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 can 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., Medicaid expansion, EZ Pass)
- Targeted investments (e.g., school construction, ed finance reform)
- 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).

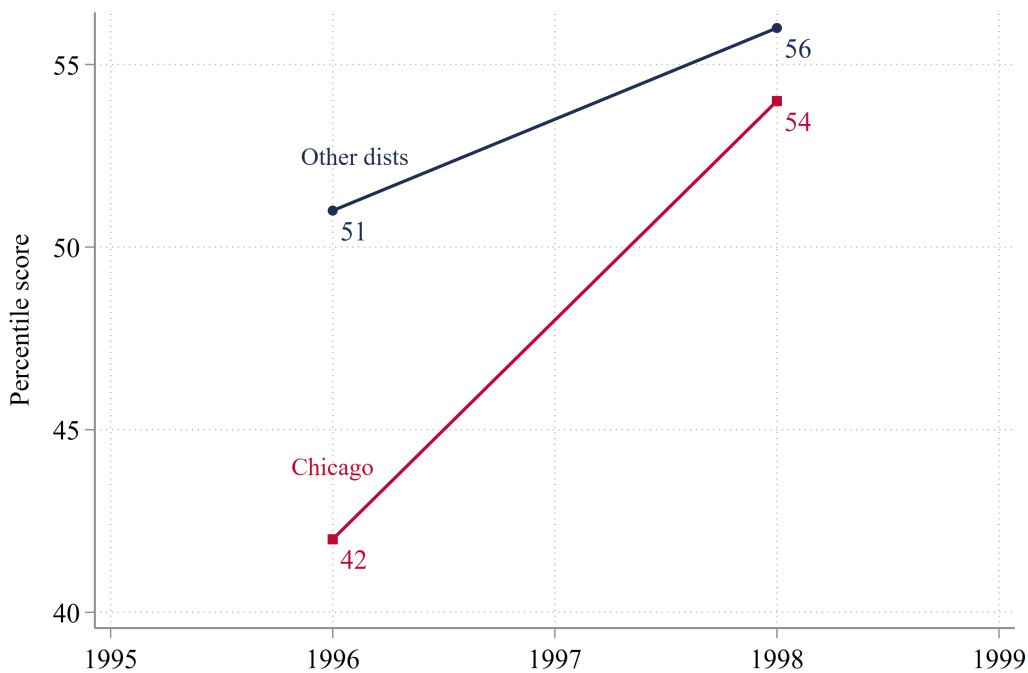
High-stakes testing in Chicago

Consider two types of 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.

An ITS design would be greatly improved by more time periods—to better establish a trend—but this is just for illustration!

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 or ITS for Chicago suggests a large *improvement*:

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

Conflicting conclusions!

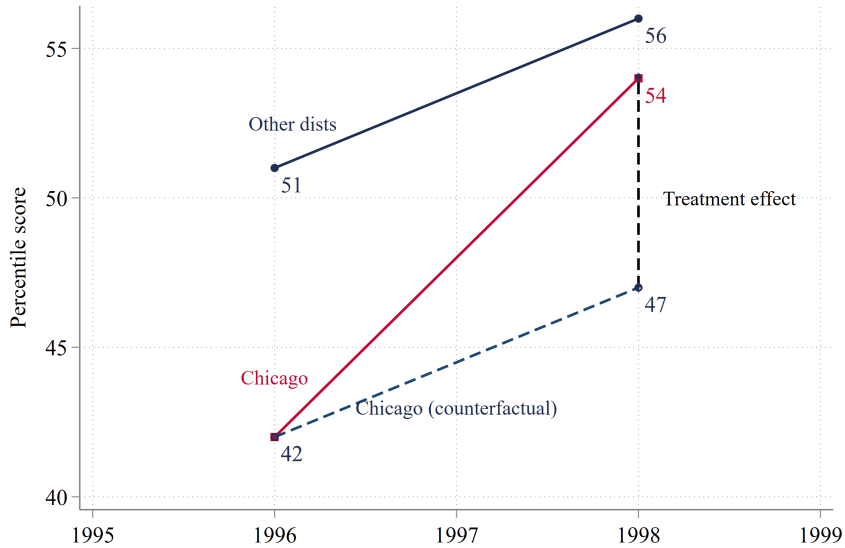
High-stakes testing in Chicago

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 that were common to all districts.

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).

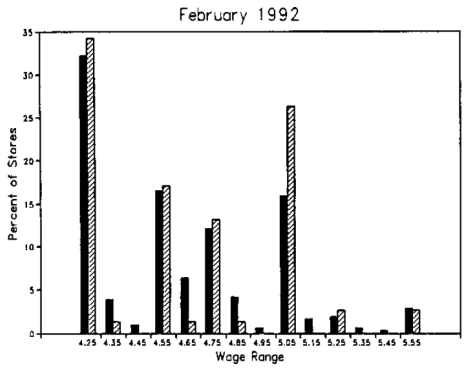
Card & Krueger (1994)

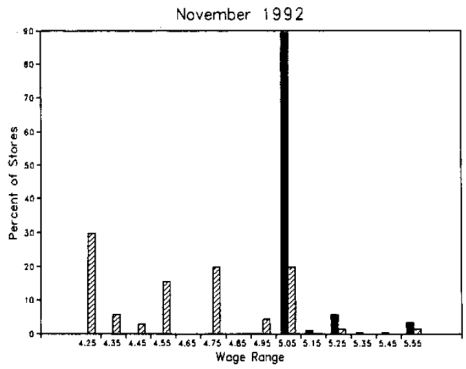
A classic DD study of the impact of the minimum wage on fast food employment (an industry likely to be affected by the minimum wage).

- NJ increased its minimum wage in April 1992, PA did not.
- Card & Krueger collected employment data 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. These kinds of checks are often important in DD studies).

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 (2x2):

- Two groups: a treated and an untreated comparison group
- 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, as we will see later.

Causal interpretation of difference-in-differences

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 apply the potential outcomes framework to a 2x2 setup:

n individuals	indexed i
2 groups	$G_i = 0$ never treated $G_i = 1$ eventually treated
2 time periods	$T_t = 0$ "pre" $T_t = 1$ "post"

Note that G is not subscripted with an t. It is a time-invariant group indicator. In the simple 2x2, the treatment indicator is $D_{it} = G_i \times T_t$. Groups could be states, counties, schools, etc.

Causal interpretation of difference-in-differences

Suppose potential outcomes for individual i at time t are given by:

$$Y_{it}(0) = \beta_0 + \beta_1 G_i + \beta_2 T_t + u_{it}$$
$$Y_{it}(1) = \beta_0 + \beta_1 G_i + \beta_2 T_t + \delta + u_{it}$$

A few things to notice here:

- There are **time-invariant** group differences in mean potential outcomes represented by β_1
- There is a **group-invariant**, common time trend represented by β_2
- The impact of the treatment δ is assumed to be the same for all i, 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

If it helps, think of β_1 as standing in for the combined effects of *all unmeasured covariates* that differ systematically between groups and do not vary over the study period.

Likewise, think of β_2 as standing in for the combined effects of *all unmeasured covariates* that change between periods but affect outcomes in the same way in both groups.

In the DD framework, treatment status D_{it} can be related to G_i (i.e., "self-selection").

Causal interpretation of difference-in-differences

A natural causal parameter of interest is the ATT:

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

$$= (\beta_0 + \beta_1 + \beta_2 + \delta) - (\beta_0 + \beta_1 + \beta_2)$$

$$= \delta$$

The ATT of interest is the difference in mean potential outcomes $Y(1)$ and $Y(0)$ in time period 1 ("post") *among those who are actually treated* (the $G = 1$ 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. The outcome we do observe is:

$$Y_{it} = Y_{it}(0) + D_{it}\left[Y_{it}(1) - Y_{it}(0)\right]$$

$$= \beta_0 + \beta_1 G_i + \beta_2 T_t + D_{it}\delta + u_{it}$$

where $D_{it} = G_i \times T_t$ as defined above.

