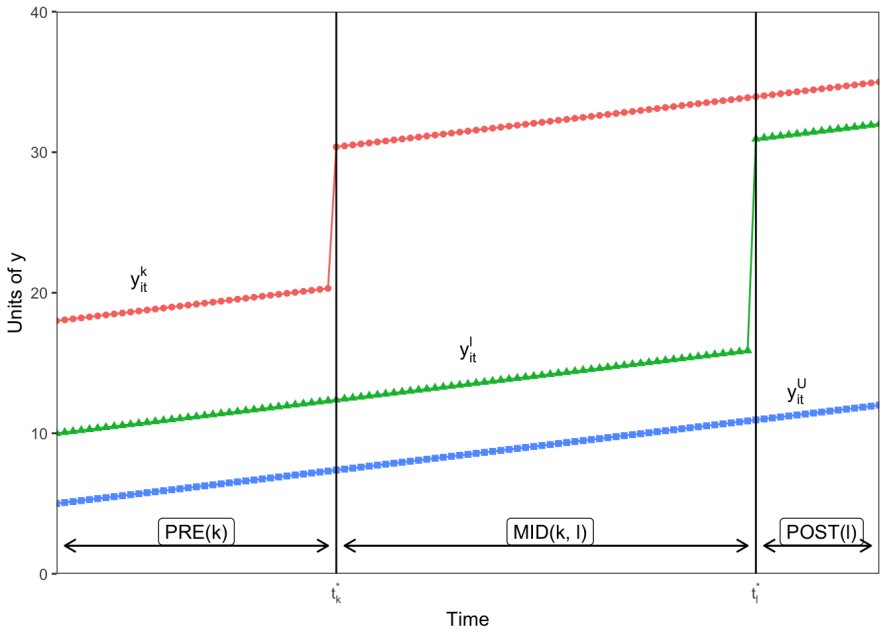
The generalized difference-in-differences model has an intuitive feel to it: changes over time for treated units are contrasted with changes over time for untreated units. Treatment may occur at different time periods, but this seems ok. The hope is that we are estimating an average treatment effect across units and time.

$$Y_{it} = \alpha_i + \lambda_t + \delta(D_i \times Post_{it}) + u_{it}$$

Recent research on two-way fixed effects models has complicated this view, and highlighted cases in which the TWFE estimator does not yield the ATT of interest.

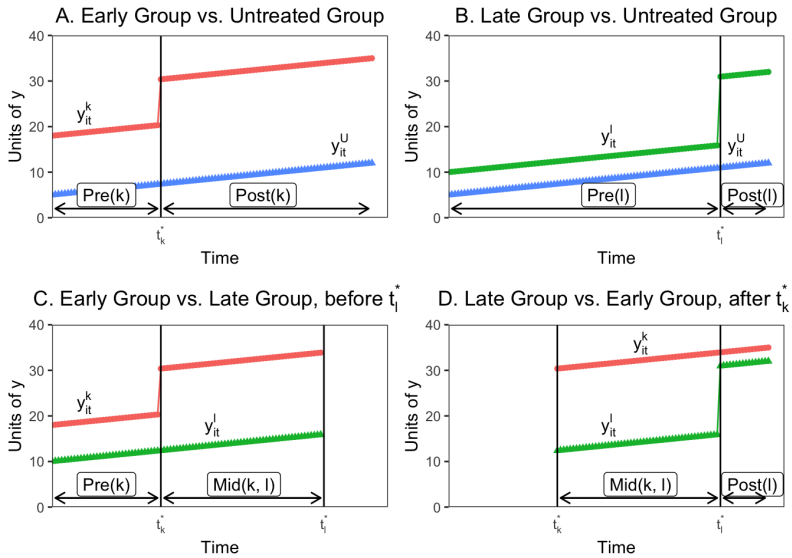
Differences-in-differences with variable timing

Goodman-Bacon (2021) points out that "early adopters" serve as a comparison group for "late adopters".



Differences-in-differences with variable timing

He also shows the generalized DD estimator is a weighted average of all possible two-group/two-period DD estimators in the data:



Weights come from group sizes *and* the share of time each group is treated.

