Difference-in-differences

Bas Machielsen

2025-06-10

The Difference-in-Differences Estimator

Foundations of Causal Inference & Basic DiD Setup

Recap: Correlation vs. Causation

The fundamental challenge in empirical work.

- **Correlation:** Two variables move together.
  - *Example:* Ice cream sales are positively correlated with crime rates.
- **Causation:** A change in one variable *causes* a change in another.
  - *Does eating ice cream cause crime?* Unlikely.
- **Confounding Variable:** A third variable affects both.
  - *Hot weather* increases both ice cream sales and the number of people outside (leading to more opportunities for crime).

Our goal is to isolate the causal effect, not just the correlation.

The Potential Outcomes Framework

Also known as the **Rubin Causal Model**.

Let's think about the effect of a treatment (e.g., a job training program) on an individual i.

- $Y_i(1)$: The potential outcome for unit i **if they receive the treatment**.
  - *Example: Person i's earnings if they attend the program.*
- $Y_i(0)$: The potential outcome for unit i **if they do NOT receive the treatment**.
  - *Example: Person i's earnings if they do not attend the program.*

The Individual Causal Effect

For any single individual *i*, the true causal effect of the treatment is the difference between their two potential outcomes:

$$\tau_i = Y_i(1) - Y_i(0)$$

- This is the pure, unadulterated effect of the treatment on that one person.
- *Example:* The increase in Person *i*'s earnings caused *only* by the training program.

The Average Treatment Effect (ATE)

Since we usually can't measure the effect for every single individual, we focus on averages.

The **Average Treatment Effect (ATE)** is the average of the individual causal effects over the entire population.

$$ATE = E[\tau_i] = E[Y(1) - Y(0)]$$

This tells us, "On average, what is the effect of this treatment for a person randomly drawn from the population?"

The Fundamental Problem of Causal Inference

This is the core challenge that all causal methods try to solve.

For any given unit *i*, we can only ever observe *one* of their potential outcomes.

- If person i takes the training program, we see $Y_i(1)$. We will never know what their earnings would have been without it, $Y_i(0)$.
- If person i does not take the program, we see $Y_i(0)$. We will never know $Y_i(1)$.

Causal inference is a **missing data problem**. The $Y_i(0)$ for the treated and the $Y_i(1)$ for the untreated are called **counterfactuals**.

Illustrating the Fundamental Problem

Unit (i)	Attends Program?	Observed Earnings	$Y_i(1)$	$Y_i(0)$
Alice	Yes (T=1)	$50,000	$50,000	**???**
Bob	No (T=0)	$40,000	**???**	$40,000
Carol	Yes (T=1)	$45,000	$45,000	**???**
David	No (T=0)	$60,000	**???**	$60,000

We can't calculate $Y_i(1) - Y_i(0)$ for anyone!

Why Simple Comparisons Fail

A naive approach might be to just compare the average earnings of those who attended the program to those who didn't.

$$\text{Naive Comparison} = E[Y|T=1] - E[Y|T=0]$$

This is almost always **wrong**. Why?

Because the people who *choose* to get treatment might be different from those who don't in ways that also affect the outcome. This is **Selection Bias**.

Selection Bias: The Hidden Difference

The simple difference can be decomposed:

$$
\begin{aligned}
E[Y|T=1]-E[Y|T=0] &= E[Y(1)|T=1] - E[Y(0)|T=0] \\
&= E[Y(1)|T=1] - E[Y(0)|T=1] \\
&\quad + E[Y(0)|T=1] - E[Y(0)|T=0] \\
\\
&= \underbrace{(E[Y(1)|T=1] - E[Y(0)|T=1])}_{\text{Average Treatment Effect on the Treated (ATT)}} \\
&\quad + \underbrace{(E[Y(0)|T=1] - E[Y(0)|T=0])}_{\text{Selection Bias}}
\end{aligned}
$$

Selection Bias: Decomposition

Hence

$$E[Y|T=1]-E[Y|T=0]=ATE+\text{Selection Bias}$$

Where: Selection Bias $= E[Y(0)|T=1]-E[Y(0)|T=0]$

- In words: Selection bias is the difference in the **no-treatment outcome** between the treated and untreated groups.
- *Job Program Example:* People who sign up for training ($T=1$) might be more motivated. Even without the program, their earnings `Y(0)` might have been higher than the less motivated group ($T=0$).