Causal interpretation of difference-in-differences

Suppose we compare the mean observed outcomes Y_{it} of the $G = 1$ and $G = 0$ groups in time period 1 (post):

$$\underbrace{E[Y|G=1,T=1]}_{\beta_0+\beta_1+\beta_2+\delta} - \underbrace{E[Y|G=0,T=1]}_{\beta_0+\beta_2}$$

$$= \delta + \underbrace{\beta_1}_{\text{selection bias}}$$

This difference does not identify δ since there are baseline differences between the $G = 1$ and $G = 0$ groups..

Causal interpretation of difference-in-differences

Alternatively, we might restrict our attention to the $G = 1$ group and do a pre-post comparison of mean observed outcomes Y_{it}:

$$\underbrace{E[Y|G=1,T=1]}_{\beta_0+\beta_1+\beta_2+\delta} - \underbrace{E[Y|G=1,T=0]}_{\beta_0+\beta_1}$$

$$= \delta + \underbrace{\beta_2}_{\text{common time trend}}$$

This is the first difference or interrupted time series (ITS). Unfortunately, this difference does not identify δ since there is an unaccounted-for time trend.

Causal interpretation of difference-in-differences

Now consider the pre-post comparison for the $G = 0$ group:

$$\underbrace{E[Y|G=0,T=1]}_{\beta_0+\beta_2} - \underbrace{E[Y|G=0,T=0]}_{\beta_0}$$

$$= \beta_2$$

The comparison group allows us to estimate the time trend!

Causal interpretation of difference-in-differences

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

$$\underbrace{E[Y|G=1,T=1]}_{\beta_0+\beta_1+\beta_2+\delta} - \underbrace{E[Y|G=1,T=0]}_{\beta_0+\beta_1} -$$

$$(\underbrace{E[Y|G=0,T=1]}_{\beta_0+\beta_2} - \underbrace{E[Y|G=0,T=0]}_{\beta_0})$$

$$= (\beta_2 + \delta) - \beta_2$$

$$= \delta$$

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

Causal interpretation of difference-in-differences

The above example in table form:

	Pre ($T=0$)	Post ($T=1$)	Diff
Never treated ($G=0$)	β_0	$\beta_0+\beta_2$	β_2
Eventually treated ($G=1$)	$\beta_0+\beta_1$	$\beta_0+\beta_1+\beta_2+\delta$	$\beta_2+\delta$
Diff	β_1	$\beta_1+\delta$	δ

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

Parallel trends assumption

The ATT is again:

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

The DD estimates:

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

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

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

Parallel trends assumption

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

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

Gathering terms, this equals:

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

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

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

Parallel trends assumption

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

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

These canceled out in our case because of how we specified potential outcomes. Keep in mind this is a model we posited for potential outcomes! It may not correspond to reality in a particular case. More on this later.

Difference-in-differences: summary thus far

To summarize:

- Changes over time in the $G = 0$ group provide the counterfactual.
- Selection into treatment related to fixed (time invariant) unobserved differences is OK.
- The outcome *levels* themselves are not important and can vary systematically by group. Only within-group *differences* are used in estimation.
- 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)
- DD is sometimes called a "comparative interrupted time series" (CITS) or nonequivalent control group pretest design. However, the CITS is usually thought of as a more general model than DD. See Section 3 of the MDRC paper by Somers et al. (2013) for a distinction between the two in the context of an educational intervention.

Regression difference-in-differences

Estimation using regression

With many units (i) in two groups observed in "pre" and "post" periods, we can use regression to estimate δ:

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 T_t + \delta(G_i \times T_t) + u_{it}$$

where $G_i = 1$ for units i who are ultimately treated (and 0 otherwise), and $T_t = 1$ for observations in the "post" period. Note 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`

Estimation using regression

You will recognize this as a **two-way fixed effects regression** where group (G_i) is the fixed effect and T_t is the time trend. It could be estimated using the fixed effects "within" estimator: `xtreg, fe` or `areg`.

Stata also has canned `did` commands (more later).

Example (2x2)

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

Estimation using regression

Equivalently, with two time periods we could estimate a regression using first differences for each observation i, subtracting Y_{i0} from Y_{i1}:

$$Y_{i1} = \beta_0 + \beta_1 G_i + \beta_2 + \delta G_i + u_{i1}$$
$$Y_{i0} = \beta_0 + \beta_1 G_i + u_{i0}$$

$$Y_{i1} - Y_{i0} = \beta_2 + \delta G_i + \epsilon_{it}$$

$$\Delta Y_i = \beta_2 + \delta G_i + \epsilon_{it}$$

The intercept here represents the common time trend β_2, and δ is the DD. The baseline differences wash out. (This is rarely done, but useful to see the equivalency).

Estimation using regression

The above model can easily accommodate more than two time periods. Continue to assume two groups and a common "post" period ($POST_t = 1$):

$$Y_{it} = \beta_0 + \beta_1 G_i + \gamma_t + \delta(G_i \times POST_t) + u_{it}$$

The γ_t are time effects—a common intercept shift in each period. The variable $POST_t$ indicates post-treatment time periods. One could put more structure on the time trend and assume linearity:

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 t + \delta(G_i \times POST_t) + u_{it}$$

where t is a linear time trend. This only makes sense, though, if the time trend is really (or approximately) linear.

Estimation using regression

The above regression models can also be extended to include time-varying covariates $\mathbf{X_{it}}$:

$$Y_{it} = \beta_0 + \beta_1 G_i + \beta_2 T_t + \delta(G_i \times T_t) + \mathbf{X_{it}}\eta + u_{it}$$

Careful thought should be put into the implications of including covariates in the model (more on this later). For example, does the parallel trends assumption hold conditional on covariates? Or unconditionally?

In-class exercise Q1

What is the effect of financial assistance on college enrollment? Replicating a simple difference-in-differences result from Murnane & Willett chapter 8 based on Dynarski (2003).

- Data: high school seniors in the National Longitudinal Surveys of Youth (NLSY) 1979-1983 (five cohorts). Includes information on college enrollment and total years of schooling.
- The data include seniors with a deceased parent. Before 1982, college students with a deceased parent were eligible for the Social Security Student Benefit Program. The program was discontinued in 1982.

How would you estimate the effect of the SSBP using a 2x2 DD design?

Generalized difference-in-differences

The examples thus far assumed a common "post" period. In practice, "treatment" often occurs for different groups at different times (e.g., policy adoption). The DD framework easily adapts: this is sometimes referred to as the "generalized difference-in-differences" model.

Suppose potential outcomes for individual i at time t are given by:

$$Y_{it}(0) = \alpha_g + \gamma_t + u_{it}$$
$$Y_{it}(1) = \alpha_g + \gamma_t + \delta + u_{it}$$

The α_g are group effects (e.g., states), γ_t captures the common time trend, and δ is a constant treatment effect. Think of there being a unique intercept for every group α_g and a yearly deviation from that intercept common to all groups, γ_t.

Generalized difference-in-differences

δ can again be estimated using a two-way fixed effects regression with group and time effects:

$$Y_{it} = \alpha_g + \gamma_t + \delta D_{igt} + u_{it}$$

where $D_{igt} = 1$ in time periods in which i's group g is treated. There is no longer a common "post" period, so timing may vary for different groups.

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

Note: we are still assuming (and estimating) a constant treatment effect.

