Unconfoundedness I: Assumptions and Matching

BUSS975: Causal Inference in Financial Research

Professor Ji-Woong Chung

Korea University

Outline

1. Introduction: what actually resolves bias
2. Confounders, covariates, and colliders
3. Identification assumptions
4. Subclassification
5. Exact matching
6. Nearest-neighbour matching
7. Inference with matching
8. Bias correction

Lecture 3b takes the same assumption further: collapse X to one number, weight instead of match, then compare with regression.

1. Introduction

Where we left off

- Lecture 2 gave a rule for **which** variables to condition on: block every backdoor path; condition on no descendant of D.
- It was silent on two things:
  - What if the variables that close the backdoors are **unobserved**? (Usually the case.)
  - Given a valid set, **how** do we actually condition — and does it matter?
- This lecture takes the optimistic branch: **suppose the confounders are observed**.
- Subclassification, matching, propensity scores and regression are all answers to "how."

Treatment assignment mechanisms resolve bias, not models

Recall Lecture 1's decomposition:

$$\underbrace{E[Y|D{=}1] - E[Y|D{=}0]}_{SDO} = ATE + \underbrace{E[Y^0|D{=}1] - E[Y^0|D{=}0]}_{\text{selection on levels}} + \underbrace{(1-\pi)(ATT - ATU)}_{\text{selection on gains}}$$

- What deletes the last two terms is a **claim about how treatment was assigned**.

What randomization actually did

- Under $(Y^1, Y^0) \perp\!\!\!\perp D$, we were permitted to *deduce*:

$$E[Y^0|D{=}1] = E[Y^0|D{=}0], \qquad E[Y^1|D{=}1] = E[Y^1|D{=}0]$$

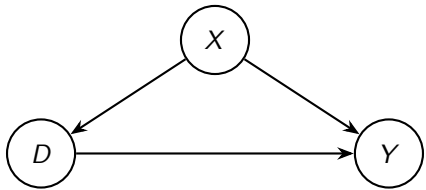
- The first kills selection on levels. Both together kill $ATT - ATU$. What is left is the ATE.
- These came from the **assignment mechanism**, not the estimator. We then *inserted* them into a statistical model.

The statistical model alone cannot resolve bias — the statistical model matched with the assignment mechanism can.

Why this framing matters

- Randomization is an **assignment mechanism**, not a statistical model. A regression is a **statistical model**, not an assignment mechanism.
- Unconfoundedness is this lecture's mechanism: treatment assigned **as if at random within strata of** X.
- A statistical model, credible *only* if you can defend that mechanism.
- Defend the mechanism first; choose the estimator second.

The bet we are making



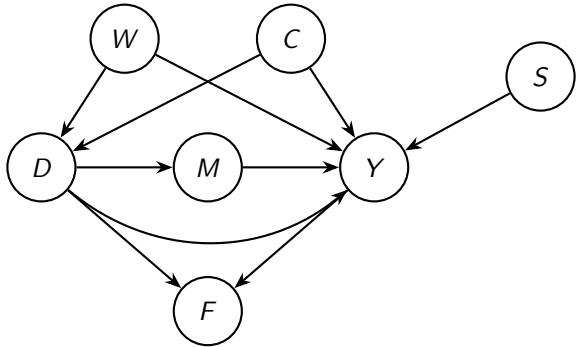
- Everything rests on one claim: **there is no dashed circle**. No unobserved U pointing at both D and Y.
- That claim is **not testable**.
- It is defended with institutional knowledge, not statistics.
- The literature calls it **unconfoundedness**, **conditional independence**, or **selection on observables**. Same thing.

2. Confounders, Covariates, and Colliders

A worked DAG: first class on the Titanic

- **Question**: did a first-class ticket (D) raise your chance of surviving (Y)?
- Raw comparison: 62.5% of first-class passengers survived versus 27.1% of everyone else — a gap of 35.4 points.
- But the *"women and children first"* rule was actually enforced, and first-class cabins were closer to the boat deck.
- **Sex** and **age** drive both who bought first class and who reached a lifeboat. Textbook confounding.

The Titanic DAG



D first class Y survival W sex C age M lifeboat F fame S swimming

- Backdoor paths: $D \leftarrow W \rightarrow Y$ and $D \leftarrow C \rightarrow Y$. Both **open**. Condition on $\{W, C\}$ and both close.
- F is a **collider**: $D \rightarrow F \leftarrow Y$. Already closed; conditioning on it *opens* it.
- M is a **mediator**: first class sat nearer the boat deck. It is *part of* the effect, not a confounder.

