Causality, OVB & Differences-in-Differences

ECON0004: Applied Econometrics

UCL

University College London

March 2026

Based on: Lecture notes (Lecture 5); Angrist & Pischke (2009) *Mostly Harmless Econometrics*.

Outline

Does University Cause Higher Wages?

We observe that university graduates earn higher wages. But is this **correlation** or **causation**?

What we want to know:

- Would a specific person's wage change if we *intervened* and sent them to university — holding everything else fixed?
- This is a **causal** question, not a descriptive one.

Difference matters for policy: If the wage premium reflects ability sorting rather than genuine human-capital accumulation, subsidising university places will not raise national productivity.

Socratic question: Why can't we simply compare average wages of graduates and non-graduates to get the causal effect?

Potential Outcomes Framework

Let $D_i \in \{0, 1\}$ be a **treatment indicator**: $D_i = 1$ if individual i attended university, $D_i = 0$ otherwise.

Potential Outcomes

- W_{1i} — wage individual i *would earn* if they attended university
- W_{0i} — wage individual i *would earn* if they did not attend university
- **Individual treatment effect:** $W_{1i} - W_{0i}$

We observe only $W_i = D_i W_{1i} + (1 - D_i) W_{0i}$.

The fundamental problem of causal inference: We can never observe *both* W_{1i} and W_{0i} for the same person. The unobserved outcome is called the **counterfactual**.

Naive Comparison and Selection Bias

The observed wage difference between graduates and non-graduates:

$$\mathbb{E}[W_i \mid D_i = 1] - \mathbb{E}[W_i \mid D_i = 0] = \mathbb{E}[W_{1i} \mid D_i = 1] - \mathbb{E}[W_{0i} \mid D_i = 0]$$

Step 1. **Add and subtract** $\mathbb{E}[W_{0i} \mid D_i = 1]$:

$$= \mathbb{E}[W_{1i} \mid D_i = 1] - \mathbb{E}[W_{0i} \mid D_i = 1] + \mathbb{E}[W_{0i} \mid D_i = 1] - \mathbb{E}[W_{0i} \mid D_i = 0]$$

Step 2. **Regroup** the two pairs:

$$= \underbrace{\mathbb{E}[W_{1i} - W_{0i} \mid D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[W_{0i} \mid D_i = 1] - \mathbb{E}[W_{0i} \mid D_i = 0]}_{\text{Selection bias}}$$

The naive comparison equals ATT *only* if $\mathbb{E}[W_{0i} \mid D_i = 1] = \mathbb{E}[W_{0i} \mid D_i = 0]$ — i.e. graduates and non-graduates would have earned the **same wages** had none of them attended university. This is almost certainly violated.

The Short Regression

Suppose the **true model** is:

$$W_i = \beta_0 + \beta_1 D_i + \beta_2 A_i + u_i$$

where A_i is individual ability (unobserved), and $\mathbb{E}(u_i \mid D_i, A_i) = 0$.

If we cannot measure A_i, we estimate the **short regression**:

$$W_i = \tilde{\beta}_0 + \tilde{\beta}_1 D_i + \tilde{u}_i$$

Question: What does $\hat{\tilde{\beta}}_1$ estimate?

Hint: The short regression lumps $\beta_2 A_i$ into the error term. If A_i is correlated with D_i, MLR.4 fails in the short regression.

OVB Derivation

True model: $W_i = \beta_0 + \beta_1 D_i + \beta_2 A_i + u_i$. The OLS estimator in the short regression (omitting A_i) is: $\hat{\hat{\beta}}_1 = \frac{\sum_i (D_i - \bar{D})(W_i - \bar{W})}{\sum_i (D_i - \bar{D})^2}$.