Generalized difference-in-differences

Implementing in Stata: can be done in multiple ways, including `xtreg`. Let $treat_{igt} = 1$ in periods in which i's group is treated.

```
xtreg y i.year i.treat, i(group) fe
```

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

```
reghdfe y i.treat, absorb(group year)
```

Generalized difference-in-differences

Stata 17+ has canned `did` commands. For example, with repeated cross-section data:

```
didregress (y) (treat), group(group) time(year)
```

With panel data on units *id* nested within *group*:

```
xtset id
xtdidregress (y) (treat), group(group) time(year)
```

There are some advantages to the canned commands:

- Stata specifies model for you, so you don't have to worry about getting interaction terms right.
- `did` has some nice post-estimation features.

Continuous treatments

All examples thus far have used a binary treatment (e.g., $D_{it} = G_i \times T_t$). It is common in DD studies to operationalize treatment as a continuous "intensity," "dosage," or "coverage" measure. For example:

- $D_{it} = 0$ if untreated and $D_{it} > 0$ if treated
- D_{it} could be an index of treatment *intensity* (e.g., law strength, new school construction per capita)
- Or, a measure of *coverage*, like the proportion of a population affected by a policy change

Swap in this continuous measure for D_{it} above and the interpretation of δ becomes the effect of a 1-unit change in this treatment intensity measure.

In-class exercise Q2

What is the effect of a lower Minimum Legal Drinking Age (MLDA) on traffic fatalities among young adults? Example from *Mastering 'Metrics* chapter 5 based on Carpenter & Dobkin (2011).

- Following the 26th Amendment (1971) some states lowered the MLDA to 18.
- In 1984, federal legislation pressured states to increase MLDA to 21. Between 1971-1984, there was a lot of variation across states and years in the MLDA.
- Was a lower MLDA associated with more traffic fatalities among 18-20 year olds?

We will use panel (*state* $\times$ *year*) data and generalized DD to address this question.

In-class exercise Q2

One could use a binary "treatment" variable here (e.g., $D_{st} = 1$ if the MLDA is below 21 in state s and year t and zero otherwise). However, this neglects some potentially interesting variation in MLDA.

We will use a continuous "dosage" measure, $LEGAL_{st}$, defined as the proportion of adults aged 18-20 who could legally drink in state s in year t. This measure also takes into account within-year changes in the MLDA (e.g., if a state raises its MLDA from 18 to 21 mid-year, $LEGAL_{st} = 0.5$).

- Min: 0 (drinking age = 21 all year)
- Max: 1 (drinking age = 18 all year)

In-class exercise Q2

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

$$Y_{st} = \alpha_g + \gamma_t + \delta LEGAL_{st} + u_{st}$$

The coefficient δ represents how much, on average, mortality rates differ when the MLDA is 18 (relative to 21), conditional on state and year. (In other words, beyond that predicted by the state and year effects).

Note: Y_{st} is mortality per 100,000 population.

Defending the parallel trends assumption

Parallel trends assumption

The DD design leans heavily on the parallel trends assumption:

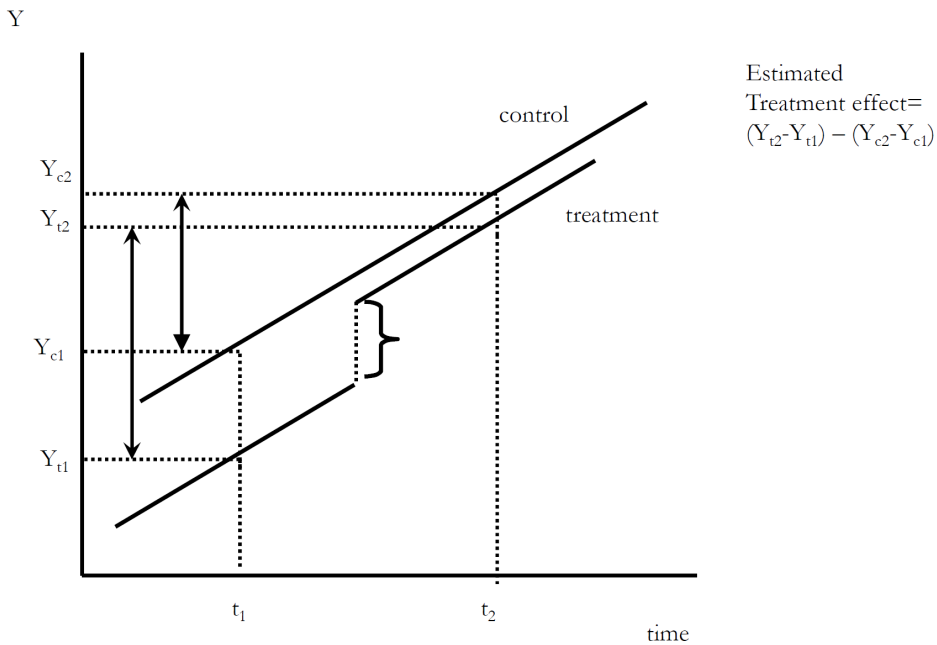
- Confounding factors varying across groups are time-invariant
- Time-varying confounding factors are group-invariant

i.e., there are no group-specific, unobserved, time varying factors that would lead to groups to follow different time paths.

Defending the parallel trends assumption is usually the most important part of a DD design.

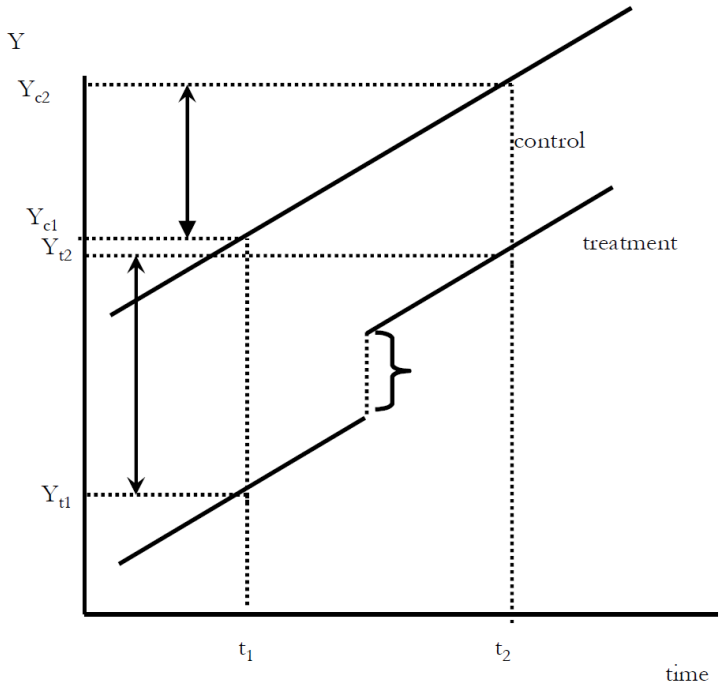
Parallel trends assumption

Parallel trends implies the time trend in the absence of treatment would be the same in both groups:



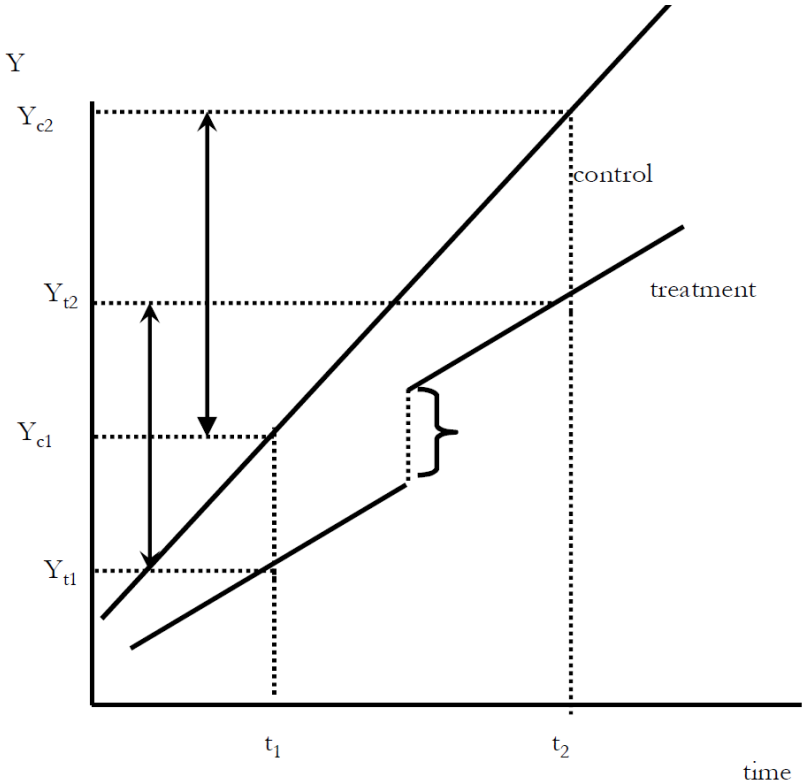
Parallel trends assumption

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



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

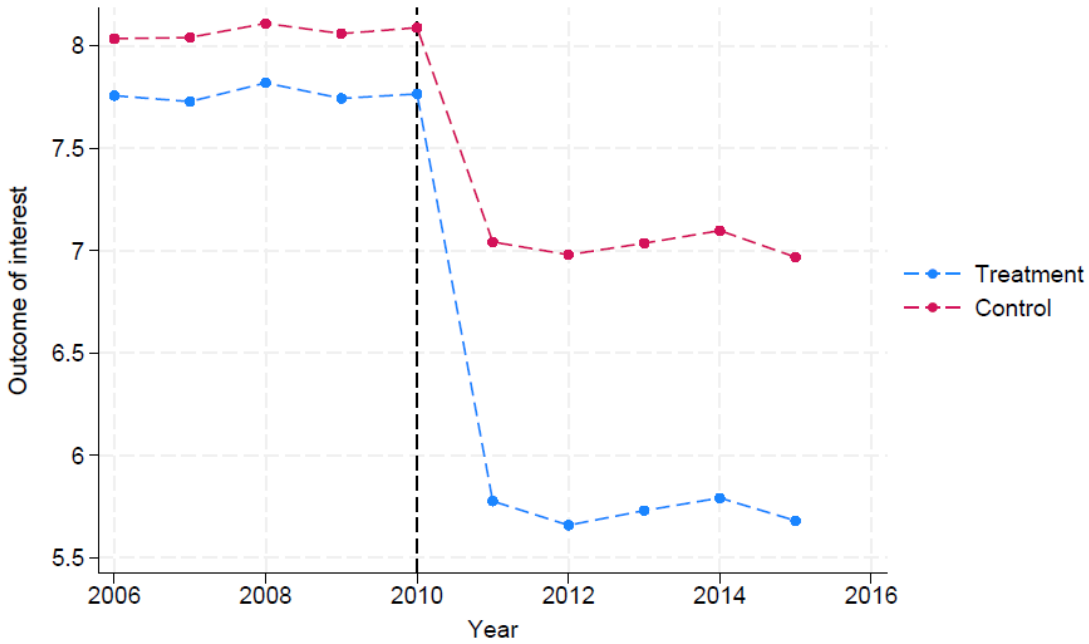
Violation of parallel trends assumption



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

Parallel trends assumption

Time trends need not be linear—here, individual year effects:



Common violations of parallel trends assumption

Common scenarios that would violate the parallel trends assumption:

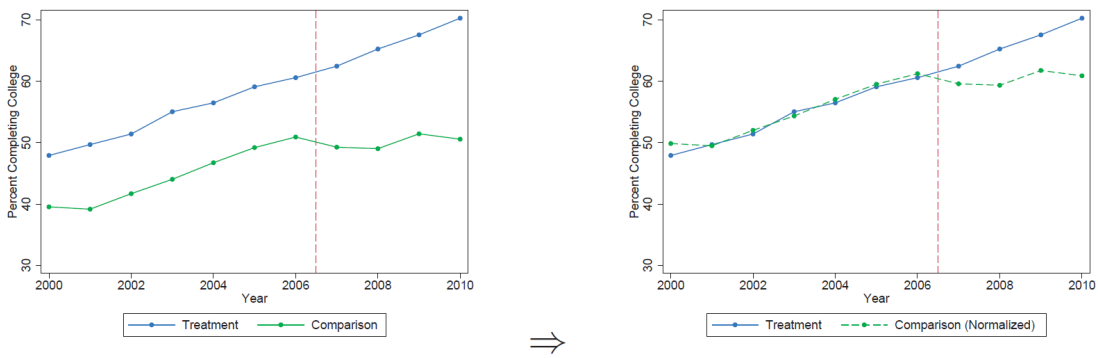
- **Targeted treatments**: often programs are targeted at subjects who are most likely to benefit from it. In many cases, the fact that a subject was on a different trajectory is what made them a good candidate for the program (e.g., a struggling student).
- **Ashenfelter's dip**: treated cases may experience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

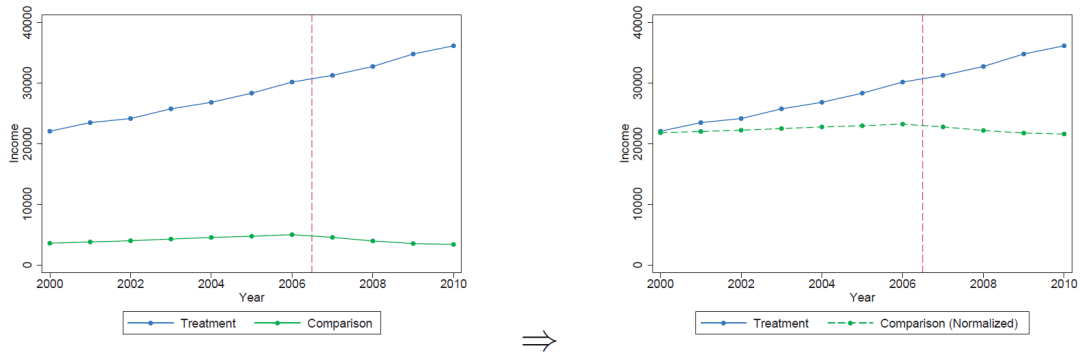
- A compelling graph: point to similar trends *prior to* treatment. Note: parallel trends *prior* to treatment are not sufficient for the parallel trends assumption, which is about the *post* period! But it helps.
- Statistical tests for differences in pre-treatment trends
- Event study regression and graph (Lecture 5)
- A placebo / falsification test
- Tests for differential trends in covariates
- Controlling for time trends directly (leans heavily on functional form)
- Triple-differences model
- Probably most important: understanding the context of your study! Ruling out potential reasons for differential time 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.

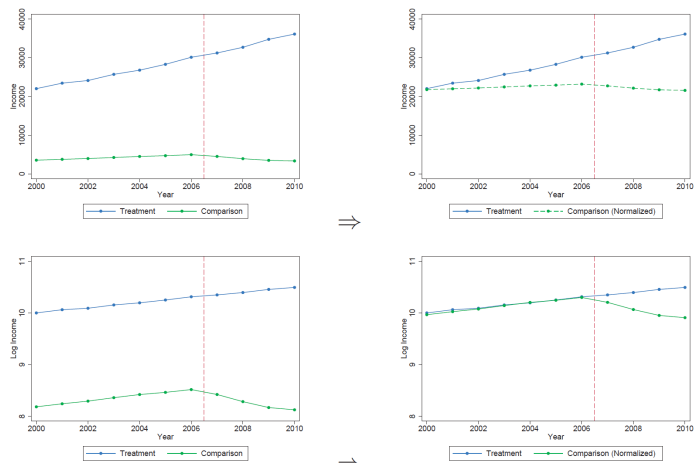
Graphical assessment of parallel trends assumption



