4. Difference-in-Differences

LPO 8852: Regression II

Sean P. Corcoran

Difference-in-differences

Difference-in-differences is a design that—in its most common (but not only) application—contrasts *changes over time* for treated and untreated groups. DD is often used with **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 typically estimated with *panel* or *repeated cross-section* data. But they also work with other data structures.

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.

High-stakes testing in Chicago

Do test-based "high-stakes" accountability policies improve student academic performance?

- A potential "natural experiment": in Chicago, the Iowa Test of Basic Skills (ITBS) became "high stakes" for students and schools in 1997. The test was administered—but was "low stakes"—prior to that year. The test is given in grades 3, 6, and 8.
- Many other districts in Illinois also regularly administered the ITBS to these grades, but the test was low stakes.

Note: this is a simplified example inspired by Jacob (2005).

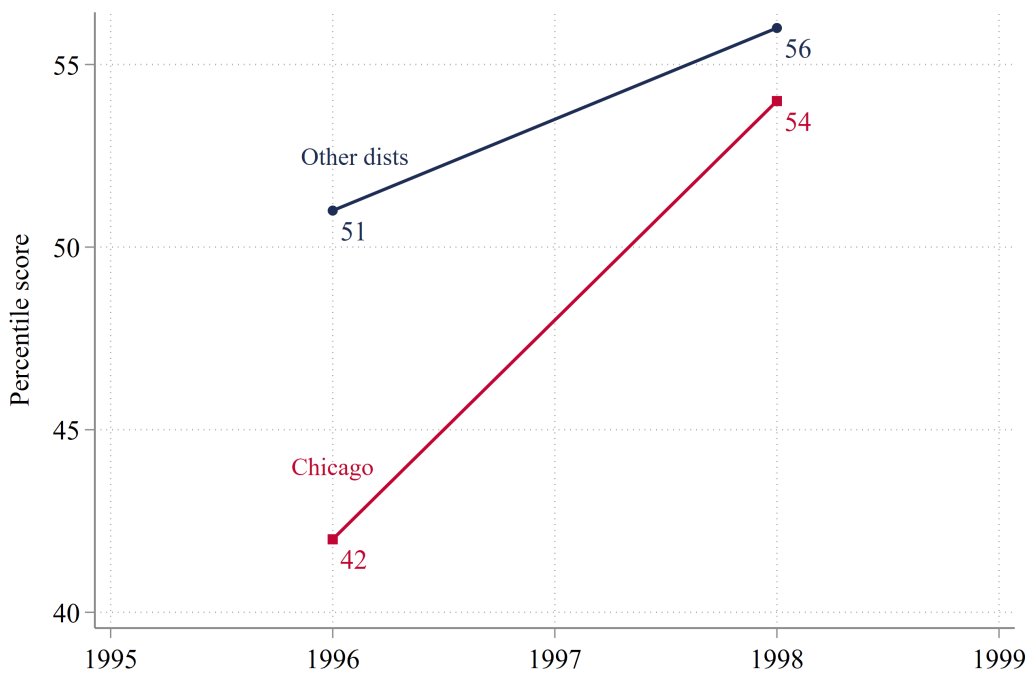
High-stakes testing in Chicago

Consider two comparisons:

- "Cross-sectional": the mean scores of Chicago 6th graders in 1998 (treated) vs. other Illinois 6th graders in 1998 (untreated).
- First difference or "interrupted time series (ITS)": the pre-to-post change in mean scores of Chicago 6th graders between 1996 and 1998.

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

High-stakes testing in Chicago



High-stakes testing in Chicago

The cross sectional comparison suggests *worse* outcomes for Chicago:

$$Y_{Chicago,1998} - Y_{Other,1998} = 54 - 56 = -2$$

The **first difference** for Chicago suggests a large *improvement*:

$$Y_{Chicago,1998} - Y_{Chicago,1996} = 54 - 42 = +12$$

Conflicting conclusions!

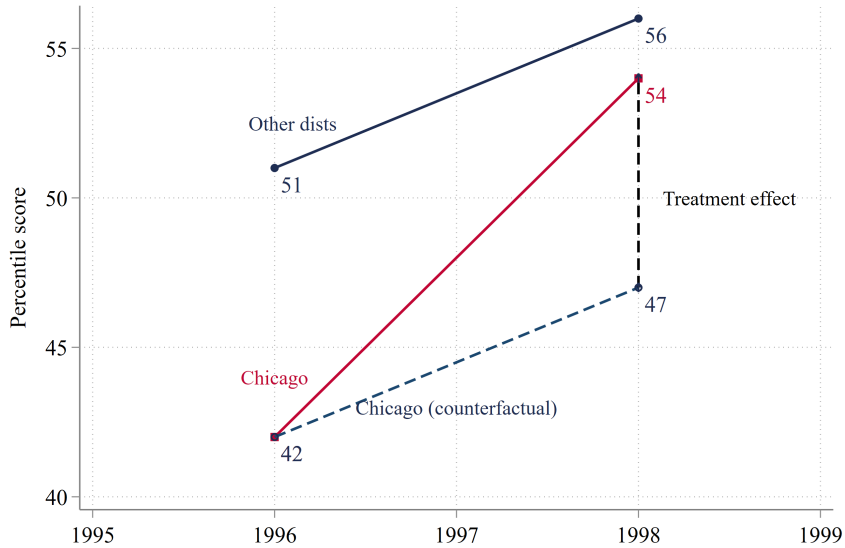
High-stakes testing in Chicago

Problems:

- The cross sectional comparison fails to recognize that Chicago 6th graders performed worse in 1996 than 6th graders in other districts did (i.e., baseline differences between treated and untreated).
- The first difference is unable to differentiate between a treatment effect for Chicago (if any) and gains between 1996 and 1998 common to all cohorts.

High-stakes testing in Chicago

Under the assumption that the change over time in other (untreated) districts represents what *would have happened* in Chicago (treated) in the absence of treatment, we can contrast *changes* in the two, or the **difference-in-differences**:



High-stakes testing in Chicago

The difference-in-differences:

$$\delta_{DD} = \underbrace{(Y_{Chicago,1998} - Y_{Chicago,1996})}_{\text{Change in Chicago}} - \underbrace{(Y_{Other,1998} - Y_{Other,1996})}_{\text{Change in other districts}}$$

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

The differencing of the two "first differences" represents the **second difference**. There was a "counterfactual" gain of 5 implied by the other districts.

High-stakes testing in Chicago

An equivalent way to write δ_{DD}:

$$\delta_{DD} = \underbrace{(Y_{Chicago,1998} - Y_{Other,1998})}_{\text{Difference "post"}} - \underbrace{(Y_{Chicago,1996} - Y_{Other,1996})}_{\text{Difference "pre"}}$$

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 Chicago and other districts for 1996 and 1998).

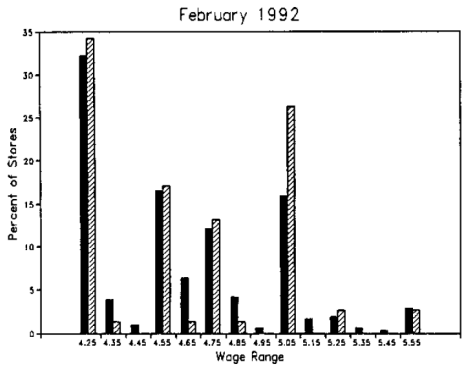
Card & Krueger (1994)

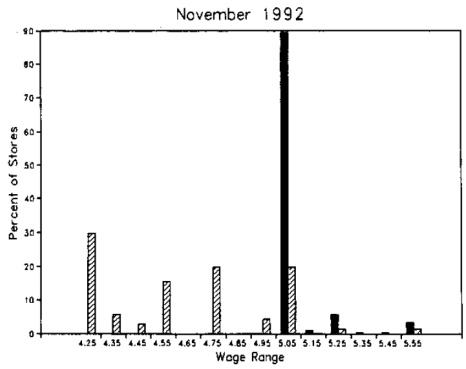
A classic DD study of the impact of the minimum wage on fast food employment (an industry likely to be affected by the minimum wage).