Introduction to Differences-in-Differences (DiD)

The Core Idea: Use data from a **pre-treatment period** to account for selection bias.

- We assume that the "selection bias" (the difference between the groups) is constant over time.
- We compare the *change* in the outcome over time for the treatment group to the *change* over time for a control group.
- The "difference in the differences" isolates the treatment effect.

The 2x2 DiD Setup

The classic setup involves two groups and two time periods.

	Before Period (Pre)	**After Period (Post)**
Treatment Group	$\hat{Y}_{T,Pre}$	$\hat{Y}_{T,Post}$
Control Group	$\hat{Y}_{C,Pre}$	$\hat{Y}_{C,Post}$

- **Treatment Group:** A group that is exposed to the policy/treatment in the "After" period.
- **Control Group:** A similar group that is *not* exposed to the treatment in either period.

Calculating the Simple DiD Estimator

We calculate two differences, then take the difference between them.

1. **First Difference (Treatment Group):** The change over time for the treated. $\Delta_T = \hat{Y}_{T,Post} - \hat{Y}_{T,Pre}$
2. **First Difference (Control Group):** The change over time for the controls. This represents the "secular trend" – what would have happened without the treatment. $\Delta_C = \hat{Y}_{C,Post} - \hat{Y}_{C,Pre}$
3. **The Difference-in-Differences:** $\tau_{DiD} = \Delta_T - \Delta_C$

DiD Under Potential Outcomes

The observed outcome, Y_{it}, is determined by the unit's group status and the time period.

- For the **treated group $(D_i = 1)$**: They are untreated at $t = 0$ and treated at $t = 1$.
  - $Y_{i0} = Y_{i0}(0)$ for $t = 0$ (pre-treatment)
  - $Y_{i1} = Y_{i1}(1)$ for $t = 1$ (post-treatment)
- For the **control group $(D_i = 0)$**: They are never treated.
  - $Y_{i0} = Y_{i0}(0)$ for $t = 0$
  - $Y_{i1} = Y_{i1}(0)$ for $t = 1$

The Estimand: The ATT

Formally, the ATT is the difference between the treated group's outcome at $t = 1$ and what their outcome *would have been* at $t = 1$ if they had not been treated.

$$\text{ATT} = E[Y_{i1}(1)|D_i = 1] - E[Y_{i1}(0)|D_i = 1] \tag{1}$$

The first term, $E[Y_{i1}(1)|D_i = 1]$, is observed as the average outcome for the treated group in the post-period, $E[Y_{i1}|D_i = 1]$.

The second term, $E[Y_{i1}(0)|D_i = 1]$, is the **counterfactual**. It is the unobservable average outcome for the treated group had they not received the treatment. The entire goal of the DiD strategy is to find a way to estimate this term.

Parallel Trends

The DiD estimator is valid under the **Parallel Trends Assumption**. This assumption states that, in the absence of treatment, the average outcome for the treated group would have changed over time by the same amount as the average outcome for the control group.

Mathematically, this is expressed in terms of the potential outcome under no-treatment, $Y_{it}(0)$:

$$E[Y_{i1}(0) - Y_{i0}(0)|D_i = 1] = E[Y_{i1}(0) - Y_{i0}(0)|D_i = 0] \tag{2}$$

The left side is the counterfactual change for the treated group. The right side is the observed change for the control group, since for them $Y_{it} = Y_{it}(0)$. This assumption allows us to use the control group to identify the counterfactual trend for the treated group.

Derivation of the DiD Estimator

First, we rearrange the parallel trends assumption (2) to solve for our unobserved counterfactual:

$$E[Y_{i1}(0)|D_i = 1] = \underbrace{E[Y_{i0}(0)|D_i = 1]}_{\text{Treated pre-treatment}} + \underbrace{(E[Y_{i1}(0)|D_i = 0] - E[Y_{i0}(0)|D_i = 0])}_{\text{Change in control group}} \tag{3}$$

This equation shows how we construct the counterfactual: we take the treated group's initial level and add the change experienced by the control group.