Four kinds of variable

Type	Example	What conditioning does
Confounder	W sex, C age	**Required.** Closes a backdoor path.
Collider	F fame	**Forbidden.** Opens a closed path.
Neutral / precision	S swimming	Optional. No bias either way; shrinks the standard error.
Mediator	M lifeboat	**Forbidden** for the total effect. **Allowed** for the partial effect.

- Only one of these four rows is about bias *reduction*. Two are about bias *creation*.
- S predicts the outcome but not treatment: free precision.
- M is subtle. Conditioning on it asks: *did first class help, holding fixed the main way it helped?*
- Press coverage *before* boarding would not be a collider.

The same DAG in corporate finance

Effect of a **bank credit line** (D) on **firm investment** (Y):

Titanic	Finance analogue	Rule
W, C sex, age	Firm size, industry, credit rating	Condition on
F fame	*Subsequent* analyst coverage	Never condition on
S swimming	Pre-period export-market demand	Optional, improves precision
M lifeboat	Amount actually drawn on the line	Depends
—	Unobserved managerial quality	The thing that sinks you

Covariate selection without a DAG

In practice you rarely have expert consensus on the full graph. Three working rules:

1. **Include variables you are confident are confounders.** Prior knowledge about the assignment mechanism is the input, not the data.
2. **Include pre-treatment variables that strongly predict the outcome.** These are confounders or precision-improvers like S, and *probably not colliders* — a variable measured before treatment cannot be caused by it.
3. **Exclude anything that might be an outcome.** If you are unsure, exclude everything measured after treatment.

- Rule 3 is most often violated.
- "Pre-treatment" enforces rule 2—a simple, verifiable timing check.

3. Identification Assumptions

Selecting the target parameter

Before estimating anything, decide *whose* effect you want.

$$ATE = E[Y^1 - Y^0], \quad ATT = E[Y^1 - Y^0 \mid D{=}1], \quad ATU = E[Y^1 - Y^0 \mid D{=}0]$$

- A polio vaccine. The three parameters answer three different policy questions.
  - ATE: what if we vaccinate **everyone**?
  - ATT: how much did the vaccine help **those who took it**?
  - ATU: what would happen if we reached **the currently unvaccinated**?
- They coincide only under constant effects. With heterogeneity they differ.
- **The estimand is a choice you make about the policy question.**

Which parameter does the question need?

Finance question	Parameter
Should the regulator extend this mandate to all firms?	*ATE*
Did the firms that adopted IFRS benefit from it?	*ATT*
Would forcing non-adopters to adopt help them?	*ATU*

- The third row is the *expansion* question — almost never the one your data answers well.
- Papers routinely estimate the *ATT* then draw conclusions requiring the *ATE* or *ATU*.

Unconfoundedness is a version of independence

- **Strong unconfoundedness:**

$$(Y^1, Y^0) \perp\!\!\!\perp D \mid X$$

- Compare Lecture 1's randomization assumption, $(Y^1, Y^0) \perp\!\!\!\perp D$, with no conditioning.
- Among units with the *same* X, treatment is as good as randomly assigned: **conditional randomization**.
- Stratum by stratum:

$$E[Y^1 \mid X{=}x] = E[Y^1 \mid D{=}1, X{=}x] = E[Y^1 \mid D{=}0, X{=}x]$$

$$E[Y^0 \mid X{=}x] = E[Y^0 \mid D{=}1, X{=}x] = E[Y^0 \mid D{=}0, X{=}x]$$

What the assumption buys: plug-in

- Y^1 and Y^0 are never both observed.
- Unconfoundedness lets us **replace a missing counterfactual with an observed outcome** from a comparable unit:

$$\underbrace{E[Y^0 \mid D{=}1, X{=}x]}_{\text{never observed}} = \underbrace{E[Y^0 \mid D{=}0, X{=}x]}_{\text{observed}} = \underbrace{E[Y \mid D{=}0, X{=}x]}_{\text{in your data}}$$

- The first equality is the **assumption**. The second is the switching equation.
- Every method in this lecture is a different way of executing that substitution.

