Matching

Professor Ji-Woong Chung
Korea University

Outlilne

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Matching Methods – Basic Idea [Part 1]

- Matching provides an intuitive way to estimate causal effects when treatment is not randomly assigned.
- Core idea:
  - For each treated unit, find one (or several) comparable untreated units that look similar in observed characteristics X.
  - These untreated "matches" serve as the counterfactual outcomes that the treated units would have experienced had they not been treated.
  - The treatment effect is then the difference in outcomes y between treated units and their matches.

Matching Methods – Basic Idea [Part 2]

- More formally:
  - For each value (or neighborhood) of X where both treated and untreated units exist:
    - Pair each treated unit with one or more untreated units that have the same (or very similar) X.
    - Compute the difference in outcomes y within each matched pair/group.
  - Averaging these within-X differences gives an estimate of the Average Treatment Effect on the Treated (ATT).
- **Key point:** Matching replaces the missing counterfactual outcome with outcomes from observational "clones" based on X.

Matching Methods – Intuition

- What assumptions does matching rely on?
  - **(1) Treatment is not randomly assigned in the raw data.**
    - If treatment were random, treated and untreated groups would be comparable without matching.
    - Matching is needed precisely because treated units differ systematically from untreated units.
  - **(2) Conditional on X, treatment is "as good as random."**
    - Also known as **Conditional Mean Independence (CMI)** or **Selection on Observables**.
    - Formally: $E[Y(0) \mid D = 1, X] = E[Y(0) \mid D = 0, X]$.
    - This means untreated outcomes after conditioning on X provide valid counterfactuals for treated units.

Matching is a "Control Strategy"

- Matching is simply one way to "control for" observable differences X between treated and untreated units.
- By conditioning on X, matching attempts to create treated and untreated groups that are comparable, mimicking a randomized experiment.
- What is another way to control for observable characteristics when estimating treatment effects?

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Matching and OLS: Not That Different

- Another way to estimate the treatment effect while controlling for covariates is simply:
  - **Run a regression of y on the treatment indicator and the relevant X's.**
  - If Conditional Mean Independence (CMI) holds given X, then OLS delivers a valid causal effect.
- To mimic matching very closely:
  - Include **dummy variables for each value (or group) of X** in the regression.
  - Then OLS estimates a treatment effect by comparing treated and untreated units *within* each X-cell.
  - This is essentially "parametric matching": OLS restricts comparisons to units that share the same X.
- **So how are matching and OLS different if they both control for X?**

Matching versus Regression

- A useful way to think about the difference:
  - **OLS is a weighted matching estimator.**
  - Both methods compare treated and untreated observations with the same covariates.
  - The key difference lies in **how the comparisons are weighted across different values of** X.
- The weighting details can get technical:
  - See Angrist & Pischke (Mostly Harmless Econometrics), Section 3.3.1.
  - But the basic idea: matching and OLS emphasize different parts of the data.

Matching vs Regression – Example [Part 1]

- Consider an example with discrete covariates X:
  - Step 1: Compute a simple matching estimator:
    - For each treated unit, find untreated units with the same X.
    - Compute average differences in their outcomes.
  - Step 2: Run OLS:
    - Regress y on the treatment indicator.
    - Add a full set of **indicator variables for every value of** X.
- Including X-indicators makes the OLS comparisons *within* each X-cell—just like matching.
- This version of OLS is extremely flexible (nonparametric in X) and conceptually close to matching.

Matching vs Regression – Example [Part 2]

- Even in this setup, the matching and OLS estimates will generally differ:
  - **Matching weights cells according to where the treated units are.**
    - If many treated units share a particular value of X, that cell receives more weight.
  - **OLS weights cells based on the amount of treatment variation within each cell.**
    - Cells with an equal mix of treated and untreated observations receive more weight.
    - Cells with only treated or only untreated observations contribute little or nothing.
- **OLS and matching use the same comparisons but emphasize different parts of the data.**

Matching vs Regression – Numerical Illustration

- Consider two covariate cells, $X = 0$ and $X = 1$, each with 100 observations:

