4. Difference-in-Differences

LPO 8852: Regression II

Sean P. Corcoran

Difference-in-differences

Difference-in-differences (DD) is a research design that—in its most common application—contrasts *changes over time* for treated and untreated groups. The approach 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 often estimated with *panel* data but can also be used with *repeated cross-sections.*

Natural experiments

Examples of natural experiments:

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

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

Chicago high-stakes testing

Do high-stakes accountability policies improve student academic performance?

- A potential "natural experiment": in Chicago, the Illinois State Aptitude Test (ISAT) became "high stakes" in 2002. The test was administered—but was "low stakes"—prior to that year. The test is given in grades 3, 5, and 8.
- This means 5th graders in 2002 were "treated" by the accountability policy while 5th graders in 2001 were not. Neither cohort was treated in 3rd grade.

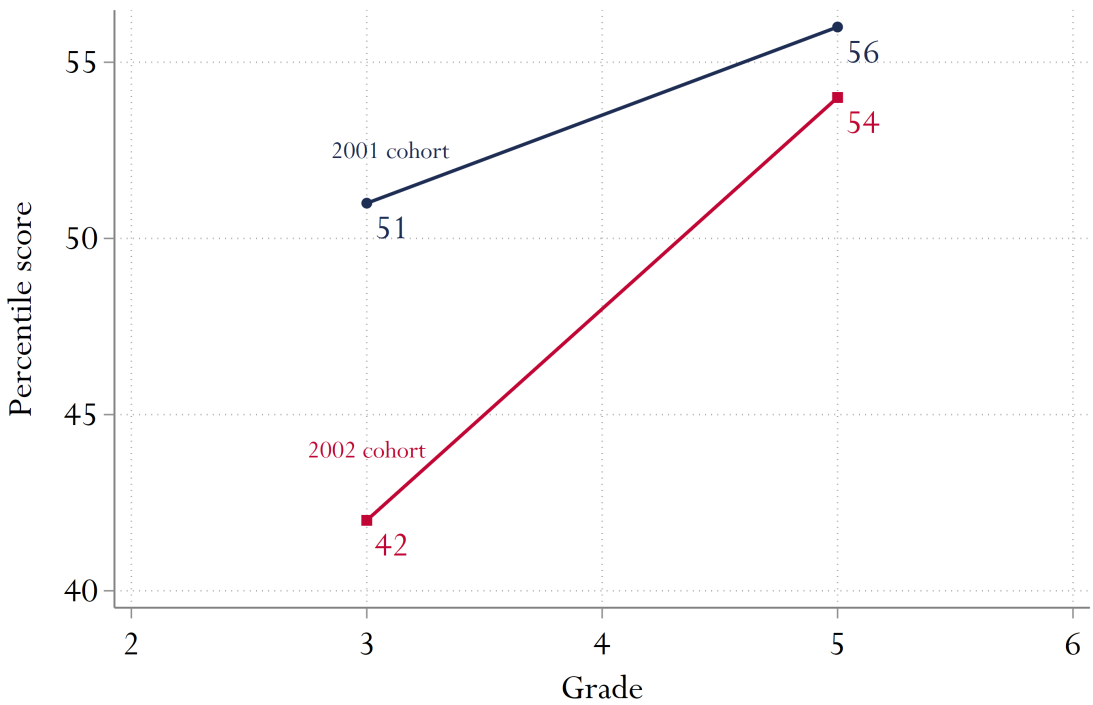
Chicago high-stakes testing

Consider two comparisons:

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

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

Chicago high-stakes testing



Chicago high-stakes testing

The cross sectional comparison of 5th grade cohorts suggests *worse* outcomes for the 2002 (accountability regime) cohort:

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

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

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

Conflicting conclusions!