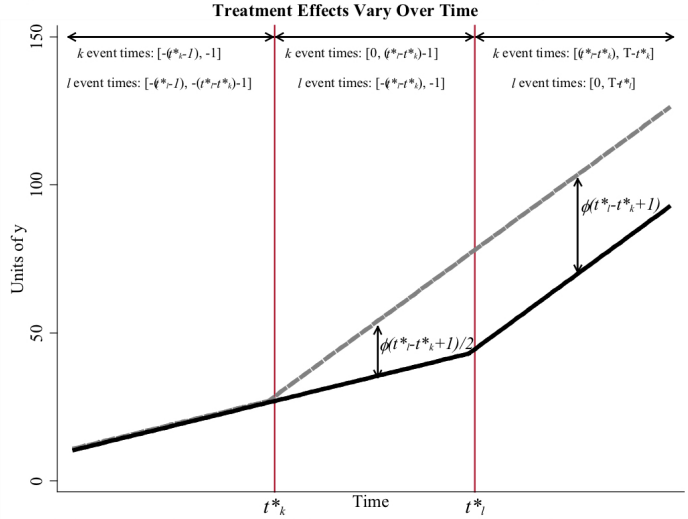
Differences-in-differences with variable timing

What are the implications of this?

- When treatment effects are homogeneous (and there are common trends), the generalized DD estimator provides the ATE. Good to go!
- However, when treatment effects are *heterogeneous*, the generalized DD estimator is a variance-weighted treatment effect that is *not* the ATE. This may not be the estimand you are interested in.
- Need to re-think the parallel trends assumption, which has to hold in all 2x2 contrasts. If a previously treated group is serving as a comparison for a later-treated group, can we assume parallel trends?
- One problematic case is when *treatment effects change over time*

Differences-in-differences with variable timing

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



Notes: The figure plots a stylized example of a timing-only DD set up with a treatment effect that is a trend-break rather than a level shift (see Meer and West 2013). Following section II.A.ii, the trend-break effect equals $\phi \cdot (t - t^* + 1)$. The top of the figure notes which event-times lie in the $PRE(k)$, $MID(k, \ell)$, and $POST(\ell)$ periods for each unit. The figure also notes the average difference between groups in each of these periods. In the $MID(k, \ell)$ period, outcomes differ by $\frac{\phi}{2}(t_\ell^* - t_k^* + 1)$ on average. In the $POST(\ell)$ period, however, outcomes had already been growing in the early group for $t_\ell^* - t_k^*$ periods, and so they differ by $\phi(t_\ell^* - t_k^* + 1)$ on average. The 2x2 DD that compares the later group to the earlier group is biased and, in the linear trend-break case, weakly negative despite a positive and growing treatment effect.

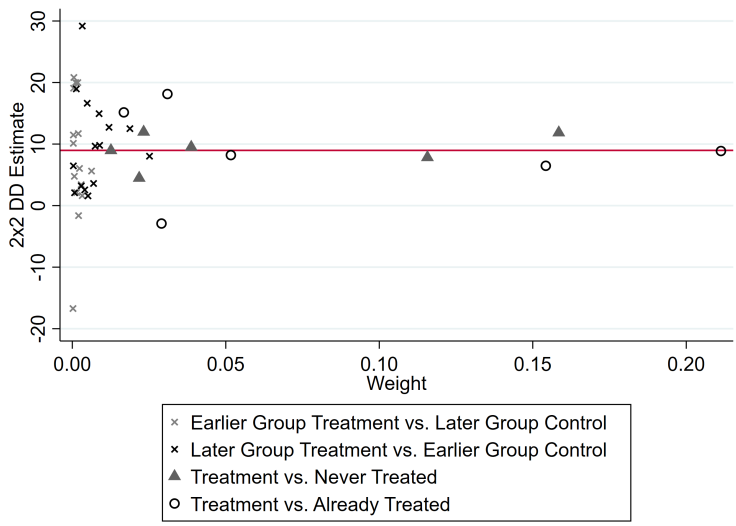
from Goodman-Bacon (2021)

Differences-in-differences with variable timing

A useful thing about the Goodman-Bacon result is that one can decompose the DD estimator into its component parts to see which "timing groups" are getting greater weight. Some treatment observations may even receive *negative* weights which can be problematic. This diagnostic can indicate whether the TWFE regression is likely to be problematic.

"Bacon decomposition"

The user-written command `bacondecomp` produces a scatterplot of the 2x2 DD estimates and their associated weights. (Must use `xtset` first). The command `ddtiming` is equivalent.



Differences-in-differences with variable timing

What to do when you have variable treatment timing: Roth et al. (2022) recommend using one of the many "heterogeneity-robust" estimators:

- Callaway & Sant'Anna (2021) - estimate treatment effects for each "treatment timing" group separately and then aggregate in a sensible way.
- de Chaisemartin and D'Haultfoeuille (2020)
- Sun and Abraham (2021)
- Cengiz, Dube, Lindner, Zipperer (2019) - stacked regression. Each treated unit is matched to 'clean' (not yet treated) controls and separate FE for each set of treated units and its control. See also Gardner (2021)
- Borusyak et al (2021) - "imputation" estimator (p 16)