	Counts		Mean outcomes	
Cell	Treated	Control	Treated $\bar{y}_1(X)$	Control $\bar{y}_0(X)$
$X = 0$	80	20	10	8
$X = 1$	20	80	6	3

- Within-cell treatment effects:

$$\hat{\tau}(X{=}0) = 10 - 8 = 2, \qquad \hat{\tau}(X{=}1) = 6 - 3 = 3.$$

- **Matching (ATT) weights by where treated units are:**

$$w_0^M = \frac{80}{80 + 20} = 0.8, \qquad w_1^M = 0.2,$$

$$\widehat{\text{ATT}}^M = 0.8 \cdot 2 + 0.2 \cdot 3 = 2.2.$$

- **OLS with X-dummies weights by treatment variation within cells:**

$$p_0 = 0.8,\ p_1 = 0.2, \quad w_j^{OLS} \propto N_j\, p_j(1 - p_j).$$

Here,

$$w_0^{OLS} \propto 100 \cdot 0.8 \cdot 0.2 = 16, \quad w_1^{OLS} \propto 100 \cdot 0.2 \cdot 0.8 = 16 \Rightarrow w_0^{OLS} = w_1^{OLS} = 0.5,$$

$$\widehat{\tau}^{OLS} = 0.5 \cdot 2 + 0.5 \cdot 3 = 2.5.$$

- Matching emphasizes cells with many treated units ($X{=}0$), while OLS emphasizes cells with more balanced treatment variation, giving equal weight to $X{=}0$ and $X{=}1$.

Matching vs Regression – Bottom Line

- Angrist & Pischke emphasize that:
  - In most applications, the numerical differences between matching and OLS tend to be small.
  - Both rely critically on the **same identifying assumption**: selection on observables (CMI).
- However, both approaches share a key limitation:
  - They only control for **observed** covariates X.
  - If unobserved variables jointly affect treatment and outcomes, both matching and OLS are biased.

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Matching – Key Limitation [Part 1]

- What distinguishes matching from research designs such as IV, natural experiments, or regression discontinuity?
  - **Matching does not introduce any new exogenous variation.**
  - It simply reweights or reorganizes the existing data based on observable characteristics X.
- Implication:
  - If the original OLS regression suffers from endogeneity, matching will generally suffer from the *same* endogeneity.
  - Matching does not "solve" the bias—it only attempts to make treated and untreated units more comparable *conditional on observed covariates*.

Matching – Key Limitation [Part 2]

- Why did we worry about OLS in the first place?
  - When treatment is not randomly assigned, simply controlling for a set of covariates X often does *not* restore exogeneity.
  - Many important omitted variables may be **unobserved**, so matching cannot adjust for them.
  - Self-selection, anticipation effects, and reverse causality all remain threats.
- **Matching inherits all of these concerns.**
  - It improves balance in observables, but *not* in unobservables.
- (Regression discontinuity is different: treatment assignment is locally as good as random by design.)

Matching – Key Limitation [Part 3]

- Important reminders:
  - Matching **cannot** fix:
    - Simultaneity or reverse causality.
    - Measurement error biases in treatment or outcome variables.
    - Omitted variable bias caused by **unobservables**.
  - Matching only controls for the variables you match on.
    - "You cannot match on what you cannot measure."

Matching – So, What Good Is It? [Part 1]

- Given its limitations, one might think matching is not very useful:
  - It is essentially another "control strategy," similar in spirit to OLS.
  - It does not create quasi-experimental variation or solve identification problems.
  - It does not overcome unobserved confounding.
- But matching **does have several practical advantages**.

Matching – So, What Good Is It? [Part 2]

- Matching can still be valuable in empirical practice:
  - **Robustness checks:** Matching provides a nonparametric benchmark against the regression estimate.
  - **Improved covariate balance:** Matching helps reduce extreme extrapolation and ensures treated and untreated units are compared only where support overlaps.
  - **Finite-sample improvement:** Matching sometimes performs better than OLS when the functional form of the regression is misspecified.
  - **Diagnostic tool:** Identifies regions of poor overlap (common support problems) that OLS masks.
- More on these benefits later.