Unconfoundedness is controversial

- Read literally: firms with identical observables **flip a coin** over a major financial decision.
- Every model of optimizing behaviour says otherwise: firms choose treatment *because* they expect to benefit — Lecture 1's selection on gains.
- This is the deepest objection, and it comes from economics, not statistics.
- Not a reason never to use these methods — a reason to state why *your* setting escapes it.

Guido W. Imbens's three defences

1. **Statistical motivation.** It is the natural starting point.
2. **It is an empirical question.** With enough institutional knowledge the relevant X may be measurable. Whether it is depends on the setting, not econometrics.
3. **Objectives differ.** The decision-maker optimizes something *other than your outcome*: a firm choosing a credit line optimizes liquidity and covenant slack, not the investment measure you study.

- Defence 3 is the strongest, and worth naming explicitly.
- It also says when the design *fails*: when the agent optimizes precisely your outcome.

Road map: two routes through this chapter

Everything ahead assumes unconfoundedness. The methods split on **what they do when a treated unit has no comparable control.**

	Route 1: interpolation	**Route 2: extrapolation**
Rests on	Common support	Functional form
Fills the gap by	It does not — it refuses	Projecting the fitted model into it
Methods	Subclassification, matching, propensity scores	Regression

Common support

- Unconfoundedness alone is not enough. You also need units to compare.
- **Strong common support** (needed for the ATE):

$$0 < \Pr(D = 1 \mid X = x) < 1 \quad \text{for all } x$$

At every value of X, both treated and untreated units must exist.

The two assumptions differ in kind

Unconfoundedness	**Common support**
Untestable	**Verifiable in your data**
Defended with institutional knowledge	Checked with a histogram
Failure $\Rightarrow$ bias	Failure $\Rightarrow$ extrapolation

- Common support is the assumption you have **no excuse** for not checking — and the one most papers skip.
- This distinction is the key to the NSW debate in Lecture 3b.

Weighting treatment effects under unconfoundedness

Put the two together and the *ATE* becomes computable.

$$
\begin{aligned}
ATE(x) &= E\big[\, Y^1 \mid X{=}x \,\big] - E\big[\, Y^0 \mid X{=}x \,\big] \\
&= \underbrace{E\big[\, Y^1 \mid D{=}1, X{=}x \,\big] - E\big[\, Y^0 \mid D{=}0, X{=}x \,\big]}_{\text{unconfoundedness — used on } \textit{both} \text{ terms}} \\
&= \underbrace{E\big[\, Y \mid D{=}1, X{=}x \,\big] - E\big[\, Y \mid D{=}0, X{=}x \,\big]}_{\text{switching equation}}
\end{aligned}
$$

$$
ATE = \underbrace{\sum_x \Big(E[Y \mid D{=}1, X{=}x] - E[Y \mid D{=}0, X{=}x] \Big) \Pr(X{=}x)}_{\text{iterated expectations; common support}}
$$

- $ATE(x)$ is the effect **within stratum** x — the CATE. The last line averages those with **population** weights.
- **Unconfoundedness is used twice**, for Y^1 and for Y^0. That is why the *ATE* needs the *strong* version.

A three-stratum example

Suppose the effect on mortality differs by age group, and the groups are equally sized:

Stratum	Effect in stratum (pp)	$\Pr(X = x)$
Young	80	1/3
Middle-aged	55	1/3
Old	100	1/3

$$\begin{aligned}\widehat{ATE} &= \frac{N_y}{N} \times SDO_y + \frac{N_m}{N} \times SDO_m + \frac{N_o}{N} \times SDO_o \\ &= \frac{1}{3} \times 80 + \frac{1}{3} \times 55 + \frac{1}{3} \times 100 \\ &= 78.33\end{aligned}$$

- If there's no old people in the treatment group, then clearly you cannot calculate SDO_o and so common support ultimately breaks the whole estimation process down.

Estimating the ATT under weaker assumptions

- A weaker assumption suffices if you only want the ATT:

$$Y^0 \perp\!\!\!\perp D \mid X$$

- Only the **untreated** potential outcome must be independent of treatment.
- **Weak common support**: Controls must exist wherever treated units do — but not conversely.

$$\Pr(D = 0 \mid X) > 0 \quad \text{for all } X \text{ such that } \Pr(D = 1 \mid X) > 0$$