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

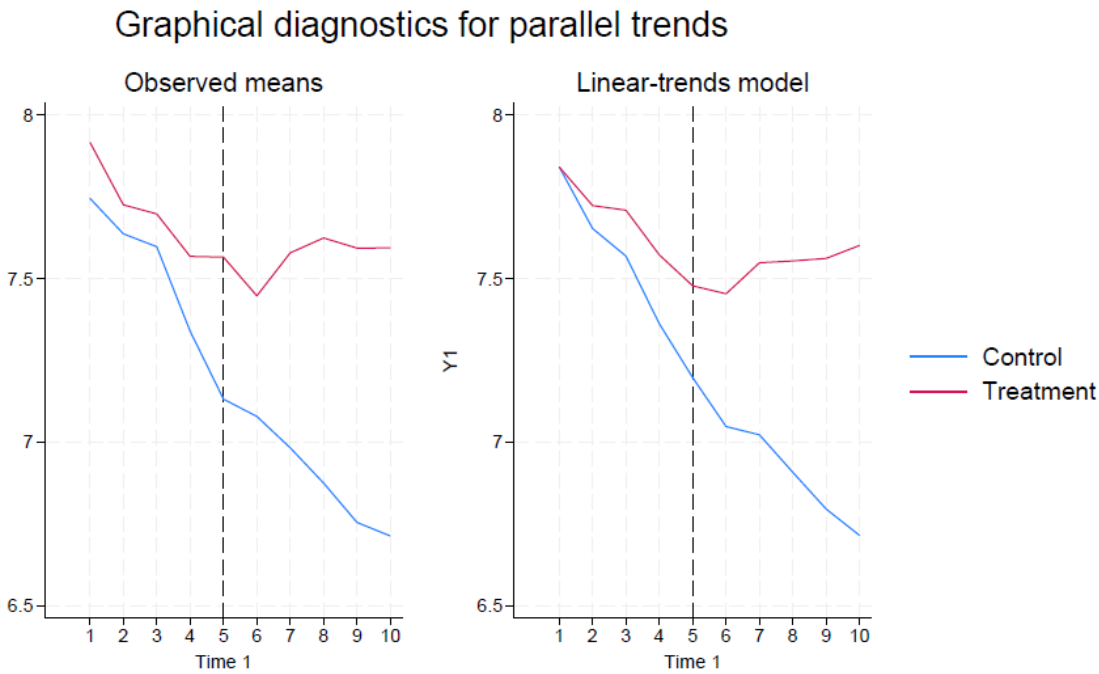
Note: if trends are parallel in levels they will *not* be parallel in logs, and vice versa!



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.

Stata `estat trendplots`

Following `didregress` or `xtdidregress`: `estat trendplots`



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 nonlinear "anticipatory" effects. See `estat granger`.

In practice, event studies are more common than these tests. (In fact `estat granger` is a special case of an event study).

Stata `estat ptrends`

Following `didregress` or `xtdidregress`: `estat ptrends`

```
. xtdidregress (y1) (treated1), group(id1) time(t1)

Treatment and time information

Time variable: t1
Control:       treated1 = 0
Treatment:     treated1 = 1
```

	Control	Treatment
Group		
id1	102	98
Time		
Minimum	1	6
Maximum	1	6

```
Difference-in-differences regression                 Number of obs = 2,000
Data type: Longitudinal

                                        (Std. err. adjusted for 200 clusters in id1)
```

y1	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
ATET						
treated1						
(Treated vs Untreated)	.4825449	.0275446	17.52	0.000	.4282281	.5368616

```
Note: ATET estimate adjusted for panel effects and time effects.

. estat ptrends

Parallel-trends test (pretreatment time period)
H0: Linear trends are parallel

F(1, 199) =  19.75
 Prob > F = 0.0000
```

Stata `estat granger`

Following `didregress` or `xtdidregress`: `estat granger`

```
. xtdidregress (y1) (treated1), group(id1) time(t1)

Treatment and time information

Time variable: t1
Control:       treated1 = 0
Treatment:     treated1 = 1
```

	Control	Treatment
Group		
id1	102	98
Time		
Minimum	1	6
Maximum	1	6

```
Difference-in-differences regression                 Number of obs = 2,000
Data type: Longitudinal

                                        (Std. err. adjusted for 200 clusters in id1)
```

y1	Coefficient	Robust std. err.	t	P>|t|	[95% conf. interval]	
ATET						
treated1 (Treated vs Untreated)	.4825449	.0275446	17.52	0.000	.4282281	.5368616

```
Note: ATET estimate adjusted for panel effects and time effects.

. estat granger

Granger causality test
H0: No effect in anticipation of treatment

F(4, 199) =   9.14
 Prob > F = 0.0000
```


Placebo/falsification tests

The DD design assumes that any change over time for the treated 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 common approach is to apply the same treatment assignments to an earlier period, before 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. Be mindful that actual treated periods are not included in the analysis.

In-class exercise Q2

For the MLDA example, estimate the effect of $LEGAL_{st}$ on:

- Non-alcohol related causes of mortality
- Mortality rates for older, unaffected cohort (age 21-24)

Tests for differential changes in covariates

It is not unusual for treated and untreated groups to be imbalanced in the "pre" period. This is OK and is a feature of DD designs. DD nets out fixed differences and focuses on within-group changes over time. Differential changes over time in key covariates, however, could be an indication of other unobserved time-varying changes specific to treated groups.

One could fit the same DD regression model but with **covariates** as the outcome. The goal here is to rule out differential changes over time for the treated groups (that are meaningful in size).

Group-specific time trends

Another approach is to augment the DD model by including group-specific linear time trends. That is, a separate linear trend for each group:

$$Y_{it} = \alpha_g + \gamma_t + \beta_g(\alpha_g \times t) + \delta D_{it} + u_{it}$$

In practice, this means dummy variables for each group and time period as well an interaction between every group and time (a linear time variable).

Can perform an F-test for joint significance of the group-specific linear time trends.

Group-specific time trends

How does this work? Consider again potential outcomes:

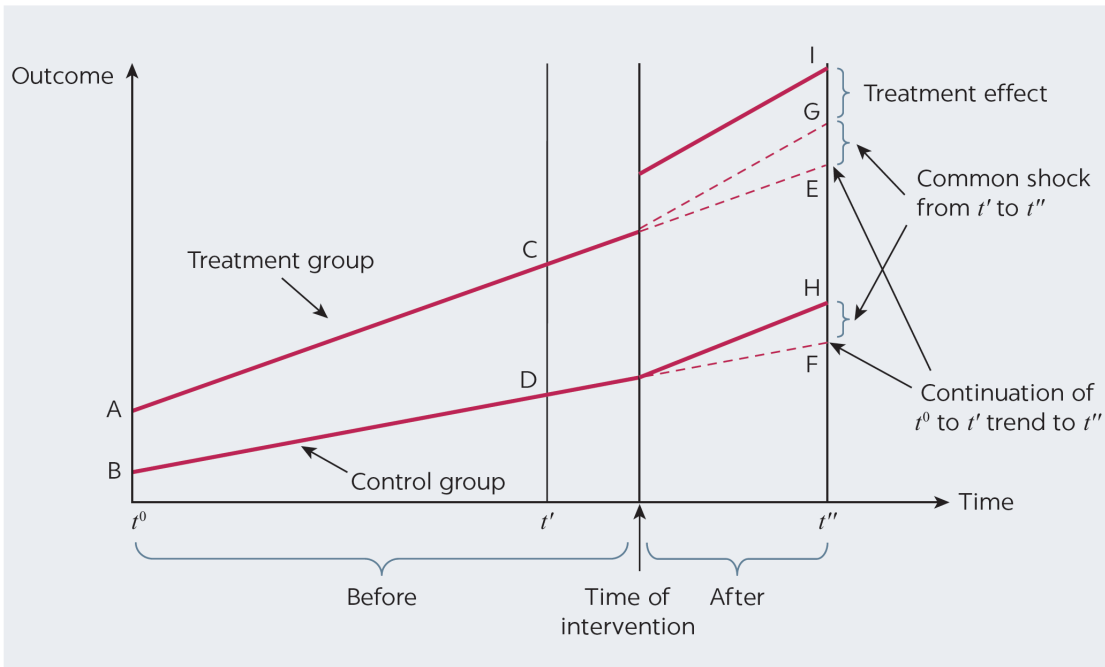
$$Y_{it}(0) = \alpha_g + \beta_g t + \gamma_t + u_{it}$$
$$Y_{it}(1) = \alpha_g + \beta_g t + \gamma_t + \delta + u_{it}$$

Potential outcomes are described by a unique intercept and linear time trend for each group ($\alpha_g + \beta_g t$) as well as a yearly deviation from this level common to every group (γ_t).