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

First, Some Notation

- We study the causal effect of a binary treatment d:
  - $d = 1$: treated
  - $d = 0$: untreated (control)
- Potential outcomes:
  - $y(1)$: outcome unit *would* experience under treatment
  - $y(0)$: outcome unit *would* experience under control
- Observed outcome:

$$y = d \cdot y(1) + (1 - d) \cdot y(0)$$

- Observable covariates:

$$X = (x_1, \ldots, x_k)$$

These are variables we will match on.

Identification Assumptions

- To estimate treatment effects using matching, we need two key assumptions:
  - **Assumption 1: Unconfoundedness (Selection on Observables)**
  - **Assumption 2: Overlap (Common Support)**
- These assumptions together ensure that:
  - Treated and control units are comparable *within* levels of X
  - There exist control units "similar enough" to every treated unit

Assumption #1 – Unconfoundedness

- Formal statement:

$$(y(0), y(1)) \perp\!\!\!\perp d \mid X$$

- Interpretation:
  - Once we condition on observable covariates X, treatment behaves "as if random."
  - No remaining unobserved selection into treatment after conditioning on X.
- Implication:
  - Within each value of X, the untreated group can stand in for the unobserved counterfactual for the treated group.
  - This makes matching possible.

"Unconfoundedness" Explained

- Stronger than Conditional Mean Independence:

$$E[y(0) \mid d = 1, X] = E[y(0) \mid d = 0, X].$$

- Equivalent regression statement:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \gamma d + u,$$

  where d is independent of error u, i.e., $d \perp u \mid X$.

- Propensity score matching and several other matching estimators require this stronger version.

- This assumption is untestable.

Assumption #2 – Overlap

- For all covariate values X,

$$0 < P(d = 1 \mid X) < 1.$$

- Interpretation:
  - Every type of unit has a positive chance of being treated and untreated.
  - Treated units must have "neighbors" in the control group with similar X.
- Why necessary?
  - Without overlap, matching cannot produce valid counterfactuals.
  - If no untreated units exist for a treated X-type, the effect is not identified for them.

"Overlap" in Practice

- Exact overlap is rare—especially when:
  - X contains continuous variables,
  - X is high-dimensional, or
  - there is strong selection into treatment.
- In practice:
  - We match on units with *similar* values of X, not identical ones.
  - This introduces small-sample bias.
- Abadie and Imbens (2008) show:
  - Nearest-neighbor matching is biased but can be bias-corrected.
  - They provide analytical corrections and variance formulas.

Average Treatment Effect (ATE)

- Under unconfoundedness + overlap:

$$\begin{aligned} ATE(X = x) &= E[y(1) - y(0) \mid X = x] \\ &= E[y \mid d = 1, X = x] - E[y \mid d = 0, X = x]. \end{aligned}$$

- This means ATE for each $X = x$ is just the difference in means.
- To obtain the population ATE:

$$ATE = \int ATE(X = x)\, f_X(x)\, dx,$$

i.e., a weighted average over the distribution of X.

(See Roberts & Whited, p. 68.)

Difficulty with Exact Matching

- Exact matching requires:

$$X_i = X_j \quad \text{for a treated-control pair.}$$

- Problems:
  - In high-dimensional X, cells become very sparse ("curse of dimensionality").
  - With continuous variables, exact equality almost never occurs.
- Result:
  - Many treated units may have no valid matches.
  - Matching estimators become biased or undefined without smoothing / nearest-neighbor methods.

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Matching on Covariates – Step #1

- The **first step** in matching: decide how to measure "closeness" between two observations in terms of their covariates.
- We choose a **distance metric**:

$$\|X_i - X_j\|$$

- Intuition:
  - This quantifies how similar observation i is to observation j based on their covariate vectors.
  - Observations with small distance are considered "good matches."
- Example: Euclidean distance

$$\sqrt{(X_i - X_j)'(X_i - X_j)}.$$

- (We will later discuss why Euclidean distance can be problematic when covariates have different scales and why Mahalanobis distance is sometimes preferred.)

Matching on Covariates – Step #2

- For each observation i, find the M closest observations *with the opposite treatment status*.
- More precisely:
  - If $d_i = 1$ (treated), match to the M nearest untreated units.
  - If $d_i = 0$ (control), match to the M nearest treated units.
- Interpretation:
  - These M observations form our estimate of what i's outcome *would have been* under the opposite treatment status.
  - Using multiple neighbors instead of just one (i.e., $M > 1$) reduces noise and variance.