Why the asymmetry makes sense

$$
\begin{aligned}
ATT(x) &= E[Y^1 \mid D{=}1, X{=}x] - E[Y^0 \mid D{=}1, X{=}x] \\
&= \underbrace{E[Y \mid D{=}1, X{=}x]}_{\text{switching equation}} - E[Y^0 \mid D{=}1, X{=}x] \\
&= E[Y \mid D{=}1, X{=}x] - \underbrace{E[Y \mid D{=}0, X{=}x]}_{\text{weak unconfoundedness}}
\end{aligned}
$$

$$
ATT = \underbrace{\sum_x \Big(E[Y \mid D{=}1, X{=}x] - E[Y \mid D{=}0, X{=}x] \Big) \Pr(X{=}x \mid D{=}1)}_{\text{common support, wherever treated units live}}
$$

- Weighting by $\Pr(X \mid D{=}1)$ — covariates **among the treated** — makes it the ATT: strata with no treated units get zero weight.
- If you can only defend the weak version, your estimand is the ATT.

4. Subclassification

Cochran's method

- The oldest estimator here, and the most transparent. Three steps:
  1. **Stratify** on X.
  2. Compute the treated-minus-control difference **within** each stratum.
  3. **Average** the strata, weighting by whatever the estimand requires.
- No functional form is assumed. Within a stratum, units are identical on X by construction.
- The weights define the estimand:

Estimand	Weight stratum k by	Interpretation
ATE	n_k / N	population share
ATT	$n_{k,\text{treated}} / N_{\text{treated}}$	treated share
ATU	$n_{k,\text{untreated}} / N_{\text{untreated}}$	control share

The Titanic: the raw comparison

- Treated: 325 first-class passengers, 62.5% survived.
- Control: 1,876 others, 27.1% survived.

$$SDO = 62.5 - 27.1 = 35.4 \text{ percentage points}$$

- Taken at face value: a first-class ticket nearly *tripled* your survival odds.
- But first class held a very different mix of people. Sex and age are confounders, and observable.
- Stratify on {sex, age}: four strata.

The four strata

Stratum	*n*	first class	survival, 1st	survival, other	difference
Adult male	1,667	175	32.6%	18.8%	13.7 pp
Adult female	425	144	97.2%	62.6%	34.6 pp
Male child	64	5	100.0%	40.7%	59.3 pp
Female child	45	1	100.0%	61.4%	38.6 pp

Three estimands, three answers

Stratum	Difference in survival	Stratification weight k		
		ATE	*ATT*	*ATU*
Adult male	13.7 pp	0.757	0.538	0.795
Adult female	34.6 pp	0.193	0.443	0.150
Male child	59.3 pp	0.029	0.015	0.031
Female child	38.6 pp	0.020	0.003	0.023
	SDO	$\widehat{ATE}$	$\widehat{ATT}$	$\widehat{ATU}$
Estimate	35.4 pp	19.6 pp	23.8 pp	18.9 pp

- Each estimate is a **weighted average of the same four differences**, read down its column: $0.757(13.7) + 0.193(34.6) + 0.029(59.3) + 0.020(38.6) = 19.6$ pp.
- The raw 35.4 pp falls to roughly **20 pp** once composition is held fixed.
- $ATT > ATE$ because first class was disproportionately adult *female*.

The curse of dimensionality

- Now delete one passenger: the single first-class female child.
- The stratum still exists — 44 girls travelled in other classes — but has **no treated unit**. The difference is undefined.

Parameter	Still identified?
ATE	**No** — needs all four differences
ATU	**No** — female children get weight 0.023
ATT	**Yes** — 23.7 pp from the three occupied strata

- One passenger of 2,201 destroys two of the three parameters. The *ATT* survives because it never needed that cell.
- This is common support failing.

Why it gets worse fast

- 1 binary covariate: 2 cells. 10 binary: $2^{10} = 1{,}024$. 1 continuous: infinitely many.
- Cells grow exponentially; your sample doesn't. Empty cells become the norm.
- Non-parametric methods **fail**: no estimate.
- Regression **doesn't fail**: it extrapolates across empty cells via functional form — and reports a number with a standard error.

5. Exact Matching

Matching as explicit imputation

- Every causal inference is an imputation problem: the counterfactual is missing and must be filled in.
- Most methods impute **implicitly**, inside a formula. Matching does it **explicitly**, unit by unit.
- Start from the ATT. Exactly **one** term is unavailable:

$$ATT = \underbrace{E[Y^1 \mid D{=}1]}_{\text{observed: } Y{=}Y^1 \text{ when } D{=}1} - \underbrace{E[Y^0 \mid D{=}1]}_{\textbf{missing}}$$

- The first term is the treated sample mean. Matching builds a stand-in for the second.