Intuitively, under the assumption that changes within groups over time are accurately described by the group-specific linear time trend and common year effect, 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

Above:

- The two "pre" periods allow us to estimate two linear time trends, one for the treated and one for the untreated. (They are not parallel).
- Had the linear time trends continued, the groups would be at E and F.
- The model allows for a common "post" time effect, which is the difference between F and H.
- We impute the common time effect to the treatment group, which would be E-G.
- Any *differential* change for the treatment group beyond that predicted by their time trend (E) and the common "post" effect (E-G) is the treatment effect (G-I).

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.

In-class exercise Q2

For the MLDA example, estimate the original model (for age 18-20) but include state-specific time trends.

Triple difference models

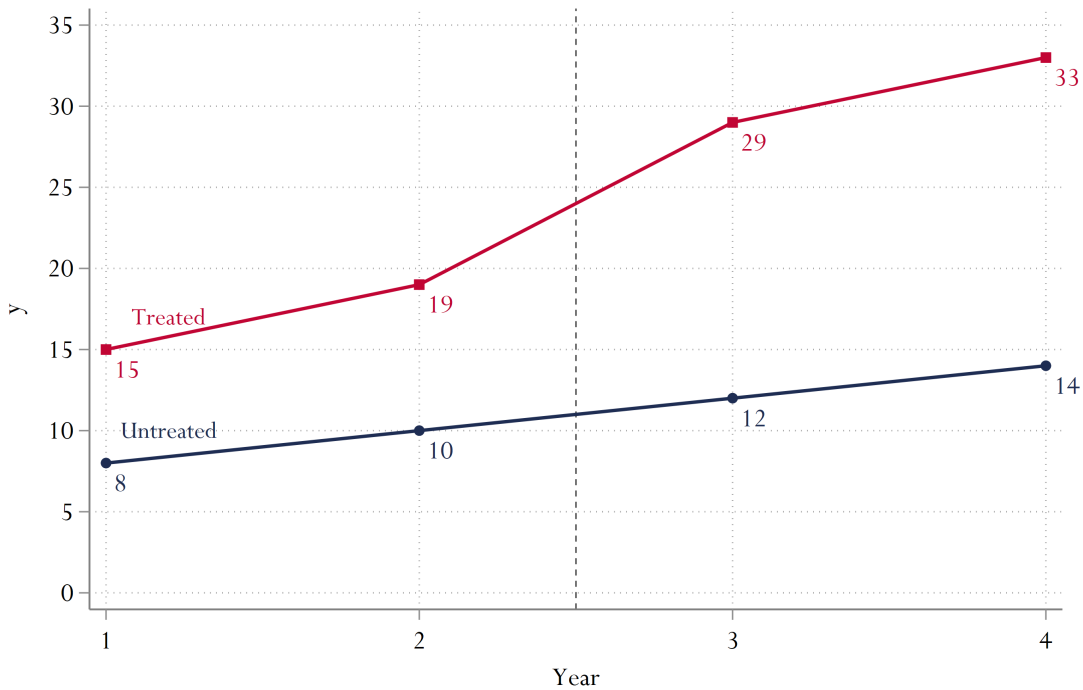
Triple difference

The **triple difference** uses an additional comparison group to remove differential time trends. This comparison group is exposed to the problematic time-varying confounder but not the treatment. 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.

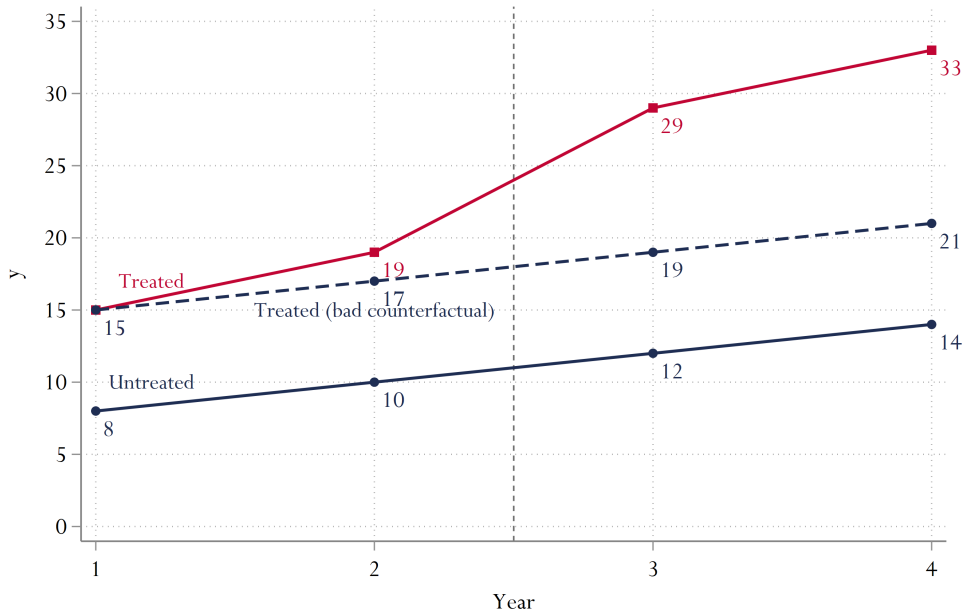
Triple difference

Stylized example: non-parallel trends



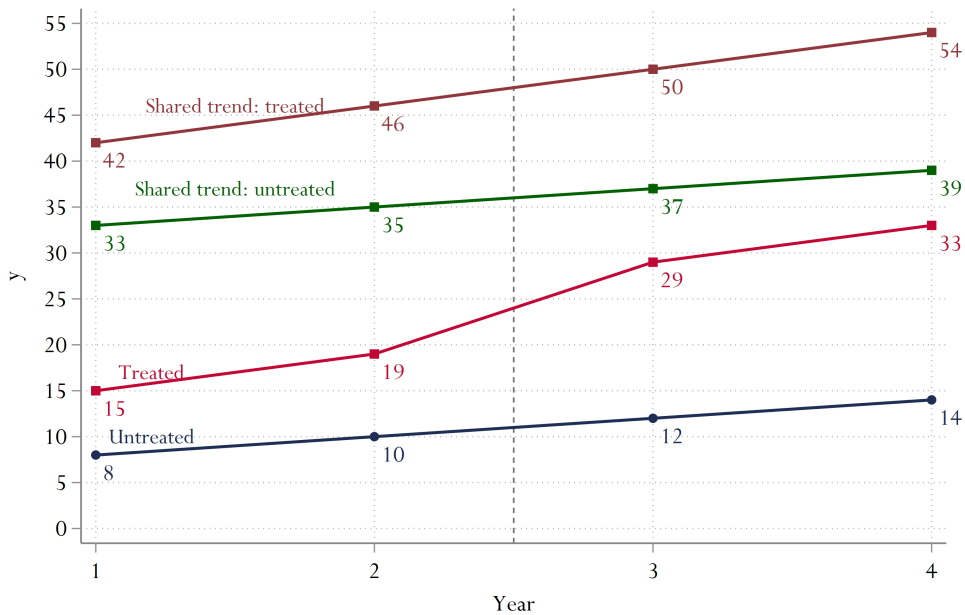
Triple difference

The untreated group is a bad counterfactual for the DD. Using means in the pre and post periods, the 2x2 DD estimate $(31 - 17) - (13 - 9) = 10$ overstates the treatment effect because of the differential time trend.