Matching on Covariates – Step #3: Notation

- Let $l_m(i)$ denote the index of the m-th closest match to unit i among observations with $d \neq d_i$.
- Example:
  - Suppose $i = 4$ is treated.
  - $l_1(4)$: closest control observation to unit 4.
  - $l_2(4)$: second closest control unit to 4, etc.
- Define the set of the M closest matches:

$$L_M(i) = \{l_1(i), l_2(i), \ldots, l_M(i)\}.$$

- This notation keeps track of which units we use to construct the counterfactual outcomes.

Matching on Covariates – Step #4

- For each unit i, construct estimates of the two potential outcomes:

$$\hat{y}_i(0) = \begin{cases} y_i & \text{if } d_i = 0 \text{ (we observe } y(0)), \\ \frac{1}{M} \sum_{j \in L_M(i)} y_j & \text{if } d_i = 1 \text{ (impute } y(0)). \end{cases}$$

$$\hat{y}_i(1) = \begin{cases} \frac{1}{M} \sum_{j \in L_M(i)} y_j & \text{if } d_i = 0 \text{ (impute } y(1)), \\ y_i & \text{if } d_i = 1 \text{ (we observe } y(1)). \end{cases}$$

- In words:
  - If unit i was treated, we observe $y(1)$ but must impute $y(0)$ using its matched controls.
  - If unit i was untreated, we observe $y(0)$ but must impute $y(1)$ using its matched treated units.

Interpretation...

$$\hat{y}_i(0) = \begin{cases} y_i & \text{if } d_i = 0, \\ \boxed{\dfrac{1}{M} \displaystyle\sum_{j \in L_M(i)} y_j} & \text{if } d_i = 1 \end{cases}$$

If observation i is treated: we observe its treated outcome $y(1)$, but we never observe its untreated outcome $y(0)$, so we construct it using the average outcome of the M closest

$$\hat{y}_i(1) = \begin{cases} \boxed{\dfrac{1}{M} \displaystyle\sum_{j \in L_M(i)} y_j} & \text{if } d_i = 0, \\ y_i & \text{if } d_i = 1 \end{cases}$$

If observation i is untreated: we observe its untreated outcome $y(0)$, but must impute $y(1)$ using the closest treated units.

Matching constructs a "synthetic twin" for each unit based on similarity in covariates.

Matching on Covariates – Average Treatment Effect (ATE)

- Once all missing potential outcomes are imputed, the ATE is:

$$\widehat{ATE} = \frac{1}{N} \sum_{i=1}^{N} \left[\hat{y}_i(1) - \hat{y}_i(0) \right].$$

- Interpretation:
  - For each unit: (observed outcome under actual treatment) minus (constructed counterfactual under the alternative).
  - Then simply average across all units.
- Under assumptions of unconfoundedness + overlap, this is a consistent estimator of the ATE.

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Matching on Propensity Score – Step #1

- Another way to implement matching is to first estimate a **propensity score** and then match on it.
- The **propensity score** is:

$$ps(X) = P(d = 1 \mid X) = E[d \mid X].$$

- Intuition:
  - Instead of matching on the full k-dimensional covariate vector X,
  - we match on a **single number** summarizing the likelihood of treatment.
- Estimation:
  - Can use Logit, Probit, OLS, machine learning models, etc.
  - Typically: Logit with flexible terms (polynomials, interactions).

Propensity Score – Step #2

- The key Rosenbaum–Rubin (1983) theorem:

$$(y(1), y(0)) \perp\!\!\!\perp d \mid X \quad \Rightarrow \quad (y(1), y(0)) \perp\!\!\!\perp d \mid ps(X).$$

- Meaning:
  - If treatment is unconfounded after conditioning on X,
  - then conditioning on the single number $ps(X)$ is **sufficient** for identification.
- So we can:
  - Match using only $ps(X)$, instead of the entire covariate vector.
  - Or run a regression of y on d and include the propensity score as a control.