Derivation of the DiD Estimator

Substitute this expression for the counterfactual (3) back into the definition of ATT:

$$\begin{aligned}\text{ATT} &= E[Y_{i1}(1)|D_i = 1] - E[Y_{i1}(0)|D_i = 1] \\ &= E[Y_{i1}(1)|D_i = 1] - \\ &(E[Y_{i0}(0)|D_i = 1] + E[Y_{i1}(0)|D_i = 0] - E[Y_{i0}(0)|D_i = 0])\end{aligned}$$

Replacing Potential Outcomes with Observables

Finally, we replace the potential outcomes with their observable counterparts:

- $E[Y_{i1}(1)|D_i = 1] = E[Y_{i1}|D_i = 1]$
- $E[Y_{i0}(0)|D_i = 1] = E[Y_{i0}|D_i = 1]$
- $E[Y_{i1}(0)|D_i = 0] = E[Y_{i1}|D_i = 0]$
- $E[Y_{i0}(0)|D_i = 0] = E[Y_{i0}|D_i = 0]$

Identification of the Estimand

Substituting these into the equation gives the DiD estimator:

$$\begin{aligned}\text{ATT} &= E[Y_{i1}|D_i = 1] - \\ &\quad (E[Y_{i0}|D_i = 1] + E[Y_{i1}|D_i = 0] - E[Y_{i0}|D_i = 0]) \\ &= (E[Y_{i1}|D_i = 1] - E[Y_{i0}|D_i = 1]) - \\ &\quad (E[Y_{i1}|D_i = 0] - E[Y_{i0}|D_i = 0])\end{aligned}$$

This is the famous “difference-in-differences” formula. It identifies the ATT under the crucial assumption of parallel trends.

Graphical Illustration of DiD

- Tbd: Graphical Example

Deconstructing the DiD Graph

- The **solid blue line** is the observed trend for the control group.
- The **solid red line** is the observed outcome for the treatment group.
- The **dotted red line** is the *counterfactual* for the treatment group, constructed by assuming its trend would have been parallel to the control group's trend.
- The **DiD effect** is the vertical distance between the actual outcome for the treatment group and its counterfactual outcome in the post-period.

DiD using a Regression Framework

We can estimate the exact same 2x2 DiD using a simple OLS regression. This is more powerful and flexible.

$$Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Post_t + \beta_3 (Treat_i \text{Ö} Post_t) + \epsilon_{it}$$

- Y_{it}: Outcome for unit i at time t.
- $Treat_i$: A dummy variable $= 1$ if unit i is in the treatment group, 0 otherwise.
- $Post_t$: A dummy variable $= 1$ if the period is "Post", 0 otherwise.
- $Treat_i \times Post_t$: An interaction term.

Interpretation of Coefficients (1/2)

Let's break down what each `\beta` represents:

$$Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Post_t + \beta_3 (Treat_i \times Post_t) + \epsilon_{it}$$

- β_0 **(Intercept):** The average outcome for the **Control Group** (`Treat=0`) in the **Pre-Period** (`Post=0`).
  - $E[Y|Treat = 0, Post = 0] = \beta_0$
- β_1**:** The average *pre-existing difference* between the treatment and control groups in the pre-period. **This is the selection bias.**
  - $E[Y|Treat = 1, Post = 0] = \beta_0 + \beta_1$
  - So, $\beta_1 = E[Y|Treat = 1, Post = 0] - E[Y|Treat = 0, Post = 0]$

Interpretation of Coefficients (2/2)

$$Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Post_t + \beta_3 (Treat_i \text{Ö} Post_t) + \epsilon_{it}$$

- β_2**:** The average change in the outcome for the **Control Group** from the pre- to the post-period. **This is the secular trend.**
  - $E[Y|Treat = 0, Post = 1] = \beta_0 + \beta_2$
  - So, $\beta_2 = E[Y|Treat = 0, Post = 1] - E[Y|Treat = 0, Post = 0]$
- β_3 **(The Interaction Term): This is the DiD estimator!** It's the additional change in the outcome for the Treatment Group, above and beyond the secular trend.
  - It is the causal effect of interest.
  - $\beta_3 = (E[Y|T = 1, P = 1] - E[Y|T = 1, P = 0]) - (E[Y|T = 0, P = 1] - E[Y|T = 0, P = 0])$

Advantages of the Regression Framework

Why bother with regression instead of just calculating the four means?