- 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 increase.

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 form of difference-in-differences:

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

The DD design can accommodate much more complicated setups.

Causal interpretation of difference-in-differences

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

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

Causal interpretation of difference-in-differences

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).

Causal interpretation of difference-in-differences

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

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 (constant treatment effect)

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

Causal interpretation of difference-in-differences

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 $t = 1$.

Causal interpretation of difference-in-differences

The (causal) ATT 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* (the $D_i = 1$ group).

Causal interpretation of difference-in-differences

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. Selection bias would be 0. 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.

Causal interpretation of difference-in-differences

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.

Causal interpretation of difference-in-differences

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!

Causal interpretation of difference-in-differences

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.

Causal interpretation of difference-in-differences

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:

Causal interpretation of difference-in-differences

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

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

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.

Parallel trends assumption

The parallel trends assumption means the pre-to-post change in Y(0) for the $D = 0$ group represents what *would have happened* to the $D = 1$ group had they not been treated.

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

Difference-in-differences: summary thus far

To summarize:

- Changes over time in the $D = 0$ group provide the counterfactual.
- Selection into treatment related to fixed (time invariant) unobserved differences is OK.
- The outcome *levels* are not important, only the within-group *differences*.
- DD can provide a consistent estimate of the ATT if the parallel trends assumption holds.

Difference-in-differences: summary thus far

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

- 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 (i) and two groups ($D_i = 0$ or $D_i = 1$) observed in "pre" and "post" periods, we can use regression to estimate δ_{DD}:

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

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

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

Regression difference-in-differences (2x2)

How does this regression map onto our earlier notation?

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

There are four expectations estimated in this regression:

$$E[Y_{it}|D_i = 0, POST_t = 0] = \alpha$$
$$E[Y_{it}|D_i = 1, POST_t = 0] = \alpha + \beta$$
$$E[Y_{it}|D_i = 0, POST_t = 1] = \alpha + \lambda$$
$$E[Y_{it}|D_i = 1, POST_t = 1] = \alpha + \beta + \lambda + \delta$$

Regression difference-in-differences (2x2)

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

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

δ is the *differential* change over time for the $D_i = 1$ group **(DD)**

Regression difference-in-differences (2x2)

More succinctly:

	Pre ($POST = 0$)	Post ($POST = 1$)	Diff 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. Note OLS will always (mechanically) estimate δ as the differential change between the $D_i = 1$ and $D_i = 0$ groups. Whether that δ can be interpreted as the (causal) ATT depends on the parallel trends assumption.

Regression difference-in-differences (2x2)

Here: some NYC schools adopted a breakfast in the classroom program in 2010. What was the impact of this program on average daily participation in breakfast?

```
. reg bkfast_part i.everbic##i.post
```

Source	SS	df	MS		
Model	6.66627777	3	2.22209259	Number of obs	= 6,160
Residual	111.439598	6,156	.018102599	F(3, 6156)	= 122.75
				Prob > F	= 0.0000
				R-squared	= 0.0564
				Adj R-squared	= 0.0560
Total	118.105875	6,159	.019176145	Root MSE	= .13455

bkfast_part	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.everbic	.0364431	.011215	3.25	0.001	.0144578	.0584285
1.post	.0004512	.0035743	0.13	0.900	-.0065557	.0074581
everbic#post						
1 1	.2219777	.0177852	12.48	0.000	.1871125	.256843
_cons	.2494476	.0022566	110.54	0.000	.2450239	.2538713

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). Alternatively, use two-way fixed effects (later section).

Regression difference-in-differences (2x2)

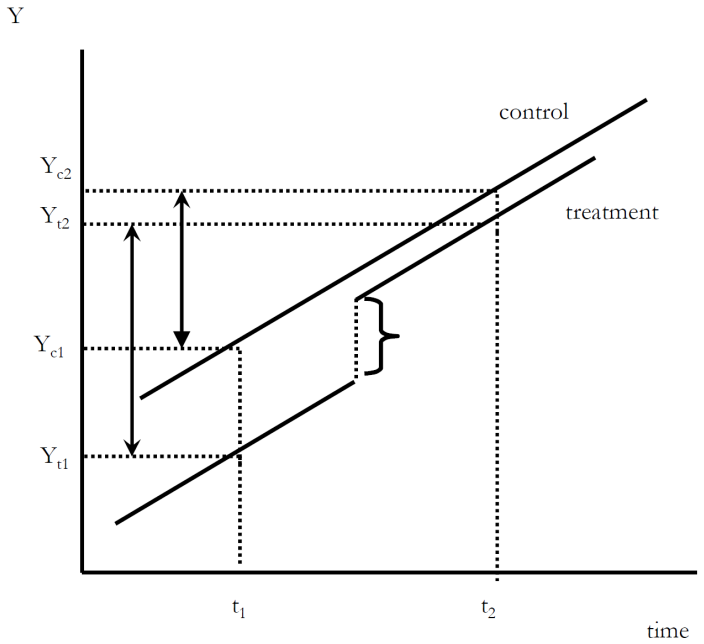
The 2x2 regression difference-in-differences can be extended to include covariates:

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

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

Parallel trends assumption

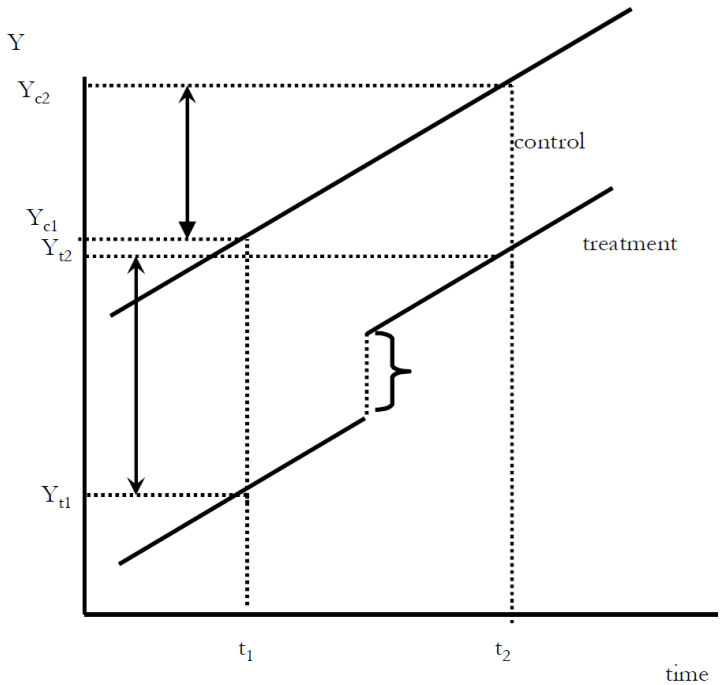
The key assumption in DD is parallel trends: that the time trend in the absence of treatment would be the same in both groups.