Replacing the missing term

Condition on X, and the assumption does its work:

$$E[Y^0 \mid D{=}1, X{=}x] = \underbrace{E[Y^0 \mid D{=}0, X{=}x]}_{\text{weak unconfoundedness}}$$

$$= \underbrace{E[Y \mid D{=}0, X{=}x]}_{\text{switching equation}}$$

$$E[Y^0 \mid D{=}1] = \underbrace{\sum_x E[Y \mid D{=}0, X{=}x] \Pr(X{=}x \mid D{=}1)}_{\text{common support: this must exist wherever treated units live}}$$

- The missing quantity is now written entirely in terms of control units.
- **Matching is simply a non-parametric estimate of** $E[Y \mid D{=}0, X{=}X_i]$ — take the controls that look like i, and average them.

The estimator, and what to do about ties

Let $\mathcal{J}(i)$ be the set of control units matched to treated unit i, and $M_i = |\mathcal{J}(i)|$.

$$\widehat{ATT} = \frac{1}{N_T} \sum_{i\,:\,D_i=1} \left(Y_i - \underbrace{\frac{1}{M_i} \sum_{j \in \mathcal{J}(i)} Y_j}_{\widehat{Y_i^0}} \right)$$

- If two controls sit at the same distance from i, use both.

The same idea for the ATE

For a control unit j, the missing half is Y^1. Let $\mathcal{I}(j)$ be the treated units matched to it, $M_j = |\mathcal{I}(j)|$.

$$\widehat{ATE} = \frac{1}{N}\Bigg[\underbrace{\sum_{i\,:\,D_i=1}\Big(Y_i - \tfrac{1}{M_i}\sum_{j\in\mathcal{J}(i)} Y_j\Big)}_{\text{treated: observed} - \text{imputed}} + \underbrace{\sum_{j\,:\,D_j=0}\Big(\tfrac{1}{M_j}\sum_{i\in\mathcal{I}(j)} Y_i - Y_j\Big)}_{\text{control: imputed} - \text{observed}}\Bigg]$$

- Every unit contributes one *estimated* treatment effect; the ATE averages all N of them.
- Equivalently $\widehat{ATE} = \frac{N_T}{N}\widehat{ATT} + \frac{N_C}{N}\widehat{ATU}$.

Exact matching and its limit

- With discrete X and full common support, exact matching and subclassification give the **same answer**.
- The difference: subclassification aggregates to stratum effects first; matching imputes unit by unit.
- Same fatal problem: with a continuous covariate, **exact matches occur with probability zero.**
- Two responses, and they define the next two sections:
  - Give up on exact equality — match to the *nearest* unit (§6), then correct that error (§8).
  - Give up on matching X itself — match on a *summary* of X (Lecture 3b, §1, Propensity Scores).

6. Nearest-Neighbour Matching

One confounder

- Drop the demand for exact equality. Match each treated unit to the control that minimizes $|X_i - X_j|$.
- The gap you accept is the **matching discrepancy**; §8 is about what it costs.
- Choices you must make and report:
  - **With or without replacement.** With replacement always takes the closest match, so it **lowers bias** and raises **variance.**
  - **How many neighbours, M.** More neighbours: less variance, more bias.
  - **Caliper**: refuse any match beyond a maximum distance. Better to drop a treated unit than match it to something unlike it.

Multiple dimensions and the distance metric

With several covariates, "closest" requires a definition.

- **Euclidean**: $\sqrt{\sum_k (X_{ik} - X_{jk})^2}$. Scale-dependent: a variable in won swamps one measured as a ratio.
- **Normalized Euclidean**: divide each covariate by its standard deviation. Fixes scale, ignores correlation.
- **Mahalanobis**: $\sqrt{(X_i - X_j)' \Sigma^{-1} (X_i - X_j)}$. Standardizes scale **and** accounts for correlation between covariates. The default choice.

- If size and leverage are highly correlated, Euclidean distance double-counts what is essentially one dimension.

7. Inference with Matching

Why the usual standard errors fail

- Matched observations are **dependent by construction**: the same control may serve several treated units, and matches were chosen *using the data*.
- Conventional standard errors treat the matched sample as randomly drawn. It was not, and they understate uncertainty.
- Abadie and Imbens (2008) show conventional standard errors are invalid for matching estimators.
- Use the **Abadie–Imbens variance estimator**, which prices in how many times each control was reused.
- Check what your software reports *by default*: several packages give the naive standard error unless told otherwise.