1. **Standard Errors:** Regression automatically provides standard errors, t-statistics, and p-values for your DiD estimate (β_3), allowing for statistical inference.
2. **Adding Covariates:** It is easy to add control variables to the model to increase precision and make the parallel trends assumption more plausible.
3. **Flexibility:** The framework is easily extended to more complex scenarios (more groups, more time periods, etc.).

Adding Covariates to the DiD Model

We can add a vector of control variables, `X`, to the regression.

$$Y_{it} = \beta_0 + \beta_1 Treat_i + \beta_2 Post_t + \beta_3 (Treat_i \text{Ö} Post_t) + \gamma' X_{it} + \epsilon_{it}$$

- **Purpose:** To control for observable characteristics that might differ between the groups and affect trends in the outcome.
- This helps strengthen the parallel trends assumption. It becomes "parallel trends *conditional on X*".
- *Example:* When studying a state-level policy, you might control for state GDP, population size, etc.

Testing the Parallel Trends Assumption

We can't prove the assumption, but we can build evidence for it. This requires data from **multiple pre-treatment periods**.

- **Method 1: Visual Inspection (Most Common)**
    - Plot the average outcomes for the treatment and control groups over time.
    - Visually check if their trends were parallel in the periods *before* the treatment was introduced.

Statistical "Tests" for Parallel Trends

If you have multiple pre-treatment periods, you can run a "placebo" test.

- **Idea:** Run a DiD analysis using only pre-treatment data. For instance, define a fake "treatment" in period t-2 and use t-3 as the "pre" period.
- **In a regression:** Interact the treatment group dummy with time-period dummies for *each pre-treatment period*. $Y_{it} = \cdots + \delta_{-2}(Treat \times PrePeriod_{-2}) + \delta_{-1}(Treat \times PrePeriod_{-1}) + \beta_3(Treat \times Post) + \epsilon_{it}$
- **Result:** The coefficients δ_{-2} and δ_{-1} should be small and statistically insignificant (not different from zero). This shows there was no pre-existing differential trend.
- Could also test this using a model like $Y_{it} = \alpha_t + \sum \delta_k(Treat \times 1(t = k)) + \beta_1 Treat$ for $k \in \{-T, \ldots, T\}$ and test the coefficients on $k = \{-T, \ldots, 0\}$

Extension: Multiple Periods & Staggered Adoption

The real world is often messier than 2x2.

- **Staggered Adoption:** Different units receive the treatment at different times.
    - *Example:* State A adopts a policy in 2010, State B in 2012, State C never does.
- The simple 2x2 Treat × Post model is no longer sufficient and can be biased.
- Modern methods (e.g., Callaway & Sant'Anna, Sun & Abraham) address these issues by using better defined control groups for each treated unit.

Visualizing Staggered DiD: Event Study Plots

These plots are standard for visualizing results from models with multiple periods.

- **X-axis:** Time relative to the treatment event (e.g., -3, -2, -1, 0, +1, +2 years).
- **Y-axis:** The estimated DiD coefficient for that relative time period.
- **Interpretation:**
  - Coefficients for **pre-periods** ($t < 0$) should be near zero (validates parallel trends).
  - Coefficients for **post-periods** ($t \geq 0$) show the dynamic causal effect of the policy over time.

Potential Pitfalls (1/3): Ashenfelter's Dip

A famous issue where the parallel trends assumption is violated in a specific way.

- **The Dip:** A notable drop in the outcome for the treatment group *just before* treatment.
- *Classic Example:* Individuals' earnings often drop right before they enter a job training program (e.g., due to job loss).
- This makes it look like the program had a huge effect, but it's really just a recovery to a normal level.

Potential Pitfalls (2/3): Policy Anticipation & Spillovers

- **Anticipation Effects:** If people know a policy is coming, they may change their behavior *before* it's officially implemented. This contaminates the "pre" period and violates parallel trends.
    - *Example:* A firm hires fewer people in anticipation of a minimum wage hike.
- **Spillover Effects:** The treatment "spills over" and affects the control group.
    - *Example:* A job fair in one city (treatment) draws workers from a neighboring city (control), affecting the control city's labor market. Your control group is no longer a valid counterfactual.