Step 1. Substitute the true model: the numerator picks up $\beta_2 \sum_i (D_i - \bar{D})(A_i - \bar{A})$, so

$$\hat{\hat{\beta}}_1 = \beta_1 + \beta_2 \underbrace{\frac{\sum_i (D_i - \bar{D})(A_i - \bar{A})}{\sum_i (D_i - \bar{D})^2}}_{\hat{\delta}_{AD}} + \underbrace{\frac{\sum_i (D_i - \bar{D})\, u_i}{\sum_i (D_i - \bar{D})^2}}_{\to 0 \text{ in expectation}}$$

Step 2. Taking expectations ($\mathbb{E}[u_i \mid D_i, A_i] = 0$): $\mathbb{E}[\hat{\hat{\beta}}_1] = \beta_1 + \beta_2 \cdot \delta_{AD}$, $\delta_{AD} = \frac{\text{Cov}(A_i, D_i)}{\text{Var}(D_i)}$.
For binary D_i: $\delta_{AD} = \mathbb{E}(A_i | D_i{=}1) - \mathbb{E}(A_i | D_i{=}0)$.

$$\text{OVB} = \beta_2 \times \left[\mathbb{E}(A_i \mid D_i = 1) - \mathbb{E}(A_i \mid D_i = 0)\right]$$

Direction of Bias

$$\mathbb{E}[\tilde{\beta}_1] = \beta_1 + \underbrace{\beta_2}_{\text{sign?}} \times \underbrace{\left[\mathbb{E}(A_i \mid D_i = 1) - \mathbb{E}(A_i \mid D_i = 0)\right]}_{\text{sign?}}$$

In the university–wage example:

- $\beta_2 > 0$: ability raises wages (positive direct effect)
- $\mathbb{E}(A_i \mid D_i = 1) > \mathbb{E}(A_i \mid D_i = 0)$: university attenders are, on average, more able

$\Rightarrow$ OVB $> 0 \Rightarrow \mathbb{E}[\tilde{\beta}_1] > \beta_1$

> The short regression **overstates** the return to university. We attribute to education the wage premium that is really due to ability.

Bias Direction: Summary

The sign of OVB depends on the signs of both β_2 (direct effect of ability on wages) and the ability gap between groups:

	$\mathbb{E}(A_i \mid D_i = 1) > \mathbb{E}(A_i \mid D_i = 0)$	$\mathbb{E}(A_i \mid D_i = 1) < \mathbb{E}(A_i \mid D_i = 0)$
$\beta_2 > 0$	**Upward bias**	Downward bias
$\beta_2 < 0$	Downward bias	**Upward bias**

University example: $\beta_2 > 0$ and $\mathbb{E}(A_i \mid D_i = 1) > \mathbb{E}(A_i \mid D_i = 0) \Rightarrow$ **upward bias**. The observed wage premium overstates the true causal effect of university.

Suppose $\beta_1 = 0.06$ (6% true return), $\beta_2 = 0.04$, and the ability gap $\mathbb{E}(A_i|D_i{=}1) - \mathbb{E}(A_i|D_i{=}0) = 10$ points.

Step 1. **Plug into the OVB formula:** $\mathbb{E}[\hat{\tilde{\beta}}_1] = 0.06 + 0.04 \times 10 = \mathbf{0.46}$ (46% estimated return)

Step 2. **Interpretation:** The short regression overstates the return by $0.46 - 0.06 = 0.40$ — more than **seven times** the true causal effect. Almost all of the estimated premium reflects ability sorting.

Even a moderate ability gap combined with a small direct effect produces **massive** upward bias. This is why naive graduate–non-graduate wage comparisons are misleading for policy.

Solutions to OVB

Solution 1: Include all relevant controls

Add A_i (or proxies) directly to the regression.