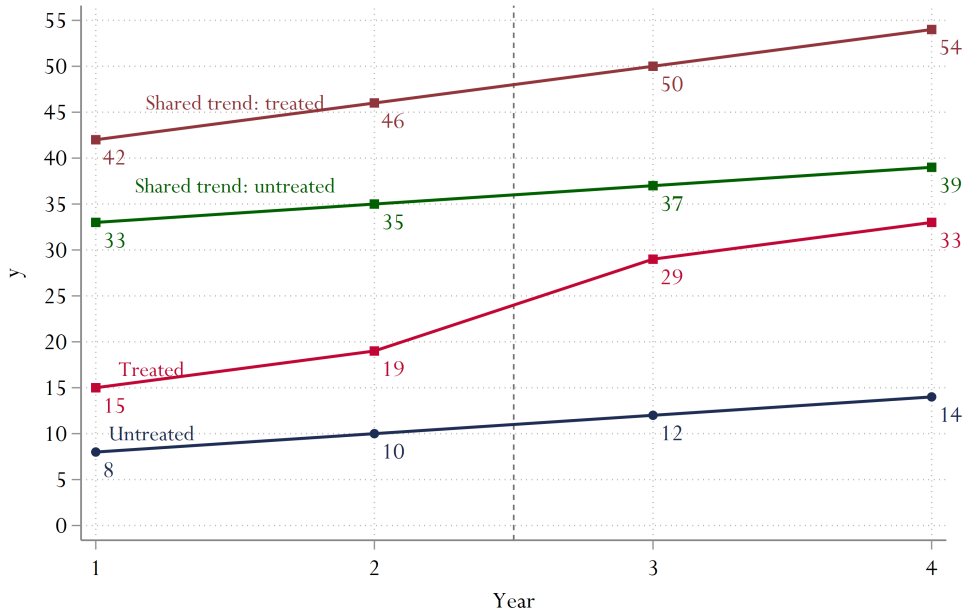
Triple difference

Suppose we have another comparison group that shares 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:

$$Y_{it} = \beta_0 + \beta_1 G1_i + \beta_2 G2_i + \beta_3 (G1_i \times G2_i) + \beta_4 POST_t$$
$$+\beta_5 (G1_i \times POST_t) + \beta_6 (G2_i \times POST_t) + \beta_7 (G1_i \times G2_i \times POST_t) + u_{it}$$

- $POST_t = 1$ in the post period (assuming 2 periods)
- $G1_i = 1$ for the original group at which level treatment is assigned
- $G2_i = 1$ for the "primary" comparison group and $G2_i = 0$ for the additional "secondary" comparison group