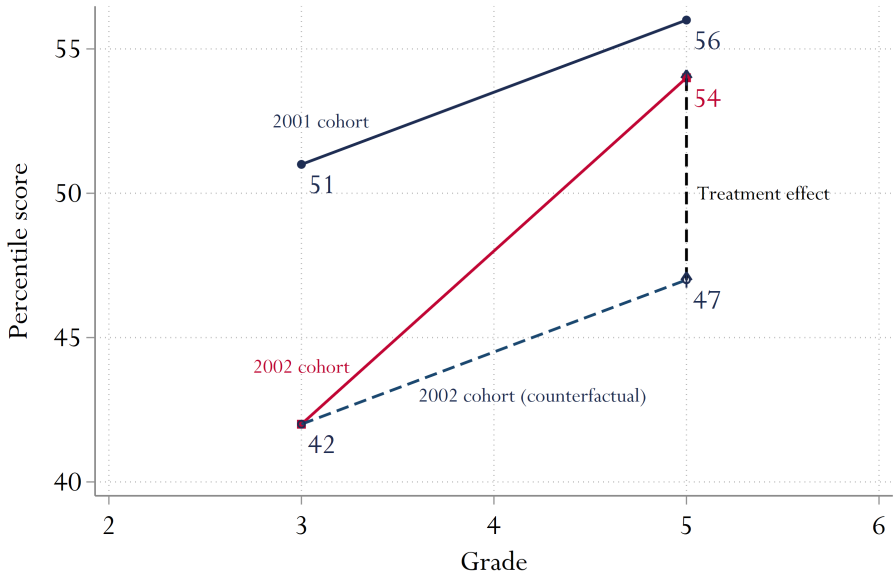
Chicago high-stakes testing

Problems:

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

Chicago high-stakes testing

Under the assumption that the change over time for the 2001 (untreated) cohort represents what *would have happened* to the 2002 (treated) cohort in the absence of treatment, we can contrast *changes* in the two, or the **difference-in-differences**:



Chicago high-stakes testing

The difference-in-differences:

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

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

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

Chicago high-stakes testing

An equivalent way to write δ_{DD}:

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

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

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

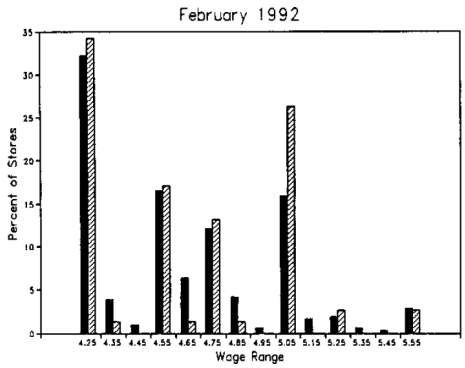
Card & Krueger (1994)

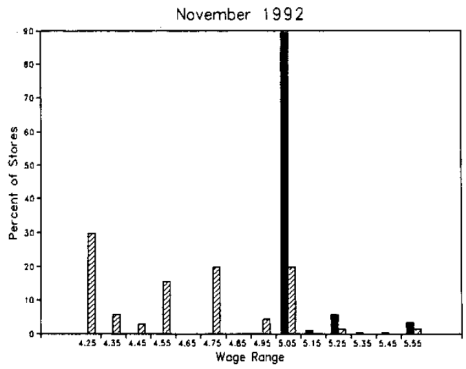
A classic DD study of the impact of the minimum wage on fast food employment (an industry likely to be affected by the minimum wage).

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

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

Card & Krueger (1994)





New Jersey Pennsylvania

FIGURE 1. DISTRIBUTION OF STARTING WAGE RATES

Card & Krueger (1994)

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

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

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

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

2x2 difference-in-differences

The two examples thus far are the simplest form of difference-in-differences:

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

Difference-in-differences estimation

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

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

Difference-in-differences estimation

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

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

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

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

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

Difference-in-differences estimation

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

A few things to note:

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

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

Difference-in-differences estimation

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

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

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

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

Difference-in-differences estimation

The causal estimand of interest is:

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

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

$$= \delta$$

That is, the mean difference in outcomes in the treated and untreated state—in the "post" period—*among those who are treated* (the $D_i = 1$ group).

Difference-in-differences estimation

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

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

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

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

Difference-in-differences estimation

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

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

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

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

Difference-in-differences estimation

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

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

$$= \lambda_1 - \lambda_0$$

The comparison group allows us to estimate the time trend!

Difference-in-differences estimation

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

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

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

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

$$= \delta$$

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

Difference-in-differences estimation

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

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

The DD estimates:

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

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

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

Difference-in-differences estimation

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

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

Gathering terms, this equals:

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

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

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

Difference-in-differences estimation

To summarize:

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

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

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

Regression difference-in-differences (2x2)

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

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

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

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

Regression difference-in-differences (2x2)

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

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

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

Regression difference-in-differences (2x2)

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

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

Regression (2x2) DD is effectively a comparison of four cell-level means. Note OLS will always (mechanically) estimated δ as the differential change between the $D_i = 1$ and $D_i = 0$ groups. Whether that δ can be interpreted as the ATT depends on the parallel trends assumption.