The Abadie–Imbens variance estimator

Let $\hat{\tau}_i = Y_i - \frac{1}{M}\sum_{m=1}^{M} Y_{j_m(i)}$ be unit i's **own** effect, so $\widehat{ATT} = \frac{1}{N_T}\sum_{i:D_i=1}\hat{\tau}_i$ is their average. K_j counts how often control j is used.

$$\hat{\sigma}^2_{ATT} \;=\; \underbrace{\frac{1}{N_T}\sum_{i\,:\,D_i=1}\left(\hat{\tau}_i - \widehat{ATT}\right)^2}_{\text{(a) spread of the unit-level effects}} \;+\; \underbrace{\frac{1}{N_T}\sum_{j\,:\,D_j=0}\frac{K_j(K_j-1)}{M^2}\,\hat{\sigma}^2(X_j)}_{\text{(b) penalty for reusing controls}}$$

$$\mathrm{SE}\left(\widehat{ATT}\right) \;=\; \sqrt{\hat{\sigma}^2_{ATT} \,/\, N_T}$$

- Term (a) is the **sample variance of the** $\hat{\tau}_i$. Alone it is the without-replacement formula: with $K_j \in \{0,1\}$, term (b) vanishes.
- $\hat{\sigma}^2(X_j) = \mathrm{var}(Y \mid X_j, D{=}0)$, the conditional outcome variance among controls at X_j.
- Here M is **fixed** — the clean $K_j(K_j{-}1)$ split needs it.

In Stata

Cunningham's training data: 10 trainees, 20 non-trainees, matched on age and gpa.

```
use https://github.com/scunning1975/mixtape/raw/master/training_inexact.dta, clear

* Minimized euclidean distance with usual variance and robust
teffects nnmatch (earnings age gpa) (treat), atet nn(1) metric(eucl) vce(iid)
teffects nnmatch (earnings age gpa) (treat), atet nn(1) metric(eucl) generate(match1)

* Minimized Mahalanobis distance with usual variance and robust
teffects nnmatch (earnings age gpa) (treat), atet nn(1) metric(maha) vce(iid)
teffects nnmatch (earnings age gpa) (treat), atet nn(1) metric(maha) generate(match2)
```

- `(earnings age gpa)` is outcome-then-covariates; `(treat)` is the treatment. `atet` asks for the *ATT*, `nn(1)` for one neighbour.
- `vce(iid)` is the constant-variance standard error. **Drop it and you get Abadie–Imbens**, Stata's default — so report the second command of each pair.
- `generate(match1)` writes each unit's matched partner back into the data. Always look at it.

Two metrics, two answers

	$\widehat{ATT}$	`vce(iid)`	**Abadie–Imbens**	distinct controls used
Euclidean	1607.5	38.7	38.7	10 of 20
Mahalanobis	1315.0	325.5	590.1	**3** of 20
Truth	1607.5			

- age has standard deviation 9.1, gpa only 0.57, so raw Euclidean distance **matches on age and ignores gpa** — the scale problem from §6. Every trainee gets a *different* control: $K_j \leq 1$, term (b) is zero, and both standard errors coincide.
- Mahalanobis makes gpa count — and gpa is noise. **Three controls serve all ten trainees**; one is used five times.

8. Bias Correction

The matching discrepancy is bias, not noise

Matching was only *approximate*: $X_{j(i)} \approx X_i$. What does the leftover gap cost?

- Abadie–Imbens (2006): More data does not help much.
- Usually you cannot get more data anyway. Abadie–Imbens (2011) is the second-best answer: **model the bias and subtract it.**

An outcome model for the missing Y^0

$$\mu^0(x) = E[\,Y \mid X{=}x, D{=}0\,] \;=\; \underbrace{E[\,Y^0 \mid X{=}x\,]}_{\text{unconfoundedness}}$$

$$\mu^1(x) = E[\,Y \mid X{=}x, D{=}1\,] \;=\; \underbrace{E[\,Y^1 \mid X{=}x\,]}_{\text{unconfoundedness}}$$

$$Y_i = \mu^{D_i}(X_i) + \varepsilon_i, \qquad E[\,\varepsilon_i \mid X_i, D_i\,] = 0$$

- The second equality on each line **is** unconfoundedness: it lets a regression fitted on controls speak about treated units.
- So far, all definition. The **assumption** arrives only when we choose a functional form for $\hat{\mu}^0$.