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
```

Or use `did` commands in Stata (see below).

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		
Model	3319	7	474.142857	Number of obs	= 16
Residual	40	8	5	F(7, 8)	= 94.83
				Prob > F	= 0.0000
				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:

- β_0 (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 in Stata

Stata's `did` commands are quite useful for DDD models since they take care of the interactions for you.

`didregress (`*y*`) (`*treat*`), group(`*group1 group2*`) time(`*time*`)`

Be sure that *treat* here is defined as "ever treated" $\times$ "primary" $\times$ "post".

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 the same county in a triple difference design.

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

In-class exercise Q2

For the MLDA example, estimate a triple difference using an unaffected age group (21-24) as the second comparison group.

Parallel trends: knowing your context

Perhaps the most important element of defending the parallel trends assumption is understanding your study context.

- What conceptual/contextual reasons might lead to a violation of common trends?
- Are the most theoretically important factors covered by your DD design?
- It is useful to name unmeasured variables that your fixed effects purportedly capture.
- Being specific can lead lead to you a stronger research design, perhaps by:
  - Excluding certain groups where common trends is unlikely to hold.
  - Constructing a synthetic control group (Lecture 6).
  - Constructing a matched comparison group prior to DD

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

Issues with Staggered Treatment Adoption Timing

Recent developments in difference-in-differences

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 average of treatment effects across units and time.

$$Y_{it} = \alpha_g + \gamma_t + \delta D_{it} + u_{it}$$

Recent research has identified previously unrecognized issues with this interpretation, and highlighted cases in which the TWFE estimator does not yield an ATT of interest.

Stylized example

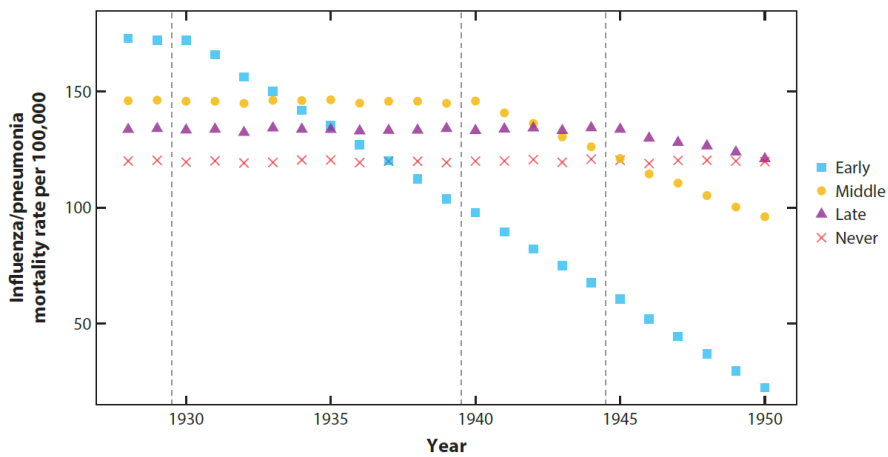


Figure 1

Stylized example: pneumonia/influenza mortality and sulfa drugs. This plot presents a stylized example inspired by Jayachandran et al. (19). Each point is the annual death rate per 100,000 population from influenza and pneumonia for a given group. Data are constructed so that there are four groups: The early-treated group sees an annual decline beginning in 1930, the middle-treated group sees the decline beginning in 1940, the late-treated group sees the decline in 1945, and the never-treated group has a constant death rate. There is also treatment effect heterogeneity. The annual decline is largest for the early-treated group (an additional 7.5 decline in the death rate in each year following treatment) and smallest (additional decline of 2.5 per year) for the latest-treated group. The middle-treated group sees a reduction of an additional 5 deaths per 100,000 in each year following treatment.

Source: Wing et al. (2024)

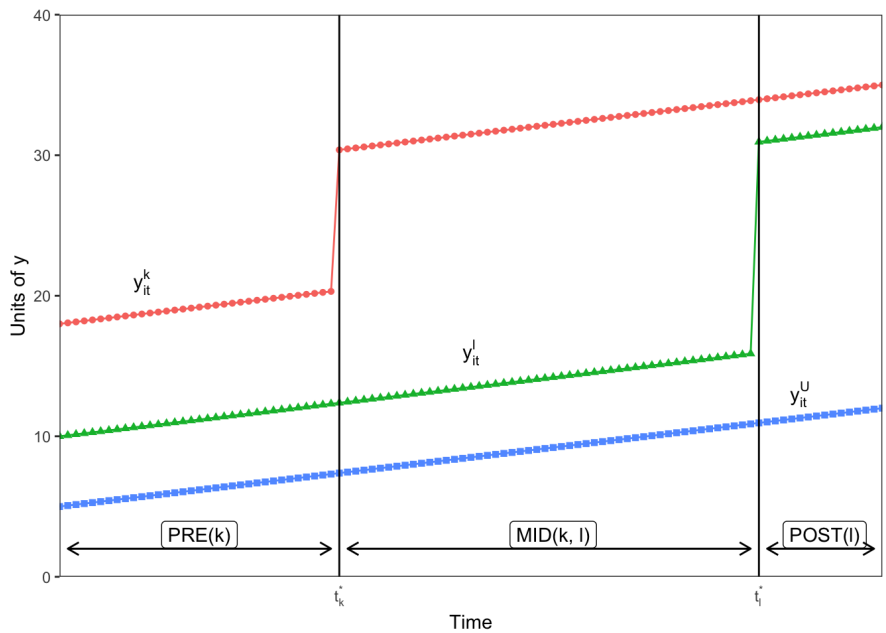
Stylized example

What is the impact of access to sulfa drugs on mortality rates from pneumonia/influenza?

- Suppose "treatment" at the state level is staggered (1930, 1940, 1945, and never).
- Treatment effects vary over time: a gradual reduction of mortality.
- Treatment effects vary by "timing group": largest effects among states treated first.
- Large baseline differences in mortality.
- Treatments are "absorbing": they turn on and stay on.

Staggered adoption

Staggered adoption "obscures distinctions between treatment vs. control, pre vs. post." Goodman-Bacon (2021) points out that "early adopters" in a TWFE design serve as a comparison group for "late adopters".



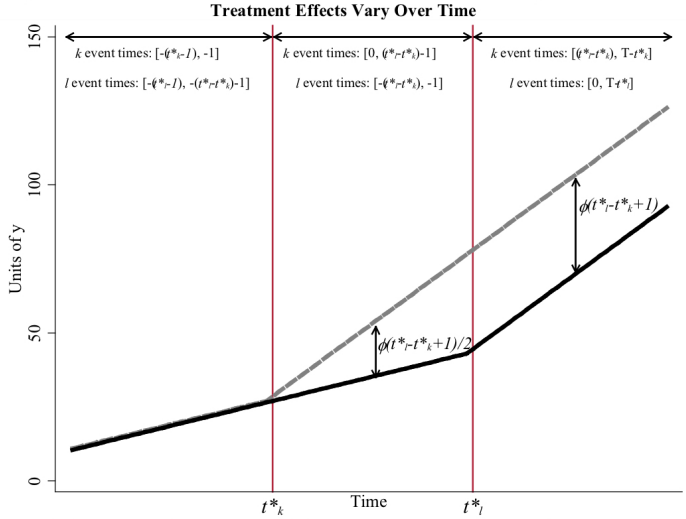
Staggered adoption: Goodman-Bacon (2021)

What are the implications of this?

- When treatment effects are *constant* (and there are common trends), the TWFE estimator provides the ATT. Good to go!
- However, when treatment effects are *heterogeneous*, the TWFE is *not* the ATT you are looking for.
  - Treatment effects could be heterogeneous in *time since treatment*
  - Treatment effects could be heterogeneous *across units*

Staggered adoption: Goodman-Bacon (2021)

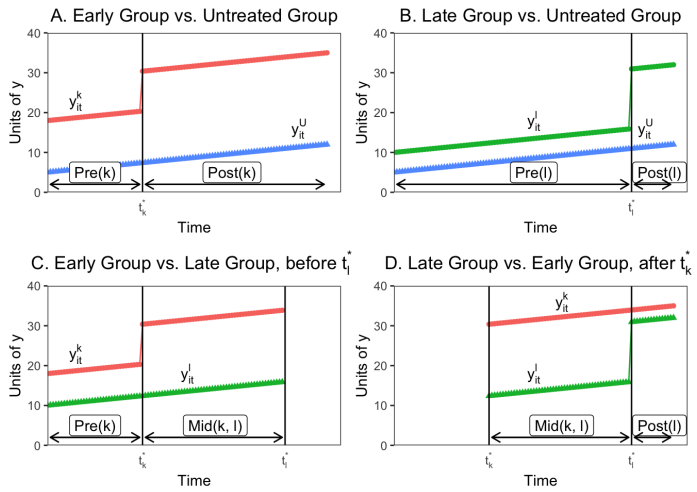
Figure 3. Difference-in-Differences Estimates with Variation in Timing Are Biased When Treatment Effects Vary Over Time



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.

Staggered adoption: Goodman-Bacon (2021)

Another insight: the TWFE estimator is a weighted average of all possible 2x2 DD estimators in the data:



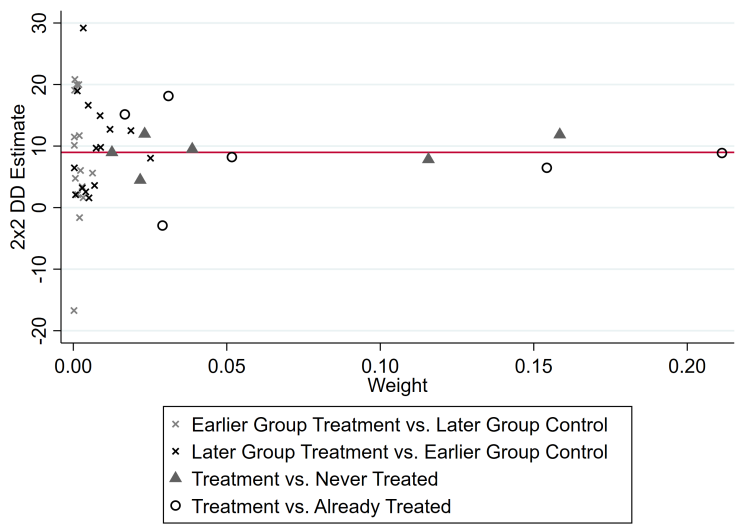
Weights come from group sizes *and* the share of time each group is treated. (Groups treated in the middle get the most weight). This is probably not the weighting you want.

Staggered adoption: Goodman-Bacon (2021)

Drawing on this insight, one can decompose the TWFE estimator into its component 2x2 parts to see which "timing groups" are getting greater weight. Some treatment observations may even receive *negative* weights which can be problematic. This diagnostic shows how 2x2 treatment effect estimates differ across groups and the weights each group receives.

"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.



Staggered adoption design

The recent literature builds on this insight and conceives of the staggered adoption design as a collection of simpler DDs ("sub-experiments"). The big question is how to combine these into one or more treatment effects of interest.

Notation:

- Individual units $i = 1, ..., N$
- Timing groups $g = 1, ..., G$
- Time periods $t = 1, ..., T$
- Timing groups are defined by when they first are exposed to treatment (A_g)
- D_{igt} is time-varying treatment indicator

Staggered adoption design

Potential outcome notation:

- $Y_{igt}(0)$ is the outcome for unit i in group g in time t if group g was *never treated*
- $Y_{igt}(a)$ is the outcome for unit i in group g in time t if group g was first treated in period a.

Now, the causal effect can vary by unit, timing group, and time period:

$$\beta_{igt}(a) = Y_{igt}(a) - Y_{igt}(0)$$

We are usually interested in the *timing group* $\times$ *time* average of these treatment effects. The recent literature views these as the key building blocks of the staggered adoption design: $ATT(g, t)$

Staggered adoption design

DD can recover $ATT(g, t)$ under the usual assumptions:

- No anticipation: the average effect of adopting the treatment in period a is equal to zero for all periods prior to a.
- Parallel trends: the average change across post-treatment periods would be the same in the treatment group and the comparison group.

Who is the comparison group? Can justify using the *never treated* and the *not yet treated* as long as the above assumptions hold. These are "clean" control groups. The DD assumptions are unlikely to hold for *already treated* groups.

Staggered adoption design

The literature now recommends a variety of "heterogeneity-robust" estimators (see Roth et al. 2022, Wing et al. 2024):

- **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

All of these approaches share an interest in making "clean" comparisons and avoiding "forbidden" ones. See Wing et al. (2024) for a useful tabular summary.

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).