Difference-in-Differences

Edicson Luna

University of Pennsylvania

Motivation

The DiD estimator

The key assumptions

Regression form and inference

Motivation

The comparison we wish we had

To measure a causal effect we would love to see the *same* unit both with and without the treatment. That counterfactual is never available.

So we look for a control group to stand in for the treated group's counterfactual. The problem, as always, is that the groups we can compare are usually not comparable: they differ for reasons that also affect the outcome.

Difference-in-differences (DiD) is a clever way to make progress even when the two groups start out different.

Two imperfect comparisons

Suppose a policy hits a treated group. We could try:

- **Before vs. after** (within the treated group).
  But other things change over time (the economy, the season). We would confound the policy with the *time trend*.

- **Treated vs. control** (at one point in time).
  But the two groups differ from the start. We would confound the policy with *fixed group differences*.

Each comparison alone is contaminated. The idea of DiD is to combine them so the contamination cancels.

The individual error component

Write the untreated potential outcome of unit i at time t as

$$Y_{it}(0) = \alpha_i + \lambda_t + u_{it}.$$

- α_i: a time-invariant component specific to the unit/group (its "type"). This is exactly what makes groups *not comparable*.
- λ_t: a common time effect (a shock hitting everyone in period t).
- u_{it}: an idiosyncratic error.

The troublemaker is α_i. If we could just *delete* it, comparability would stop being a problem.

The trick: difference out α_i

Take the change over time *within* a group. The stubborn α_i does not depend on t, so it cancels:

$$Y_{i,\text{post}}(0) - Y_{i,\text{pre}}(0) = \underbrace{(\lambda_{\text{post}} - \lambda_{\text{pre}})}_{\text{common time effect}} + (u_{i,\text{post}} - u_{i,\text{pre}}).$$

The individual component is gone! But a common time effect λ remains.

So we take *one more difference*, this time across groups: since λ_t is common to treated and control, it cancels too. That second difference is the whole idea.

The DiD estimator

The 2×2 setup

The simplest DiD has two groups and two periods:

	Before	**After**
Treated	$\bar{Y}_{T,\text{pre}}$	$\bar{Y}_{T,\text{post}}$
Control	$\bar{Y}_{C,\text{pre}}$	$\bar{Y}_{C,\text{post}}$

Only the treated group is exposed to the policy, and only in the "after" period. The control group is never treated but lives through the same time period.

The double difference

The DiD estimator is the difference of the two "before→after" changes:

$$\hat{\delta} = \left(\bar{Y}_{T,\text{post}} - \bar{Y}_{T,\text{pre}}\right) - \left(\bar{Y}_{C,\text{post}} - \bar{Y}_{C,\text{pre}}\right).$$

Why does this work? Plug in $Y(0) = \alpha + \lambda + u$ for the control and for the treated counterfactual:

- The first difference kills each group's α.
- The second difference kills the common λ.

What survives is the treatment effect on the treated group. The control group's change is our estimate of the time effect, which we subtract from the treated group's change.

An easy example: the minimum wage

Card and Krueger (1994) studied a classic question: does raising the minimum wage destroy jobs?

- In April 1992, New Jersey raised its minimum wage from \$4.25 to \$5.05. (*treated*)
- Neighboring Pennsylvania left it at \$4.25. (*control*)

They measured employment at fast-food restaurants in both states, *before* and *after* the increase. PA plays the role of NJ's counterfactual: what NJ employment would have done absent the hike.

The example, in numbers

Full-time-equivalent employment per restaurant (approximate):

	Before	**After**	**Change**
NJ (treated)	20.4	21.0	$+0.6$
PA (control)	23.3	21.2	-2.1

$$\hat{\delta} = (+0.6) - (-2.1) = +2.7.$$

Employment in PA fell, but in NJ it did not. Relative to the control, the minimum-wage increase did *not* reduce employment — a famous and surprising result.

The key assumptions

Assumption 1: Parallel trends

DiD uses the control group's change as the treated group's counterfactual change. This is legitimate only if:

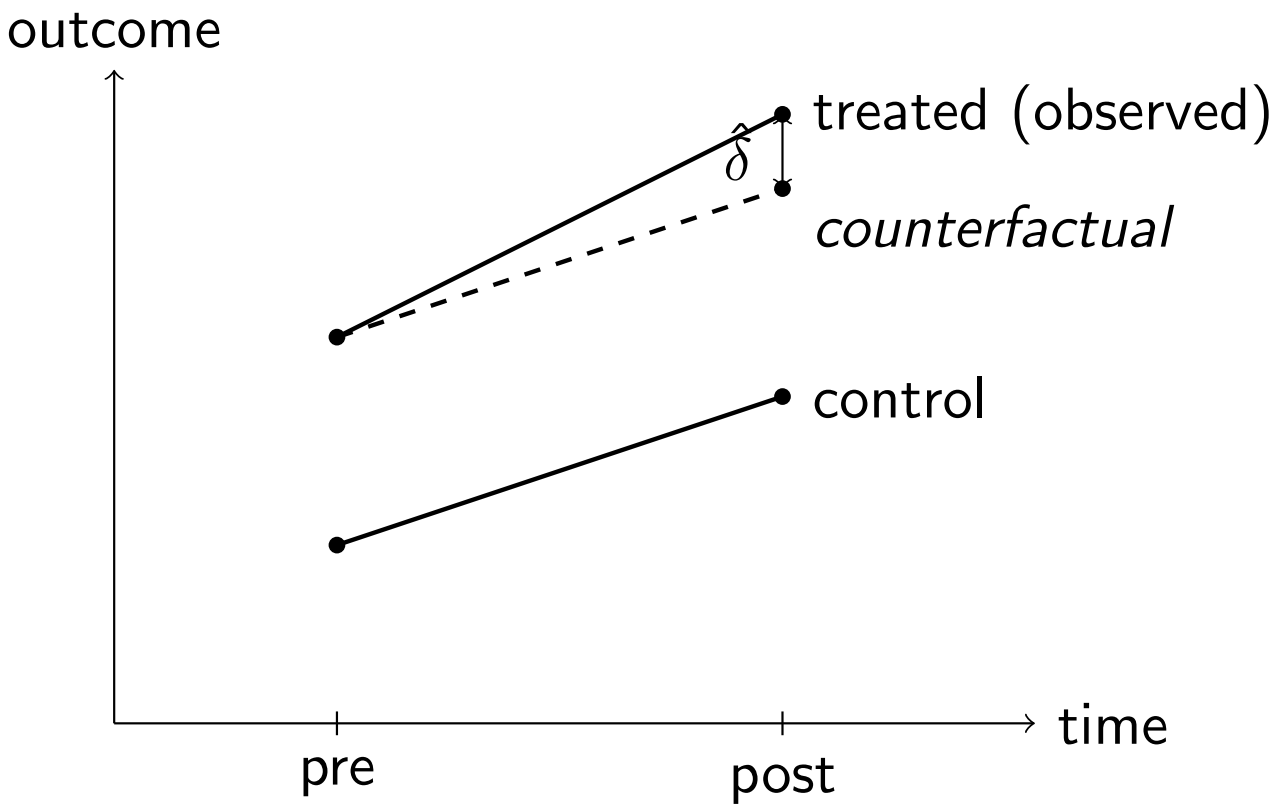
Absent the treatment, treated and control would have followed the same trend.