Regression difference-in-differences (2x2)

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

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

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

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

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

Regression difference-in-differences (2x2)

The 2x2 regression model can also include covariates:

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

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

Example: Dynarski (2003)

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

- Table 8.1 from Murnane & Willett on next page begins with the ITS design, focusing only on survivors (a first difference)

- Data: NLSY high school seniors who would be eligible for benefits just before (N=137) and after (N=54) the policy change.

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

Example: Dynarski (2003)

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

(a) *Direct Estimate*

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

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

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

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

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

Example: Dynarski (2003)

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

Example: Dynarski (2003)

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

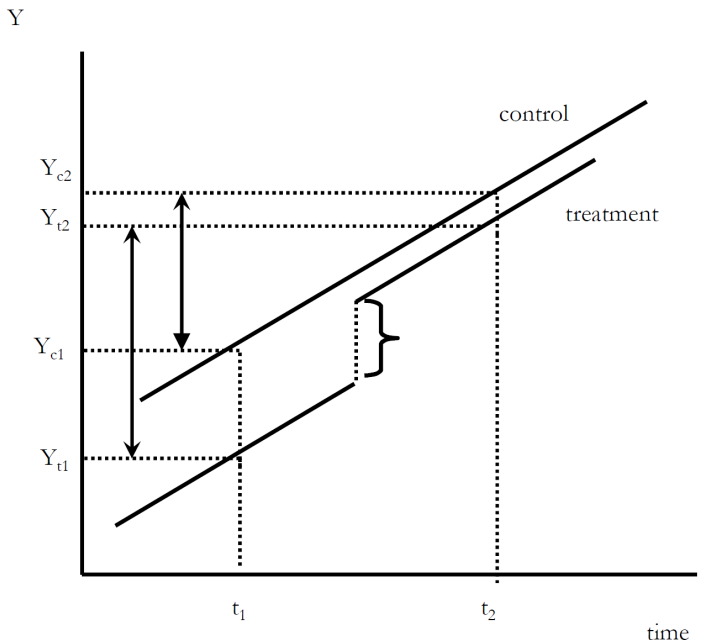
H.S. Senior Cohort	Number of Students	Was Student's Father Deceased?	Did H.S. Seniors Receive an Offer of SSSB Aid?	Avg Value of *COLL* *(standard error)*	Between-Group Difference in Avg Value of *COLL*	"Difference in Differences"	
						Estimate *(standard error)*	*p*-value
1979–81	137	Yes	Yes *(Treatment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

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

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

Parallel trends assumption

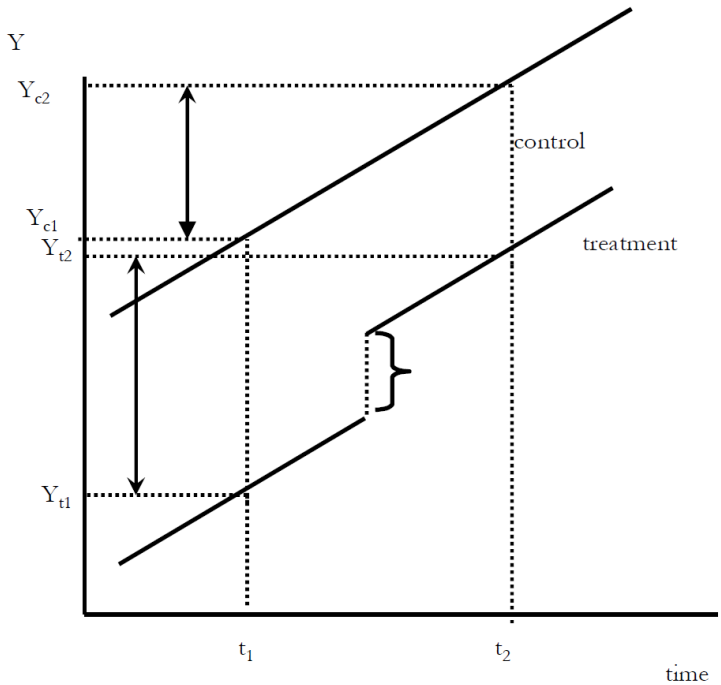
The key assumption in DD is parallel trends: that the time trend in the absence of treatment would be the same in both groups.