Potential Pitfalls (3/3): Other Limitations

- **Functional Form:** The basic DiD model assumes the treatment effect is a constant, additive shift.
- **Data Requirements:** Requires panel data (tracking the same units over time) or repeated cross-sectional data.
- **Bad Control Group:** The entire method relies on finding a credible control group that satisfies the parallel trends assumption. This is often the hardest part.

Summary: DiD is Powerful, but Use with Care

- **What it does:** DiD provides a powerful and intuitive way to estimate causal effects by controlling for time-invariant unobserved differences (selection bias).
- **The Golden Rule:** The **Parallel Trends** assumption is everything. You must convince yourself and your audience that it is plausible.
- **Best Practices:**
  - Use a regression framework for S.E.s and covariates.
  - Always visually inspect pre-treatment trends.
  - Run statistical pre-trend tests if you have the data.
  - Be aware of pitfalls like anticipation, spillovers, and Ashenfelter's dip.

Staggered Adoption

Many modern DiD settings involve **staggered treatment adoption**, where different units get treated at different times.

The standard tool for this has been the two-way fixed effects (TWFE) regression:

$$Y_{it} = \alpha_i + \lambda_t + \beta^{TWFE} \cdot D_{it} + \epsilon_{it}$$

where $D_{it} = 1$ if unit i is treated at time t.

Recent research (Goodman-Bacon, 2021; de Chaisemartin & D'Haultfœuille, 2020) shows that $\hat{\beta}^{TWFE}$ is a weighted average of all possible 2x2 DiD estimators.

Crucially, some of these comparisons are "bad": they use already-treated units as controls for later-treated units.

Why Bad Comparisons Are A Problem?

Imagine Cohort 2010 gets treated in 2010 and Cohort 2012 gets treated in 2012.

- To estimate the effect on Cohort 2012 in the year 2012, TWFE implicitly uses Cohort 2010 as part of the "control" group.
- But Cohort 2010 has already been treated for two years!

This is only valid if treatment effects are constant across cohorts and time. If effects are **heterogeneous** or **dynamic** (e.g., they grow over time), this comparison is contaminated.

- The resulting $\hat{\beta}^{TWFE}$ can be a meaningless average, sometimes with negative weights, and may not represent any true ATT.

Sun and Abraham (2021) Solution

The core idea is to avoid aggregation and bad comparisons.

Step 1: Define Cohorts

- A cohort, $G = g$, is the group of all units that are first treated at the same time period g.
- Let $G = \infty$ (or $G = C$) be the **never-treated** group.

Step 2: Estimate Cohort-Specific ATTs

- Instead of one β, estimate a separate effect for *each cohort* g at *each time period* ℓ.
- Use only **clean controls**: units that are not yet treated. The never-treated group $(G = C)$ is the cleanest control.

Identification of Cohort-Specific ATT

The estimand of interest is the Average Treatment Effect for cohort g at calendar time ℓ (where $\ell \geq g$). We denote this $\text{ATT}(g, \ell)$.

For each pair (g, ℓ), $\text{ATT}(g, \ell)$ is identified by a simple 2x2 DiD comparing cohort g to the clean control group (C) between the pre-treatment period $(g-1)$ and period ℓ:

$$\begin{aligned}\widehat{\text{ATT}}(g, \ell) = &\Big(E[Y_\ell | G = g] - E[Y_{g-1} | G = g]\Big) \\ &- \Big(E[Y_\ell | G = C] - E[Y_{g-1} | G = C]\Big)\end{aligned}$$

Key Assumption: A cohort-specific parallel trends assumption. In the absence of treatment, the outcome for cohort g would have evolved in parallel to the outcome for the never-treated (or not-yet-treated) group.

Bias-variance trade-off

The S&A framework is flexible. You can choose your control group to manage the bias-variance trade-off:

- **Option A: Never-Treated Controls.** This is the cleanest option, requiring the weakest parallel trends assumption (only vs. never-treated). It is preferred if the group is large enough.
- **Option B: Not-Yet-Treated Controls.** The estimator can use all currently untreated units as a time-varying control group. This **substantially increases statistical power** and reduces variance. The cost is a slightly stronger (but still plausible) parallel trends assumption against these not-yet-treated units.

Questions?