- Works if A_i is measurable (e.g. IQ test scores, grades)
- Often impractical: many confounders are unobservable
- Even with proxies, residual OVB remains if measurement is imperfect

Solution 2: Find exogenous variation in the treatment

Find a source of variation in D_i that is **uncorrelated with A_i**:

- **Random assignment:** explicitly randomise D_i (RCT)
- **Quasi-experiments:** exploit "natural" randomisation in D_i (e.g. policy changes, geographic discontinuities)

OVB: Step-by-Step Proof

True model: $W_i = \beta_0 + \beta_1 D_i + \beta_2 A_i + u_i$

Step 1. Take conditional expectations for each treatment group:

$$\mathbb{E}(W_i \mid D_i = 1) = \beta_0 + \beta_1 + \beta_2\, \mathbb{E}(A_i \mid D_i = 1)$$

$$\mathbb{E}(W_i \mid D_i = 0) = \beta_0 + \beta_2\, \mathbb{E}(A_i \mid D_i = 0)$$

Step 2. Subtract to get what OLS estimates in the short regression:

$$\begin{aligned}\hat{\hat{\beta}}_1 &= \mathbb{E}(W_i \mid D_i = 1) - \mathbb{E}(W_i \mid D_i = 0) \\ &= \beta_1 + \beta_2 \left[\mathbb{E}(A_i \mid D_i = 1) - \mathbb{E}(A_i \mid D_i = 0)\right]\end{aligned}$$

Step 3. The second term is **omitted variable bias**; it vanishes only if $\mathbb{E}(A_i \mid D_i = 1) = \mathbb{E}(A_i \mid D_i = 0)$.

How Randomisation Eliminates OVB

Suppose D_i is assigned **randomly** (e.g. by lottery).

Random assignment ensures $D_i \perp A_i$, so:

$$\mathbb{E}(A_i \mid D_i = 1) = \mathbb{E}(A_i \mid D_i = 0)$$

Substituting into the OVB formula:

$$\mathbb{E}[\tilde{\beta}_1] = \beta_1 + \beta_2 \times \underbrace{[\mathbb{E}(A_i \mid D_i = 1) - \mathbb{E}(A_i \mid D_i = 0)]}_{=0} = \beta_1$$

> Random assignment **balances** the treatment and control groups on *all* characteristics — observed and unobserved. The short regression now recovers the true causal effect.

Case Study: Perry Preschool Programme

Setup:

- 1962–67, Ypsilanti, Michigan
- 123 disadvantaged African-American children aged 3–4
- Randomly assigned to high-quality preschool (treatment) or no preschool (control)
- Followed up at ages 27 and 40

Key findings (age 27):

- Treatment group: higher earnings, employment, graduation rates
- Lower crime rates, welfare dependency

Why trust this? Random assignment ensures that any difference in adult outcomes is **causally attributable** to the preschool programme — not to family background, ability, or neighbourhood effects.

Limits of Randomised Controlled Trials

RCTs are the **gold standard** for causal inference, but face real constraints:

- **Ethics:** It may be unethical to withhold beneficial treatments from a control group
- **Cost:** Long-horizon RCTs (tracking subjects for decades) are enormously expensive
- **Compliance:** Subjects may not follow their assigned treatment (the "intention-to-treat" vs "treatment on the treated" distinction)
- **External validity:** Results from a specific sample may not generalise to other populations or contexts
- **Feasibility:** We cannot randomise macroeconomic policies, legal regimes, or historical events

> When RCTs are infeasible, economists look for **quasi-experiments** — natural or policy-induced variations in treatment that approximate random assignment.

Quasi-Experiments

Quasi-Experiment

A research design that exploits *as-good-as-random* variation in treatment arising from natural events or policy changes, rather than deliberate randomisation.

Examples:

- A minimum wage law applies to one state but not a neighbouring state (Card & Krueger, *AER* 1994)
- Eligibility for a programme is determined by a date-of-birth cutoff
- A natural disaster affects some firms but not others

Key assumption: Unobserved differences between treated and untreated units are **stable over time**. If we can track outcomes before and after treatment, we can use the untreated units as a counterfactual.