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

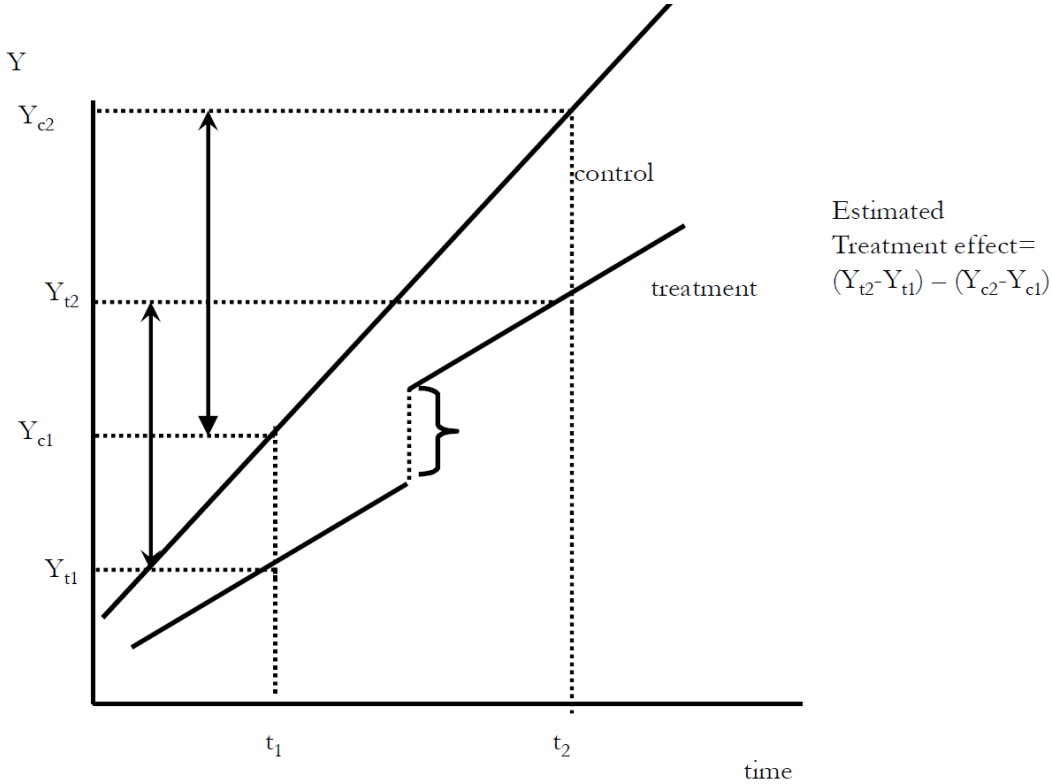
Parallel trends assumption

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



Common violations of parallel trends assumption

Common scenarios that would violate the parallel trends assumption:

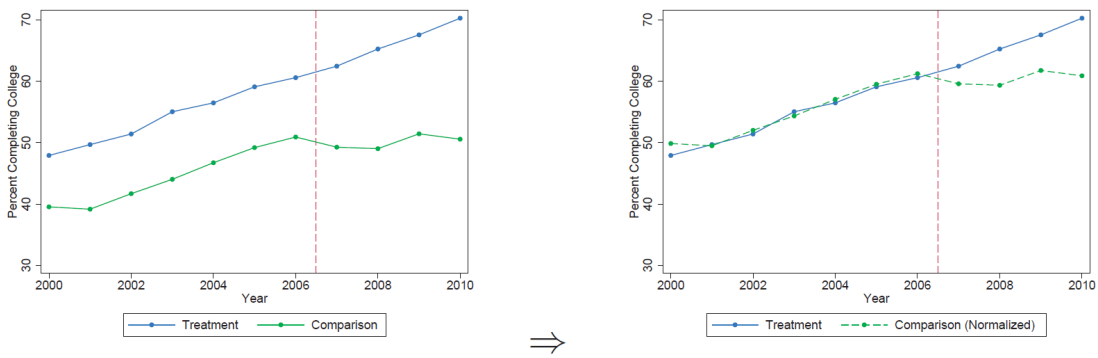
- **Targeted treatments**: often programs are targeted at subjects who are most likely to benefit from it. In many cases, the fact that a subject was on a different trajectory is what made them a good candidate for the program (e.g., a struggling student).
- **Ashenfelter's dip**: treated cases may experience 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.

Parallel trends assumption

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

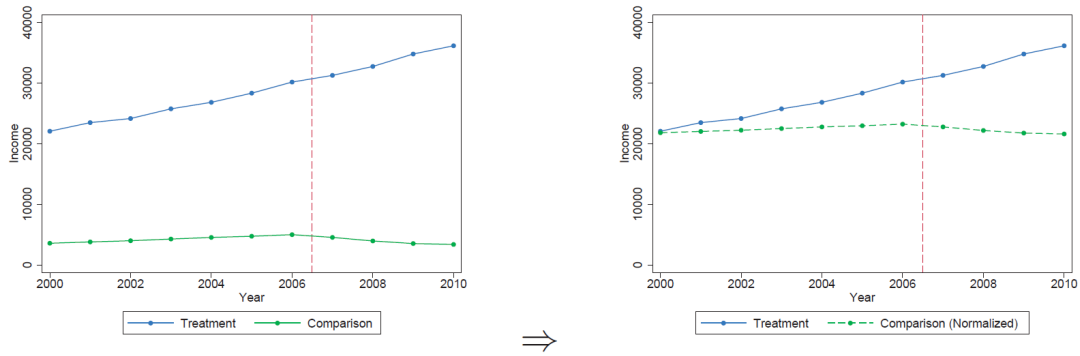
- A compelling graph: point to similar trends *prior to* treatment. Note: parallel trends *prior* to treatment are neither necessary nor sufficient for the parallel trends assumption, which is about the *post* period!
- Event study regression and graph
- Statistical tests for differences in pre-treatment trends
- A placebo / falsification test
- Controlling for time trends directly (leans heavily on functional form)
- Triple-differences model
- Probably most important: understanding the context of your study! Ruling out reasons for non parallel-trends.

Graphical assessment of parallel trends assumption



The graph on the right ("normalized") subtracts baseline difference between a treated and untreated comparison group to better visualize the trend differences. In Stata 17+ see `didregress` post-estimation command `estat trendplots`. Also easy to do yourself.

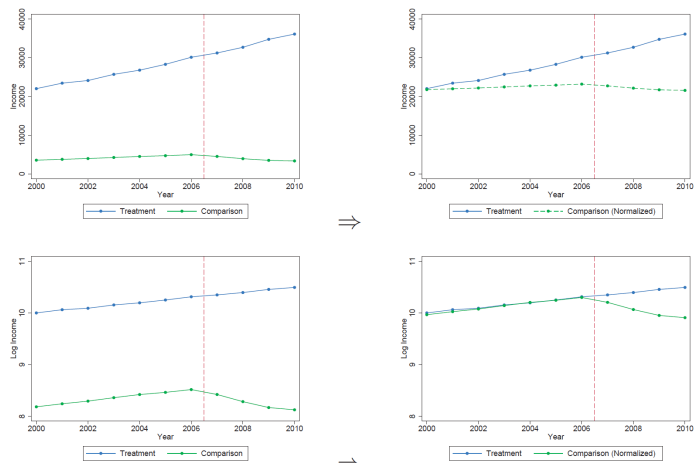
Graphical assessment of parallel trends assumption



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

Graphical assessment of parallel trends assumption

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



If your outcome variable is in levels and does not satisfy parallel trends, a log transformation may help (if appropriate for your outcome). Above, the top panels are in levels; the bottom panels are in logs.

Covariates and the 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.

Statistical tests for differences in pre-treatment trends

Two tests (easily implemented in Stata 17+ `did` commands, though not hard to code):

1. **Differential linear trend for the treated**: add to the DD model separate linear time trends for the ever-treated group, pre- and post-treatment. Conduct an F-test for significance of the pre-treatment linear trend. This assesses whether the treated group was on a differential trend prior to treatment. See `estat ptrends`.
2. **Granger-type test**: add to the DD model a full set of interactions between pre-treatment years and ever-treated. Conduct an F-test for the joint significance of these interactions. This assesses "anticipatory" effects. See `estat granger`.

In practice, event studies are more common than these tests.