Where the bias hides

Substitute $Y_i = \mu^{D_i}(X_i) + \varepsilon_i$ into $\widehat{ATT} = \frac{1}{N_T}\sum_{D_i=1}(Y_i - Y_{j(i)})$, subtract the true ATT, then add and subtract $\frac{1}{N_T}\sum_{D_i=1}\mu^0(X_i)$:

$$\widehat{ATT} - ATT = \underbrace{\frac{1}{N_T}\sum_{D_i=1}\left(\mu^1(X_i) - \mu^0(X_i) - ATT\right)}_{\text{(1) sampling variation in } \textit{which} \text{ units got treated}} + \underbrace{\frac{1}{N_T}\sum_{D_i=1}\left(\varepsilon_i - \varepsilon_{j(i)}\right)}_{\text{(2) mean-zero noise}} + \underbrace{\frac{1}{N_T}\sum_{D_i=1}\left(\mu^0(X_i) - \mu^0(X_{j(i)})\right)}_{\text{(3) } \textbf{matching bias}}$$

- Scaled by $\sqrt{N_T}$, rows (1) and (2) average out.
- Row (3) does not. Its summands are **signed**: if the nearest control is systematically older or smaller, every term carries the same sign.

Two things make the bias big

$$\text{Matching bias} \;=\; E\big[\,\mu^0(X_i) - \mu^0(X_{j(i)}) \,\big|\, D = 1\,\big]$$

- Read it as a product of two quantities, and it is zero if **either** is:
  - **How bad the matches are** — the size of $X_i - X_{j(i)}$. Perfect matching kills it.
  - **How steep μ^0 is** — the heterogeneity of Y^0 across strata of X. A flat μ^0 kills it too.
- Match a 13-year-old to a 15-year-old and nothing is lost *unless* Y^0 differs between 13- and 15-year-olds.
- For the ATT it is Y^0 that is missing and imputed, so only μ^0 has to be modelled.

The Abadie–Imbens procedure, in five steps

1. **Match.** Nearest neighbour matching; form $\hat{\delta}_i = Y_i - Y_{j(i)}$ for each treated unit.
2. **Regress.** OLS on the **matched control units only**: $Y_j = \alpha + \beta X_j + e_j$.
3. **Predict.** $\hat{\mu}^0(X_i) = \hat{\alpha} + \hat{\beta} X_i$ and $\hat{\mu}^0(X_{j(i)}) = \hat{\alpha} + \hat{\beta} X_{j(i)}$.
4. **Subtract** the estimated bias:

$$\widehat{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \Big[(Y_i - Y_{j(i)}) - (\hat{\mu}^0(X_i) - \hat{\mu}^0(X_{j(i)})) \Big]$$

5. **Report an Abadie–Imbens standard error** (§7, Inference with Matching). The correction changes the estimate, not the reuse problem.

- Step 2 never touches a treated outcome: $\hat{\mu}^0$ is fitted where Y^0 is *observed*, then transplanted. Unconfoundedness permits the transplant.

Ten workers

Unit	Worker	Y^1	Y^0 / $\widehat{Y^0}$	$\hat{\delta}_i$	Age	D
1	Andy	\$200	\$195	\$5	23	1
2	Betty	\$250	\$225	\$25	27	1
3	Chad	\$150	\$140	\$10	29	1
4	Doug	\$300	\$195	\$105	22	1
5	Edith	—	\$225		27	0
6	Fred	—	\$500		32	0
7	Gina	—	\$200		26	0
8	Hank	—	\$190		26	0
9	Inez	—	\$180		28	0
10	Janet	—	\$140		29	0

- Andy (23) and Doug (22) both match Gina and Hank, the 26-year-olds: $\widehat{Y^0} = \frac{200+190}{2} = 195$. Betty and Chad match exactly.
- $\widehat{ATT} = (5 + 25 + 10 + 105)/4 = \$36.25$. Fred and Inez are never used.
- Both inexact matches point the same way: the match is *older* than the trainee.

Steps 2–4, worked through

Step 2, over the four *matched* controls — Edith $(27, 225)$, Gina $(26, 200)$, Hank $(26, 190)$, Janet $(29, 140)$: $\hat{\mu}^0(\text{age}) = 683.75 - 18.33\,\text{age}$.