Differences-in-Differences: Intuition

Setup:

- Two groups: **Treatment** (affected by policy) and **Control** (not affected)
- Two periods: **Pre** (before policy) and **Post** (after policy)

The naive approach: compare treatment group post vs. pre.

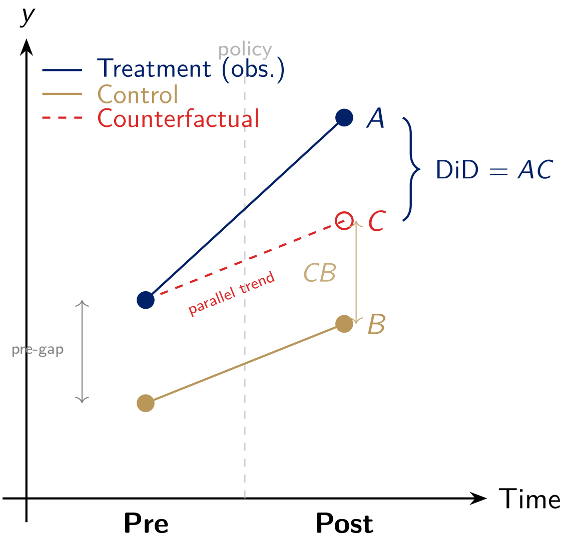
- Problem: outcomes may change for reasons *unrelated* to the policy (common trends, macroeconomic shocks)

The DiD fix:

1. Compute the post-minus-pre difference for the *treatment* group
2. Subtract the post-minus-pre difference for the *control* group

> DiD **differences out** any time trend common to both groups. What remains is attributed to the treatment.

DiD: Visual Illustration



Reading the DiD Diagram

Key points:

- A = Treatment post-period outcome
- B = Control post-period outcome
- C = **Counterfactual** (parallel-trend)
- Filled dots = observed pre-period outcomes

DiD estimator:

$$\hat{\delta} = AC = AB - CB$$

AB = naive post-difference; CB = counterfactual gap (pre-period gap under parallel trends).

Why not just use AB? The treatment group was *already* above the control group before the policy ($CB > 0$). DiD subtracts this pre-existing gap to isolate the **policy-induced** change.

The DiD 2×2 Table

Let μ_{dt} denote the average outcome for group $d \in \{0, 1\}$ in period $t \in \{0, 1\}$:

	Pre $(t = 0)$	**Post** $(t = 1)$	**Change**
Treatment $(D = 1)$	μ_{10}	μ_{11}	$\mu_{11} - \mu_{10}$
Control $(D = 0)$	μ_{00}	μ_{01}	$\mu_{01} - \mu_{00}$
Difference	$\mu_{10} - \mu_{00}$	$\mu_{11} - \mu_{01}$	

$$\hat{\delta}^{\text{DiD}} = (\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00}) = (\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00})$$

Both rearrangements give the same answer: post-difference minus pre-difference (or equivalently, treatment-change minus control-change).

DiD Assumptions

DiD produces an unbiased estimate of the treatment effect under three conditions:

SUTVA — Stable Unit Treatment Value Assumption

The potential outcomes for each unit are unaffected by the treatment status of other units (no spillovers). Treatment is consistently defined across units.

Common (Parallel) Trend

Absent the treatment, both groups would have followed the *same trend* over time: $\mathbb{E}[W_{0,\text{post}} - W_{0,\text{pre}} \mid D = 1] = \mathbb{E}[W_{0,\text{post}} - W_{0,\text{pre}} \mid D = 0]$.

No Compositional Change

The composition of the treatment and control groups does not change in response to the treatment (no selective sorting into groups).

DiD Assumptions: No Anticipation & Testing

No Anticipation

Units do *not* change behaviour *before* treatment. For all $t < t^*$: $\mathbb{E}[W_{0i,t} \mid D = 1] = \mathbb{E}[W_{0i,t} \mid D = 0]$. Anticipatory behaviour (e.g. pre-emptive hiring before a minimum wage rise) contaminates the pre-period baseline.