Introduction to event studies

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 (*leads*) and K periods after treatment (*lags*). 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 a generally accepted practice 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 $t=0$)

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.

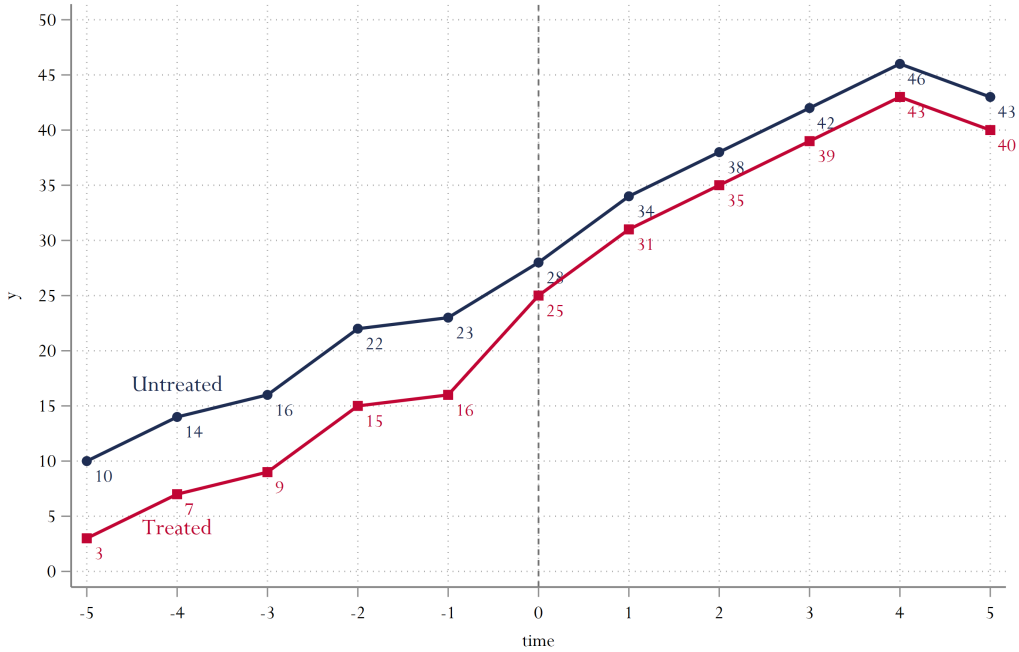
If the coefficients on the pre-treatment (lead) interaction terms are significantly different from zero, this suggests a non-parallel trend before treatment.

The **event study graph** is a plot of these γ_j, 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 observed Y in each time period for two groups:



Event study

In this stylized example, **pre**-treatment:

$$\lambda_{-2} = 22 - 23 = -1$$
$$\lambda_{-3} = 16 - 23 = -7$$
$$\lambda_{-4} = 14 - 23 = -9 \ \ldots \text{ etc.}$$
$$\gamma_{-2} = (15 - 16) - (22 - 23) = \mathbf{0}$$
$$\gamma_{-3} = (9 - 16) - (16 - 23) = \mathbf{0}$$
$$\gamma_{-4} = (7 - 16) - (14 - 23) = \mathbf{0} \ \ldots \text{ etc.}$$

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.

$\gamma_0 = (25 - 16) - (28 - 23) = \mathbf{4}$

$\gamma_1 = (31 - 16) - (34 - 23) = \mathbf{4}$

$\gamma_2 = (35 - 16) - (38 - 23) = \mathbf{4}$ .. etc.

The treatment effect of 4 appears in time 0 and remains constant in subsequent periods.

Event study

If there are two groups and treatment occurs for everyone in the $D_i = 1$ group in the same time period, this is very easy to implement in Stata (assume *year* is the time period):

```
reg y i.evertreated##i.year
```