Formally, in terms of untreated potential outcomes,

$$\mathbb{E}[Y_{\text{post}}(0) - Y_{\text{pre}}(0) \mid \text{Treated}] = \mathbb{E}[Y_{\text{post}}(0) - Y_{\text{pre}}(0) \mid \text{Control}].$$

Groups may start at different *levels* (that is the α_i, which we difference out). What we require is a common *trend*.

Parallel trends, in a picture



The counterfactual (dashed) is parallel to the control. The gap between it and the observed treated line is the effect $\hat{\delta}$.

Parallel trends is about the counterfactual

The assumption concerns a line we never observe: how the treated group *would* have moved without treatment. So it can never be tested directly.

In practice we build confidence in it indirectly:

- Check **pre-trends**: if we have several periods before treatment, do the groups already move in parallel?
- Look for other control groups or a placebo period.

Parallel pre-trends are *evidence* for the assumption, not a proof of it.

Assumption 2: No anticipation

The second main assumption is **no anticipation**: units do not respond to the treatment *before* it actually arrives.

If NJ restaurants had started cutting staff in the "before" period because they *knew* the hike was coming, the pre-period would already be contaminated, and $\bar{Y}_{T,\text{pre}}$ would not be a clean baseline.

Together, parallel trends and no anticipation are what let the control group's change stand in for the treated group's missing counterfactual.

Some fine print

A few more conditions travel quietly with DiD:

- Common shocks / no other policy. No *other* event should hit one group and not the other at the same time (e.g. a local recession in PA only).
- **Stable composition.** With repeated cross-sections, the groups should not change who they are made of between periods.
- **No spillovers (SUTVA).** The control group is not indirectly affected by the treatment.

All of these are really versions of one idea: apart from the treatment, the two groups share the same time path.

Regression form and inference

DiD as a regression

The double difference is exactly the interaction coefficient in a simple regression. Let $T_i = 1$ for the treated group and $\text{Post}_t = 1$ in the after period:

$$Y_{it} = \beta_0 + \beta_1 T_i + \beta_2 \text{Post}_t + \delta (T_i \times \text{Post}_t) + \epsilon_{it}.$$

Reading off the four cells:

- β_1 captures the fixed group gap (the α).
- β_2 captures the common time change (the λ).
- δ is the DiD estimator: the extra movement in the treated group, after netting both out.

Why the regression form is useful

Writing DiD as OLS is not just convenient notation:

- We can add control variables to make parallel trends more plausible.
- We can allow many groups and many periods by replacing T_i and Post_t with group and time fixed effects (the "two-way fixed effects" estimator).
- We get standard errors for free from the OLS machinery.

And because it is just OLS, everything we proved about large-sample behavior applies directly.

Back to asymptotic theory

The DiD estimator $\hat{\delta}$ is an OLS coefficient, so the asymptotic results carry over. Under our assumptions,

$$\hat{\delta} \xrightarrow{p} \delta, \qquad \sqrt{n}\,(\hat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, V).$$

Consistency gives us the right target; asymptotic normality lets us build confidence intervals and test $H_0 : \delta = 0$.

But there is a subtlety in *how* we estimate V here that we must not ignore.

The inference caveat: clustering

In DiD we observe the same groups repeatedly over time. The errors u_{it} within a group are **serially correlated** — they are far from i.i.d.

Ignoring this makes the usual standard errors too small, so we reject H_0 far too often (Bertrand, Duflo and Mullainathan, 2004).

The fix is **clustered standard errors**, allowing arbitrary correlation within each group. The relevant asymptotics is then in the number of clusters (groups), not the number of observations — so DiD with very few groups is dangerous.

Takeaway

- When groups are not comparable, model the untreated outcome as $Y_{it}(0) = \alpha_i + \lambda_t + u_{it}$.
- Differencing over time removes the unit component α_i; differencing across groups removes the common time effect λ_t. What is left is the DiD estimator.
- It is valid under parallel trends and no anticipation (plus common shocks, stable composition, no spillovers).
- $\hat{\delta}$ is just an OLS interaction coefficient: consistent and asymptotically normal — but use clustered standard errors, since observations repeat over time.