Testing the assumptions:

- **Pre-trends test:** Estimate DiD for each pre-treatment period; require $\hat{\beta}_3 \approx 0$ throughout.
- **Placebo in time:** Assign treatment one period earlier; a non-zero effect signals a pre-existing trend.
- **Placebo group:** Apply DiD to an unaffected group; a non-zero effect flags spurious common shocks.

Parallel trends is *untestable* post-treatment. Pre-trend tests are necessary evidence, not sufficient proof.

DiD as a Regression

Stack all observations (i = individual, t = period). We now write the treatment group indicator as T_i (the same role as D_i above; T for "treated group" is the standard DiD convention).

$$y_{it} = \beta_0 + \beta_1 \, T_i + \beta_2 \, \text{Post}_t + \beta_3 \, (T_i \times \text{Post}_t) + \epsilon_{it}$$

Variable	**Role**
$T_i \in \{0,1\}$	Treatment indicator (1 = treatment group)
$\text{Post}_t \in \{0,1\}$	Period indicator (1 = post-treatment period)
$T_i \times \text{Post}_t$	Interaction term
$\hat{\beta}_1$	Pre-existing level difference (selection)
$\hat{\beta}_2$	Common time trend
$\hat{\beta}_3$	**DiD estimator: the causal treatment effect**

Proof: $\hat{\beta}_3$ is the DiD Estimator

Step 1. Compute conditional means from $y_{it} = \beta_0 + \beta_1 T_i + \beta_2 \text{Post}_t + \beta_3 (T_i \times \text{Post}_t) + \epsilon_{it}$:

	Pre	**Post**
Treated	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
Control	β_0	$\beta_0 + \beta_2$

Step 2. **First differences:** Treated: $\beta_2 + \beta_3$; Control: β_2

Step 3. **DiD:** $(\beta_2 + \beta_3) - \beta_2 = \beta_3$. ✓

$\hat{\beta}_3 = (\mu_{11} - \mu_{10}) - (\mu_{01} - \mu_{00})$ — exactly the 2×2 DiD estimate.

Why Regression DiD Beats the 2×2 Table

The regression formulation offers important advantages over the simple table:

1. **Additional controls:** We can add covariates X_{it} to the regression to absorb residual confounders and improve precision:

$$y_{it} = \beta_0 + \beta_1 T_i + \beta_2 \text{Post}_t + \beta_3 (T_i \times \text{Post}_t) + \gamma' X_{it} + \epsilon_{it}$$

2. **Standard errors:** Regression gives $\text{se}(\hat{\beta}_3)$ directly, enabling hypothesis tests and confidence intervals for the treatment effect.
3. **Multiple periods:** Panel data with many time periods (two-way fixed effects) can be accommodated in the regression framework.

Socratic question: The parallel trend assumption is untestable for the post-period. How might you provide evidence for it using *pre-period* data?

Lecture 5 Summary

1. **Causal questions** require a counterfactual. Naive comparisons combine the causal effect with **selection bias**.
2. **OVB formula:** $\mathbb{E}[\tilde{\beta}_1] = \beta_1 + \beta_2\big[\mathbb{E}(A_i \mid D_i = 1) - \mathbb{E}(A_i \mid D_i = 0)\big]$. Bias direction depends on signs of β_2 and the ability gap.
3. **Random assignment** eliminates selection bias by balancing groups on all characteristics. RCTs are the gold standard but face ethical and practical limits.
4. **Quasi-experiments** exploit natural variation to approximate randomisation. **DiD** uses a control group to subtract time trends from the treatment group's change.
5. **DiD regression** ($T_i \in \{0,1\}$ = treatment group indicator): $y_{it} = \beta_0 + \beta_1 T_i + \beta_2 \text{Post}_t + \beta_3(T_i \times \text{Post}_t) + \epsilon_{it}$. $\hat{\beta}_3$ is the causal estimate under the parallel-trend assumption.