This is a full factorial of treatment group status and time period—includes main effects for *evertreated* and each year, 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.evertreated##ib2005.year
```

Event study example

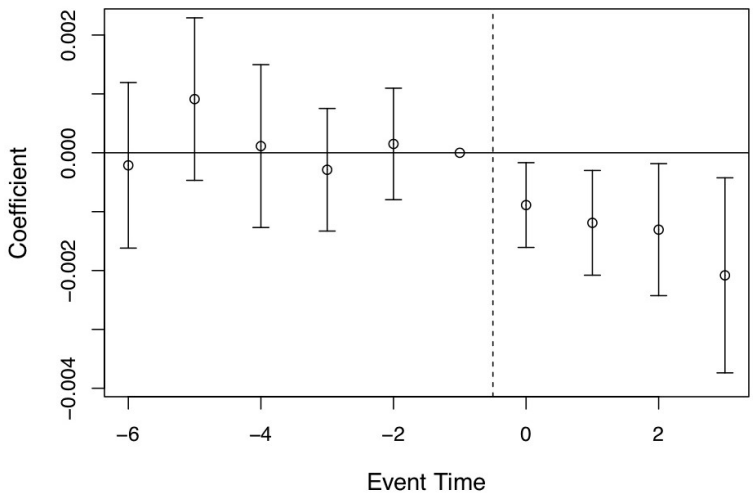
The following figures are from Miller et al. (QJE 2021), via the *Mixtape*. The authors estimate the impact of state expansion of Medicaid under ACA on the annual mortality rates of older persons under 65 in the U.S.

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

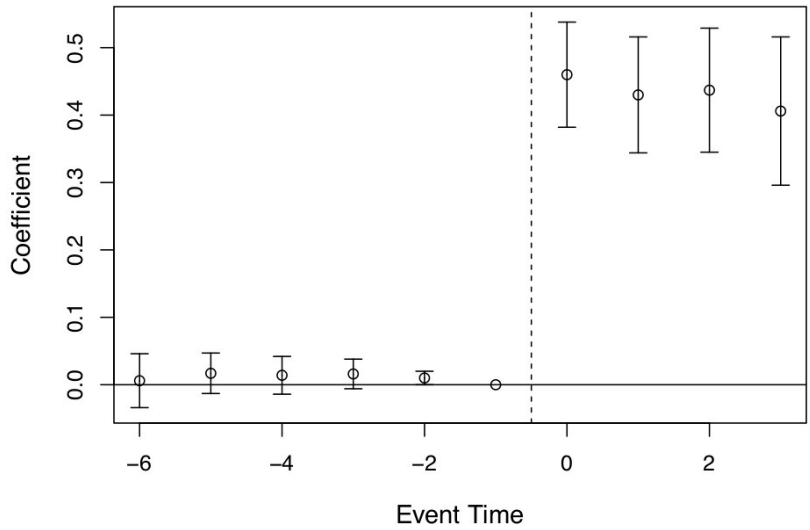
Plotted points are event study coefficients, shown with 95% confidence intervals. (Time zero is the first year of expansion). Outcome: mortality rate



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

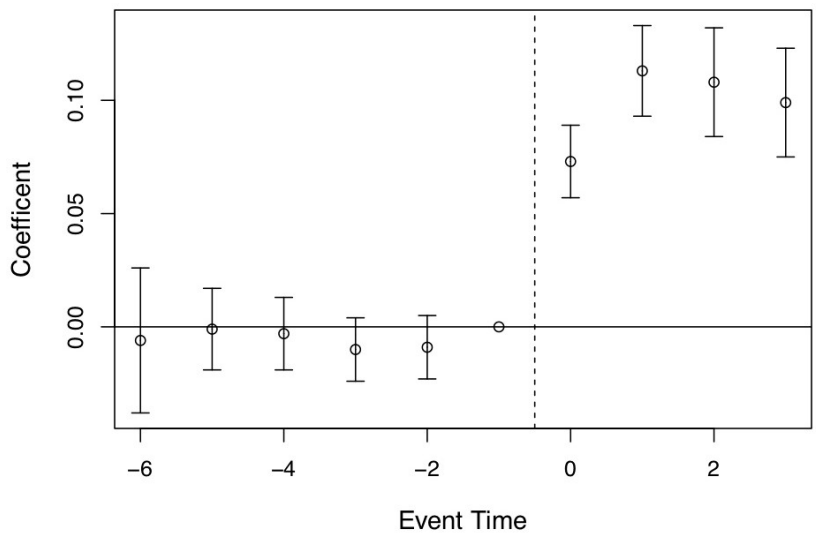
Event study example

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**



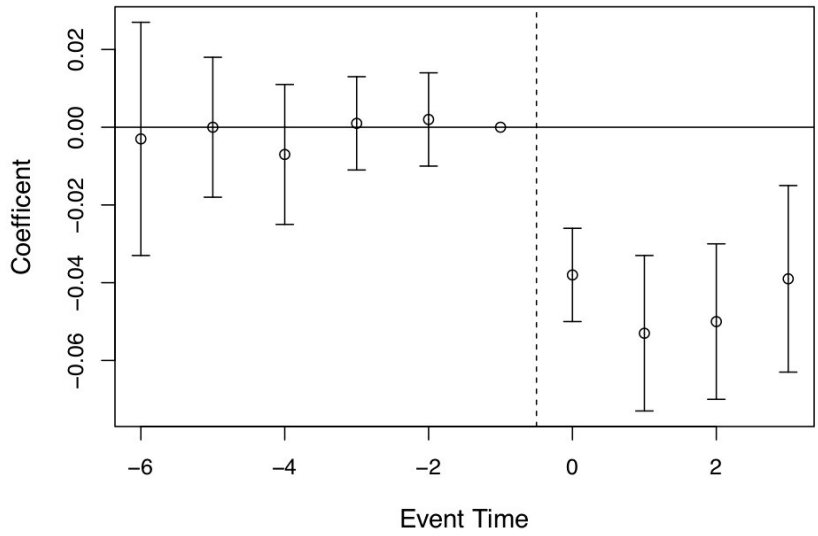
Event study example

Here: Medicaid **coverage rates**



Event study example

Here: **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 using `eventdd`

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.group`*`, timevar(`*`timetoevent`*`)`

This syntax estimates an OLS model with the *group* main effect included in the covariates (see also next slide). 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 lead, and so on. This variable should be **missing for groups that are never treated**. See Clarke and Schythe (2020).

Event study using eventdd

With lots of groups (or panel data) you can have `eventdd` estimate a fixed effects model specification:

`eventdd` *`y x1 x2`*`, timevar(`*`timetoevent`*`) method(fe, absorb(state))`

Here the command uses `xtreg` where the variable *state* is used as the fixed effect.

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 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, 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.

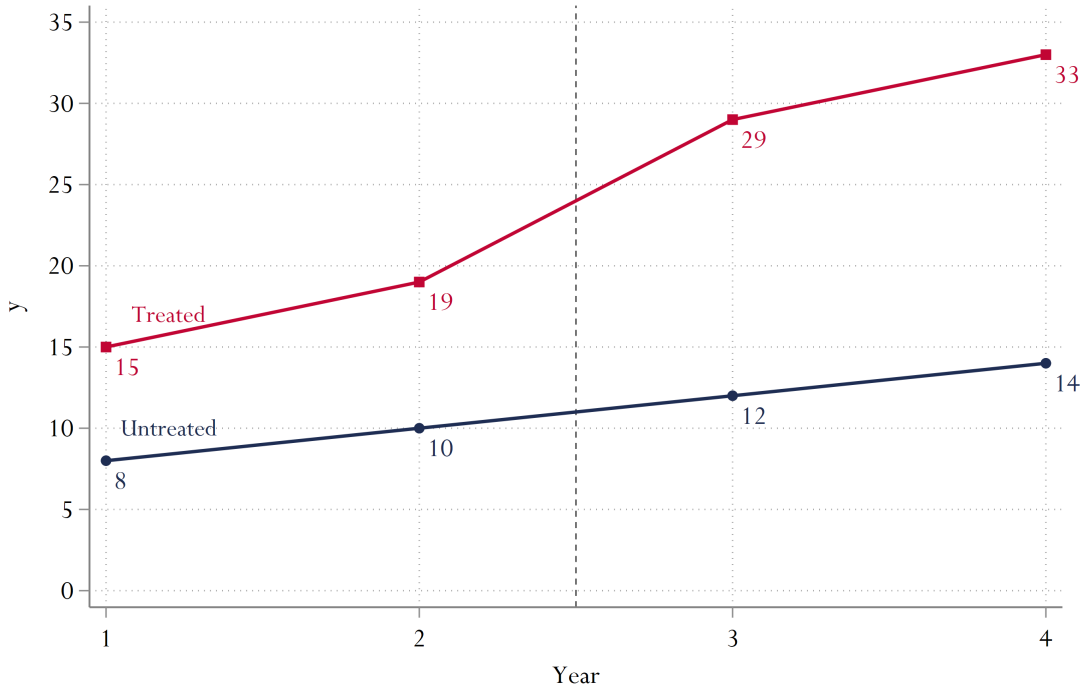
Triple difference

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.

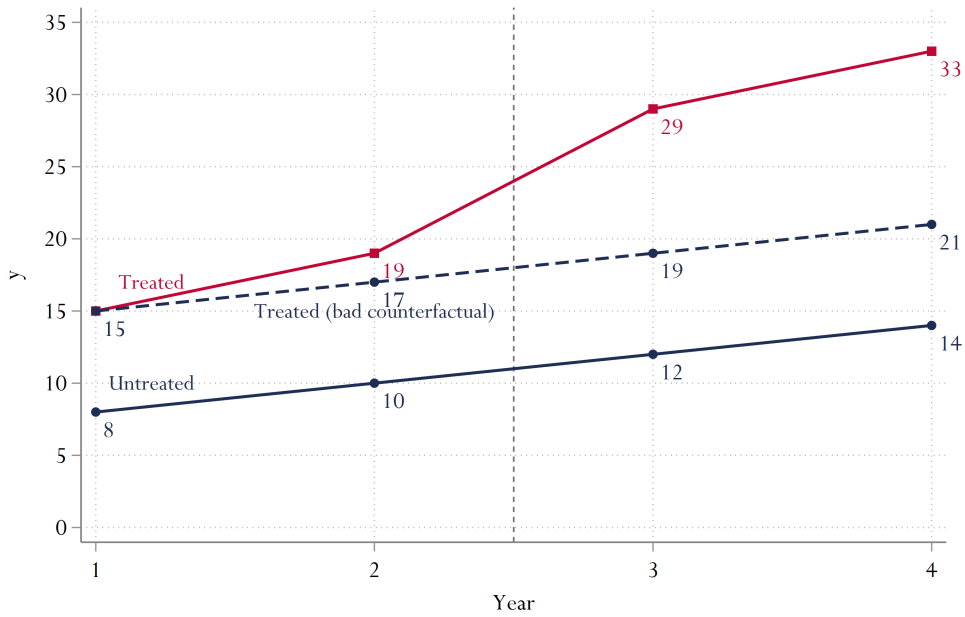
Triple difference

Stylized example: non-parallel trends



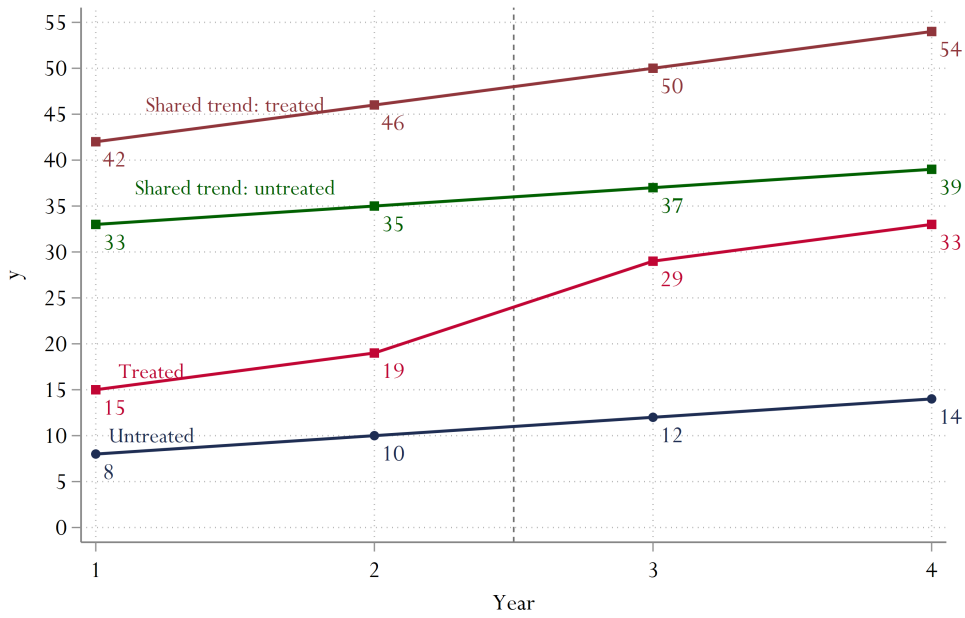
Triple difference

The untreated group is a bad counterfactual for the DD. Using means in the pre and post periods, the DD estimate $(31 - 17) - (13 - 9) = 10$ overstates the treatment effect due to the differential time trend.



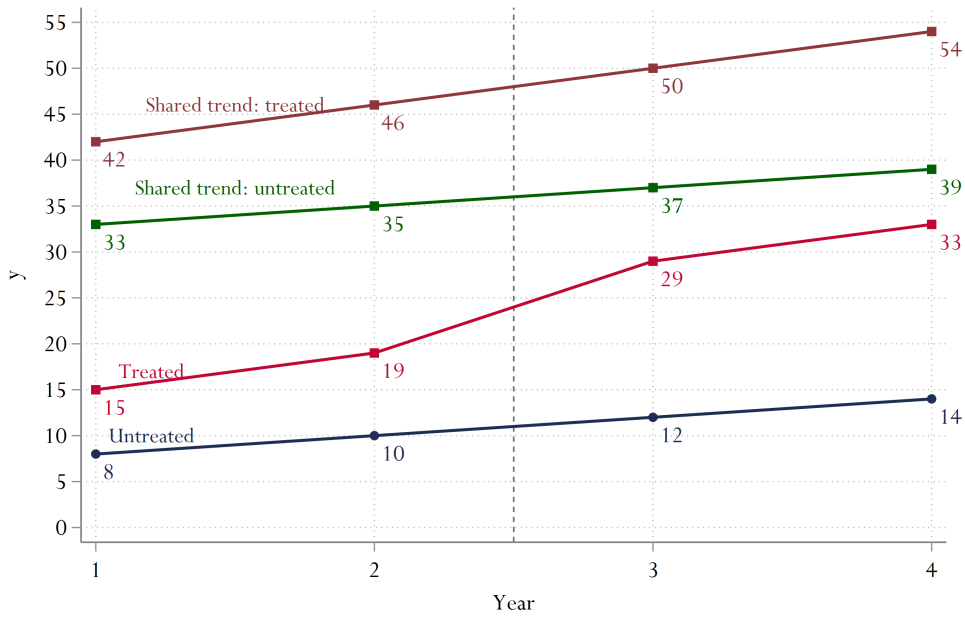
Triple difference

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 using means pre and post): $(52 - 44) - (38 - 34) = 4$. Subtract from the original DD estimate to isolate the treatment effect: $10 - 4 = 6$



Regression triple difference

We can use regression to estimate the triple difference:

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

- $POST_t = 1$ in the post period (still assuming 2 periods)
- $D_i = 1$ for the "ever treated" units within the focal and additional comparison groups
- $G_i = 1$ for the focal treated and untreated group ("primary"), while $G_i = 0$ for the additional comparison groups ("secondary")

There are 3 indicator variables, three 2-way interactions, and one 3-way interaction.

Regression triple difference

In Stata, for the stylized example above:

```
reg y i.evertreatgroup i.primary i.primary#i.evertreatgroup
i.post i.evertreatgroup#i.post i.primary#i.post
i.evertreatgroup#i.primary#i.post
```

Or:

```
reg y i.evertreatgroup##i.primary##i.post
```

See the .do file on Github that walks you through this stylized example.

Regression triple difference

```
.    reg y i.evertreatgroup i.primary i.evertreatgroup#i.primary i.post ///
>    i.evertreatgroup#i.post i.primary#i.post i.evertreatgroup#i.primary#i.post
```

Source	SS	df	MS	Number of obs	=	16
				F(7, 8)	=	94.83
Model	3319	7	474.142857	Prob > F	=	0.0000
Residual	40	8	5	R-squared	=	0.9881
				Adj R-squared	=	0.9777
Total	3359	15	223.933333	Root MSE	=	2.2361

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.evertreatgroup	10	2.236068	4.47	0.002	4.843618	15.15638
1.primary	-25	2.236068	-11.18	0.000	-30.15638	-19.84362
evertreatgroup# primary 1 1	-2	3.162278	-0.63	0.545	-9.292225	5.292225
1.post	4	2.236068	1.79	0.111	-1.156382	9.156382
evertreatgroup# post 1 1	4	3.162278	1.26	0.242	-3.292225	11.29223
primary#post 1 1	0	3.162278	0.00	1.000	-7.292225	7.292225
evertreatgroup# primary#post 1 1 1	6	4.472136	1.34	0.217	-4.312764	16.31276
_cons	34	1.581139	21.50	0.000	30.35389	37.64611

Regression triple difference

Interpretation:

- α (34): mean outcome in the "pre" period for the never treated (secondary)
- β_1 (10): difference between the ever treated (secondary) and never treated (secondary) in the pre period
- β_2 (-25): difference between the never treated (primary) and never treated (secondary) in the pre period
- β_3 (-2): if -25 was the difference in the pre period between the never treated (primary) and never treated (secondary), this is how *different* the difference is in the pre period between the ever treated (primary) and ever treated (secondary)

Regression triple difference

Interpretation, cont.

- β_4 (4): the change from pre to post for the never treated (secondary)
- β_5 (4): the *differential* change from pre to post for the ever treated, secondary (vis a vis the never treated, secondary). Think of this as the diff-in-diff for the secondary group.
- β_6 (0): *differential* change from pre to post for the never treated (primary) and never treated (secondary). This is zero since these groups have the same time trend (by design for this example).
- β_7 (6): the **triple difference**. If *post*evertreatgroup* is the diff-in-diff for the secondary group, then this is how *different* the diff-in-diff is for the primary group.

Triple difference

Two example studies:

- 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 also Olden & Møen (2022) for more on the triple difference method.

Generalized difference-in-differences

Most examples thus far have assumed a common "post" 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-differences 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

In *Mastering 'Metrics* chapter 5 what is the effect of a lower Minimum Legal Drinking Age (MLDA) on traffic fatalities among young adults?

- 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*) and DD to address this question.

Note: this example is based on Carpenter & Dobkin (2011).

Generalized difference-in-differences

Their TWFE (generalized DD model) for mortality rates by motor vehicle accidents (Y_{st}) in state s in year t:

$$Y_{st} = \beta_s + \gamma_t + \delta(D_s \times POST_{st}) + u_{st}$$

which is shorthand for:

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

Generalized difference-in-differences

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

- $STATE_{ks} = 1$ 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_j is a *year effect*.
- Covariates X_{st} may be included to control for other time-varying factors associated with Y and treatment.

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.
- The year effects γ_j capture trends in the outcome common to all states, unrelated to treatment.
- There need not be a common post-treatment period. $POST_{st} = 1$ for years in which a state is treated (with timing specific to their case).
- The coefficient on the interaction (δ) represents how much, on average, outcomes *differ* in treated 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.

Generalized difference-in-differences

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

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

Alternatively, can use user-written `reghdfe` which accommodates multiple fixed effects:

```
reghdfe y i.evertreated#i.post, absorb(state year)
```

In Stata 17+ can use `did` commands. For example, with panel data for states, where *treated* is a time-varying treatment variable (like the interaction of *evertreated* and *post*):

```
xtdidregress (y x) (treated), group(state) time(year)
```

Generalized difference-in-differences

Using our notation for *potential outcomes* for a state k in year t:

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

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.

Treatment as "intensity" or "dosage"

Treatment need not be binary in the generalized DD. Rather, could operationalize treatment as a continuous "intensity" or "dosage" measure. In the MLDA example:

$$Y_{st} = \alpha + \delta Intensity_{st} + \sum_{k=2}^{50} \beta_k STATE_{ks} + \sum_{j=2}^{T} \gamma_j YEAR_{jt} + u_{st}$$

Here, $Intensity_{st}$ is a time-varying measure of the dose of treatment in state s in year t. (It is equal to zero if untreated). Another 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 (and more) from the MLDA study reported in *Mastering 'Metrics*.

- Generalized DD using two-way fixed effects
- Placebo test using other outcomes, age groups
- Triple difference
- Event study
- Group-specific time trends

Note the treatment here is an "intensity" measure ranging from 0 to 1: what proportion of young adults aged 18-20 could legally drink in state s in year t? The measure also accounts for partial-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):

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

Group-specific time trends

How does this work? Consider again the evolution of potential outcomes in two conditions (untreated and treated):

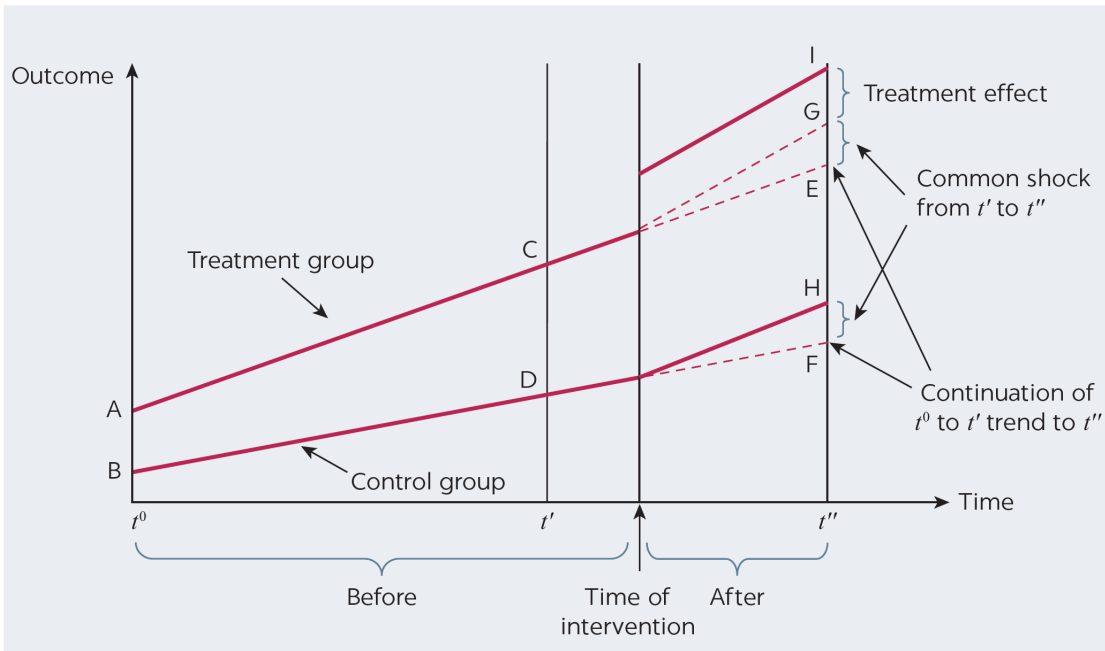
$$Y_{0kt} = \alpha + \beta_k + \gamma_t + \theta_k t$$
$$Y_{1kt} = \alpha + \beta_k + \gamma_t + \theta_k t + \delta$$

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 *differential* change over time associated with treatment.

Group-specific time trends

FIGURE 12.1 Three periods, with nonparallel trends



Source: Original figure for this publication.

Source: Glewwe & Todd (2022) chapter 12.

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 lean heavily on the linearity assumption and are generally *less precise*. This specification in practice is more often used as a robustness check than a primary model specification.

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 "treated" 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.

Difference-in-differences in other contexts

What would the regression model be?

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

- $D_i = 1$ for villages in the program ($= 0$ for non-program villages)
- $HP_p = 1$ for high-poverty households ($= 0$ for low-poverty households)

Note we are back in the simple 2x2 DD framework.

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 with different thresholds.
- 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.

Tyler, Murnane, and Willett (2000)

What would the regression model be?

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

- $D_i = 1$ if individual i lives in a state with a low GED passing threshold
- $LS_s = 1$ if individual i had a low score but above the passing threshold for their state

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...."

Difference-in-differences with variable timing

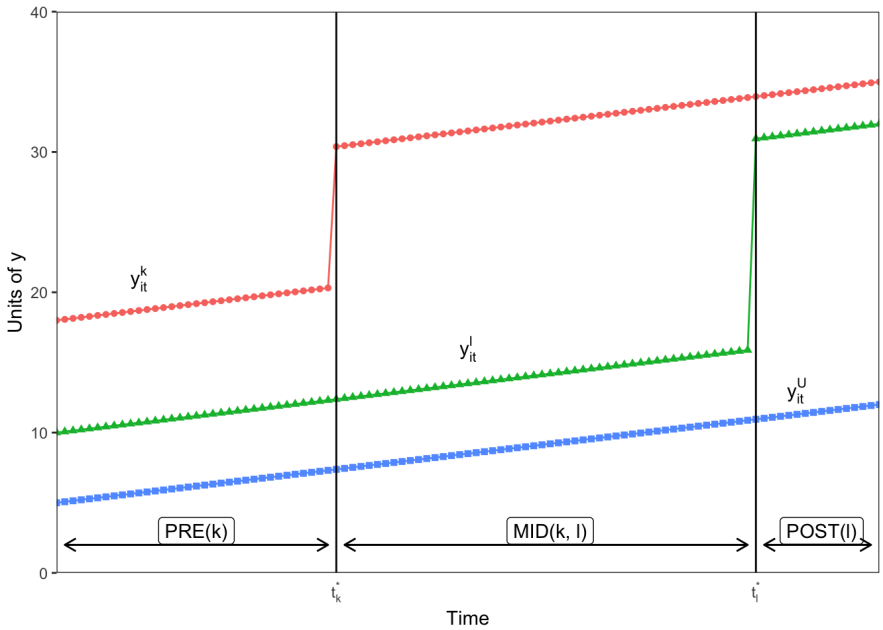
The generalized difference-in-differences model (TWFE) has an intuitive feel to it: changes over time for treated units are contrasted with changes over time for untreated units. Treatment may occur at different time periods, but this seems ok. The hope is that we are estimating an ATT across units and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

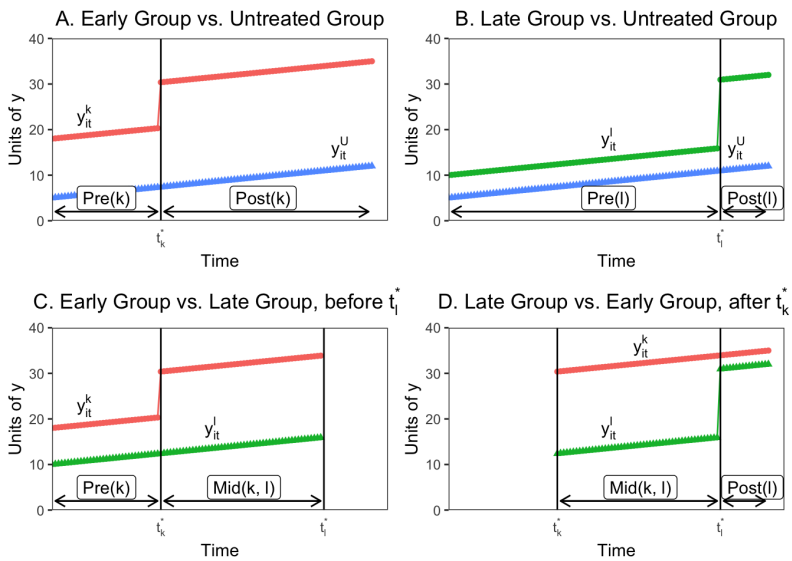
Difference-in-differences with variable timing

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



Difference-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.

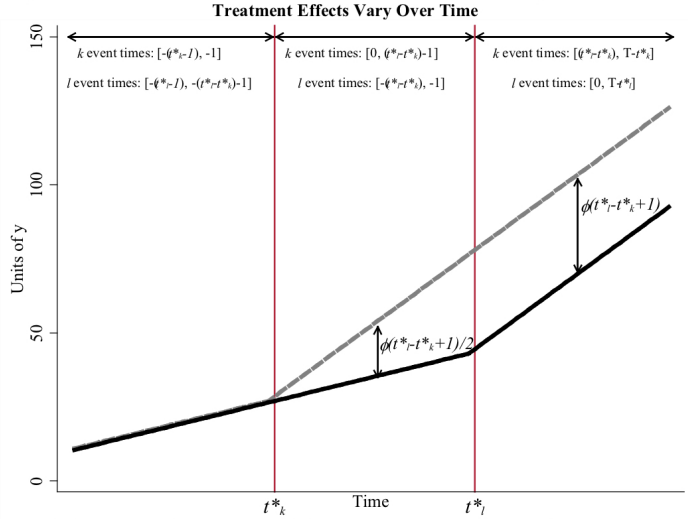
Difference-in-differences with variable timing

What are the implications of this?

- When treatment effects are homogeneous (and there are common trends), the generalized DD estimator provides the ATT. Good to go!
- However, when treatment effects are *heterogeneous*, the generalized DD estimator is a variance-weighted average treatment effect that is *not* the ATT and probably not the estimand you are interested in.
  - Treatment effects could be heterogeneous in *time since treatment*
  - Treatment effects could be heterogeneous *across units*
- Need to re-think the parallel trends assumption, which has to hold in <u>all</u> 2x2 contrasts. If a previously treated group is serving as a comparison for a later-treated group, can we assume parallel trends?

Difference-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.

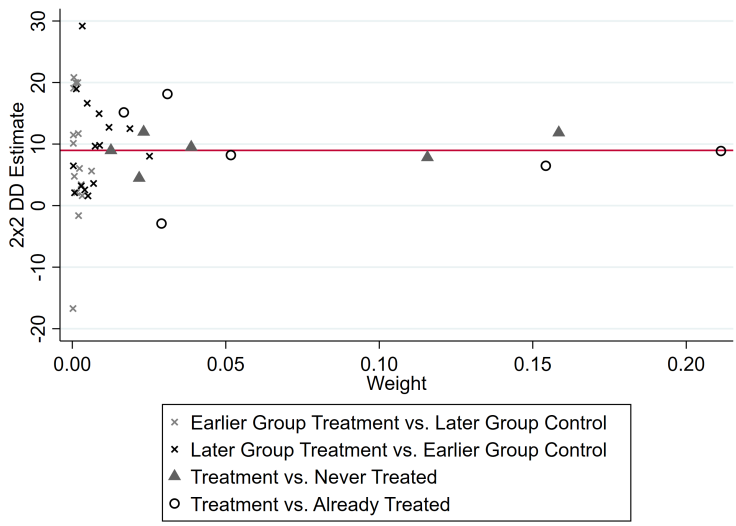
from Goodman-Bacon (2021)

Difference-in-differences with variable timing

A useful result from Goodman-Bacon 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.



Difference-in-differences with variable timing

The TWFE estimator is problematic in part because it is an average of many pre-post comparisons. If treatment effects are heterogeneous (either across time or across units), it can be hard to interpret.

What about event studies, which parse out separate treatment effects by year? This provides sensible estimates if the only source of heterogeneity is *across time* (and units are observed with differential time since treatment). If heterogeneity is *across treatment cohorts*, coefficients are hard to interpret (see Sun & Abraham 2021).

Difference-in-differences with variable timing

What to do? Roth et al. (2022) recommend using one of the many "heterogeneity-robust" estimators:

- **Callaway & Sant'Anna (2021)** - estimate ATTs 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

These approaches share an interest in making "clean" comparisons and avoiding "forbidden" ones.

Callaway and Sant'Anna (2021)

Estimate all possible clean ATTs and then aggregate them. In Stata: `csdid` (also install `drdid`). Define g as the timing group identifier (first year of treatment), which is equal to 0 if never treated.

`csdid` *y x*`, ivar(`*varname*`) time(`*year*`) gvar(`*g*`) method(drdid estimator) [notyet]`

For $t > g$ ATT is estimated as (NT=never treated):

$$[E(Y_{g,t}) - E(Y_{NT,t})] - [E(Y_{g,g-1}) - E(Y_{NT,g-1})]$$

For $t < g$ ATT is estimated as (NT=never treated):

$$[E(Y_{g,t}) - E(Y_{NT,t})] - [E(Y_{g,t-1}) - E(Y_{NT,t-1})]$$

Note the latter are period to period comparisons (pre-treatment)

Callaway and Sant'Anna (2021)

`csdid` reports all of the $ATT(g,t)$ estimates. With a small number of timing groups and periods, could report these individual estimates.

It is more likely you will want to aggregate these to one ATT estimate. The post-estimation command `estat` can produce these.

- `estat event`: event study estimates (period by period)
- `estat simple`: simple weighted average of ATT estimates (weighting by group size)
- `estat group`: aggregation by group

Callaway and Sant'Anna (2021)

Aggregation is more complicated when there are covariates and the parallel trends assumption holds only conditional on covariates. Here Callaway and Sant'Anna provide several estimators: outcome regression (OR), inverse probability weighting (IPW), or doubly robust (DR/AIPW).

These rely on propensity scores for treatment in period g conditional on X.

Callaway and Sant'Anna (2021)

Great resources on `csdid`:

- Stata conference presentation: `http://fmwww.bc.edu/RePEc/scon2021/US21_SantAnna.pdf`
- Programmer website: `https://friosavila.github.io/playingwithstata/main_csdid.html`

See also Sun and Abraham (2020) and `eventstudyinteract` Stata implementation for another estimator that weights timing group ATTs.

Stacked event study

Also in the aim of creating "clean" comparisons, Cengiz et al (2019) introduced the idea of a "stacked" regression.

- Each treatment timing group is matched to "clean" (never treated) controls in its own dataset.
- These datasets are stacked.
- Regression is estimated that includes unit by stack fixed effects, time by stack fixed effects, so comparisons are "within-stack"

Implement in Stata using `stackedev` (by Josh Bleiberg).