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

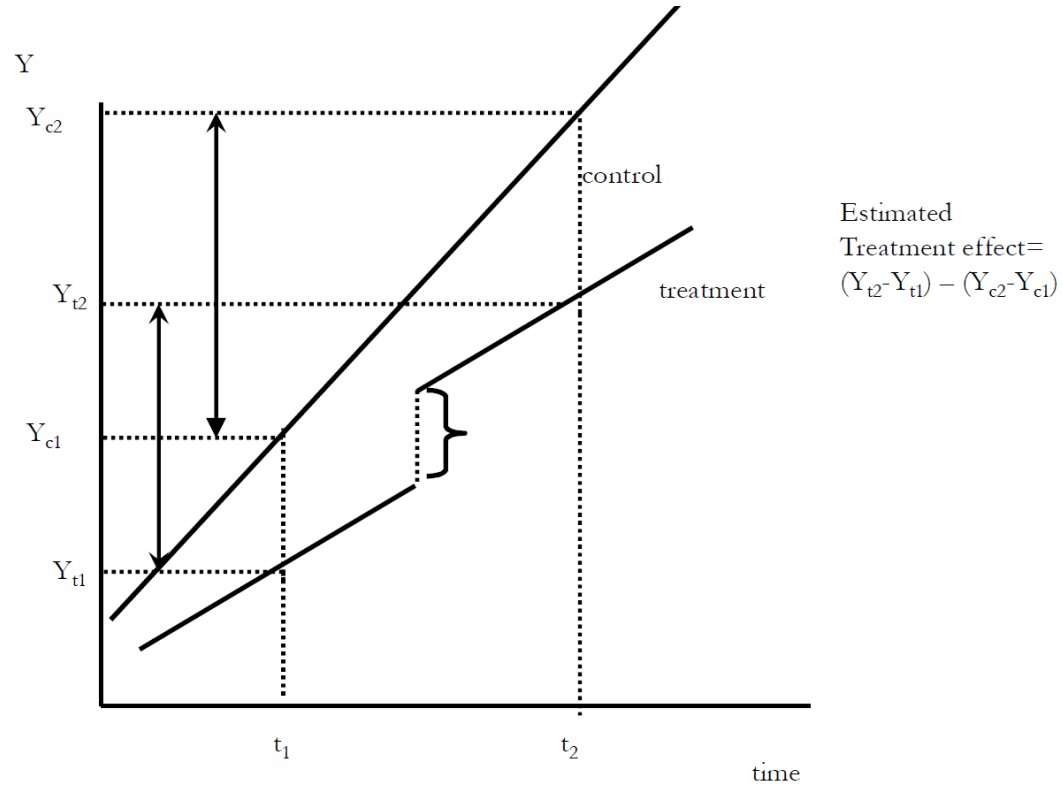
Parallel trends assumption

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



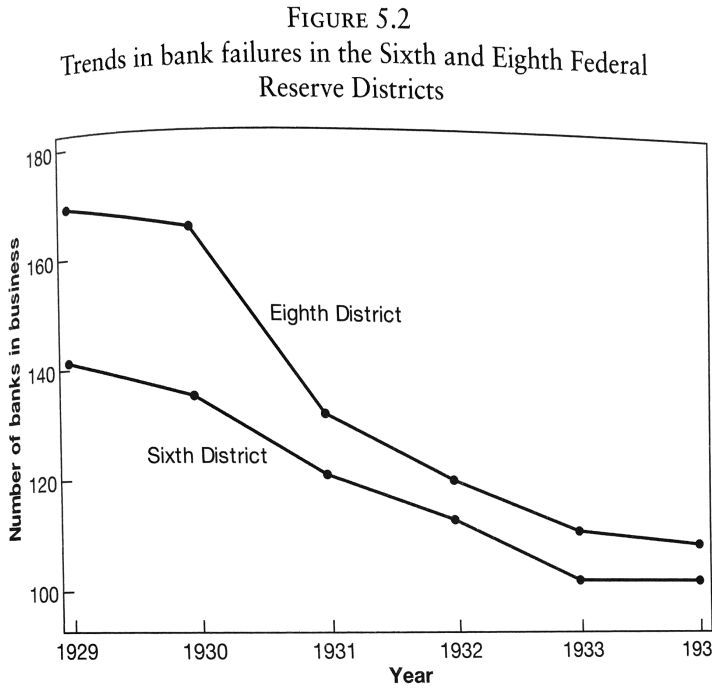
Parallel trends assumption

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

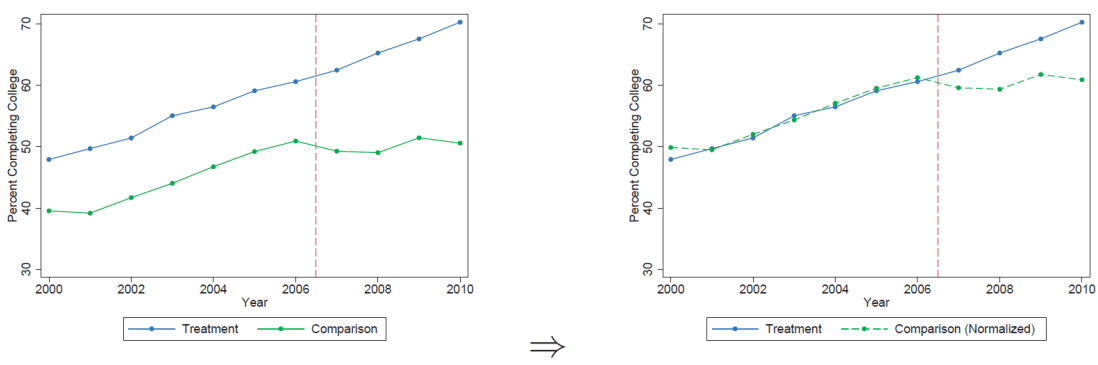
- A compelling graph: pointing to similar trends prior to the treatment. Note: common trends prior to treatment are neither necessary nor sufficient for parallel trends assumption!
- Event study regression and graph
- A placebo / falsification test