Worker	Age	Match age	$\hat{\delta}_i$	$\hat{\beta}(X_i - X_{j(i)})$	$\hat{\delta}_i^{BC}$
Andy	23	26	\$5	\$55.0	−\$50.0
Betty	27	27	\$25	\$0	\$25.0
Chad	29	29	\$10	\$0	\$10.0
Doug	22	26	\$105	\$73.3	\$31.7

$\widehat{ATT} = \$36.25 \longrightarrow \widehat{ATT}^{BC} = \$4.17.$

- Exact matches get **zero** correction. Only the two bad matches move.
- $\hat{\beta} < 0$: among *these four* controls earnings fall with age. Refit on all six and Fred $(32, \$500)$ flips it to $\hat{\beta} = +43.7$, giving $\widehat{ATT}^{BC} = \$112.6$.
- One outlier. The "small" regression is doing enormous work — the method's exposure.

What the correction does and does not do

- It converts a *matching* problem into a *much easier regression* problem. It does not eliminate the regression.
- It is honest about its assumption: you must name a functional form for $\hat{\mu}^0$, and the ten-worker example shows how much that moves the answer.
- It does **not** repair the matches.
- Report the raw and bias-corrected estimates side by side. A large gap is a **diagnostic**: your matches were poor and the outcome model is load-bearing.

Wrapping Up

Summary [Part 1]: the assumption

- **Assignment mechanisms resolve bias, not models.** The SDO decomposition is an identity; only a claim about how treatment was assigned deletes its bias terms.
- **Covariates come in four kinds**: confounders (include), colliders (never), precision covariates (optional), mediators (*depends on the estimand*: forbidden for the total effect, required for the partial). Pre-treatment timing is your best protection.
- **Unconfoundedness**, $(Y^1, Y^0) \perp\!\!\!\perp D \mid X$: randomization *within* strata of X. Untestable; defended institutionally, and Imbens's third defence is strongest in finance.
- **Common support** is a *separate* assumption, and unlike unconfoundedness it is **checkable**. It is asymmetric: the ATT needs only $e(X) < 1$, the ATE both ends.
- **Subclassification** makes the estimand a choice about weights. On the Titanic: $SDO = 35.4$ pp, but $ATE = 19.6$, $ATT = 23.8$, $ATU = 18.9$.
- The **curse of dimensionality** is structural, not a small-sample nuisance. One passenger of 2,201 killed the ATE and ATU; the ATT survived.

Summary [Part 2]: matching, and what it costs

- **Matching** is explicit imputation: replace the missing Y_i^0 with a comparable control's outcome. Exact matching is the honest version; it fails on continuous X.
- **Nearest-neighbour matching** buys feasibility with a **matching discrepancy** that is *signed*, not noise.
- **That bias is a product**: how bad the matches are, times how steep μ^0 is. Either factor at zero kills it.
- **Two separate repairs.** The *variance* estimator fixes inference: reusing controls makes comparisons dependent, and $K_j(K_j - 1)$ prices that in. *Bias correction* fixes the estimate, subtracting a modelled $\hat{\mu}^0(X_i) - \hat{\mu}^0(X_{j(i)})$.
- **Scale matters.** Euclidean and Mahalanobis are different estimators, not different implementations: 1607.5 against 1315.0, truth 1607.5.
- **Report raw and bias-corrected side by side.** A large gap says the matches were poor and the outcome model is load-bearing.
- Next: the same assumption, executed by collapsing X to **one number**.

References

- Cunningham, S. *Causal Inference: The Remix*, Ch. 5.
- Cochran, W. (1968). The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies. *Biometrics* 24, 295–313.
- Imbens, G. (2004). Nonparametric Estimation of Average Treatment Effects under Exogeneity. *Review of Economics and Statistics* 86, 4–29.
- Abadie, A. and G. Imbens (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica* 74, 235–267.
- Abadie, A. and G. Imbens (2008). On the Failure of the Bootstrap for Matching Estimators. *Econometrica* 76, 1537–1557.
- Abadie, A. and G. Imbens (2011). Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics* 29, 1–11.
- Lin, Z., P. Ding and F. Han (2023). Estimation Based on Nearest Neighbor Matching: From Density Ratio to Average Treatment Effect. *Econometrica* 91, 2187–2217.
- Angrist, J. and J.-S. Pischke (2009). *Mostly Harmless Econometrics*, Ch. 3. Princeton University Press.