- Controlling for time trends directly (leans heavily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

Federal Reserve policy and bank failures

From Mastering 'Metrics chapter 5—treatment in 1930.

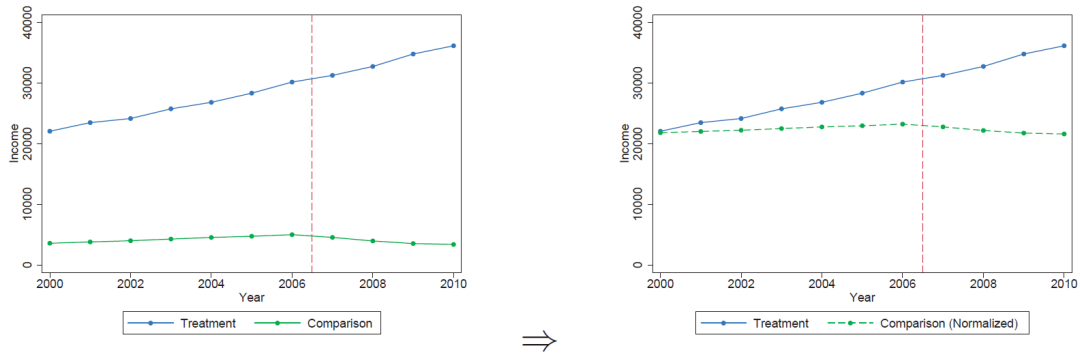


Checking the parallel trends assumption (1)



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

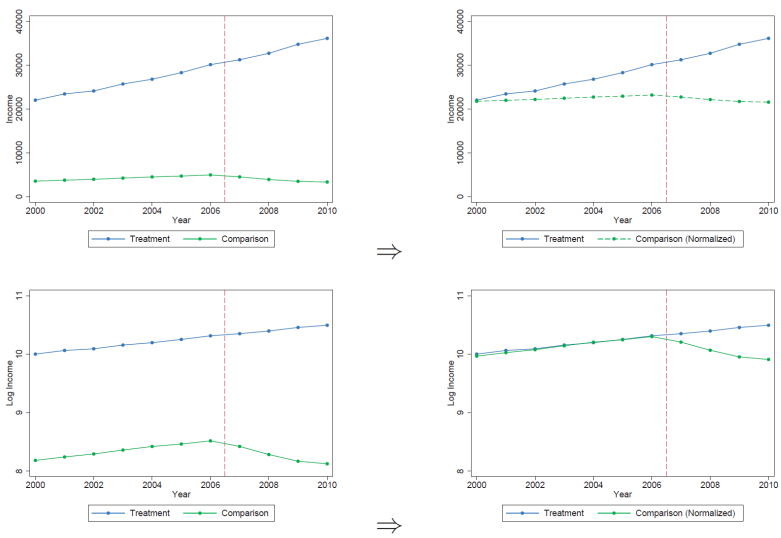
Checking the parallel trends assumption (2)



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

Checking the parallel trends assumption (3)

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



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