Propensity Score – Step #3

- Estimate $ps(X_i) = P(d_i = 1 \mid X_i)$ for every observation.
- Common approaches:
  - Logit (most common), Probit, or even OLS.
  - Add nonlinear or interaction terms for continuous covariates.
  - In large samples: ML methods
- The fitted value $\widehat{ps}(X_i)$ gives the predicted probability of treatment for unit i.

Tangent About Step #3

- Include only covariates that predict treatment d.
- Why?
  - Variables unrelated to treatment assignment only add noise to the model.
  - Excluding irrelevant variables can improve the finite sample performance of matching.
- Practical implication:
  - Economic logic, institutional knowledge, and theory should guide which variables enter the propensity score.
  - This is not necessarily the same set of covariates that predict outcomes y.

Matching on ps(X) – Remaining Steps...

- After estimating $ps(X)$, repeat the matching procedure but use the **difference in propensity scores** as the distance metric.
- Example:
  - If unit i is untreated,
  - choose M treated observations whose propensity scores are closest to $ps(X_i)$.
- This creates a one-dimensional matching problem that avoids the curse of dimensionality associated with high-dimensional X.

Propensity Score – Advantage #1

- Propensity scores reduce subjective choices in covariate matching.
  - No need to choose a multivariate distance metric.
  - No need to standardize or rescale variables.
  - No need to decide how to weight each covariate.
- Because matching is done on a single dimension, it is more transparent and easier to implement.
- Also helps avoid the "curse of dimensionality": Matching in high-dimensional X becomes sparse; matching on $ps(X)$ does not.

Propensity Score – Advantage #2

- Instead of matching, we can directly estimate the ATE using the inverse probability weighting (IPW) formula:

$$ATE = E\left[\frac{d_i - ps(X_i)}{ps(X_i)(1 - ps(X_i))} y_i\right].$$

- This uses the propensity score to reweight observations so treated and untreated groups "look like" each other.
- See Angrist–Pischke (Section 3.3.2) for an intuitive interpretation based on residuals and orthogonality.

Why the IPW Estimator Works

- Key identity under unconfoundedness:

$$E[y(1)] = E\left[\frac{d_i}{ps(X_i)} y_i\right], \qquad E[y(0)] = E\left[\frac{1-d_i}{1-ps(X_i)} y_i\right].$$

- Intuition:
  - Treated units are "rare" when $ps(X)$ is small, so they receive larger weights.
  - Untreated units are "rare" when $1 - ps(X)$ is small, so they receive larger weights.
  - Weighting by the inverse propensity score constructs a **pseudo-population** where treatment is as-good-as random.
- $ATE = E[y(1)] - E[y(0)]$.
- Combine the two expressions:

$$ATE = E\left[\frac{d_i}{ps(X_i)} y_i\right] - E\left[\frac{1-d_i}{1-ps(X_i)} y_i\right].$$

- Rearranging:

$$ATE = E\left[\frac{d_i - ps(X_i)}{ps(X_i)(1-ps(X_i))} y_i\right].$$

- IPW reweights each unit so that treated and untreated groups replicate the distribution of X in the population.
- After weighting, treatment is independent of X — just like a randomized experiment.

But There Is a Disadvantage (Sort of)?

- Sometimes variables do not predict treatment d but do improve prediction of the outcome y.
- Including these variables in covariate matching (or in regression) improves precision:
  - Same logic as adding controls in OLS to reduce residual variance.
- But if such variables do not affect treatment, they *should not* enter the propensity score.
- Angrist and Hahn (2004) show:
  - Using $ps(X)$ and *not* adding irrelevant covariates can yield **better finite-sample properties**.
  - Matching on the full set of X may actually increase noise.

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Practical Considerations

- Matching involves a large number of choices that meaningfully affect results:
  - **Distance metric**: How do we measure similarity in X?
  - **Number of matches**: 1 nearest neighbor or several?
  - **With or without replacement**: Can the same unit be reused?
  - **Which covariates** X to match on?
  - **Covariate vs. propensity score matching**: Which dimension to match on?
- These decisions directly affect bias, variance, and overlap—so matching requires careful design.

Choice of Distance Metric [Part 1]

- Simple Euclidean distance:

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)'(X_i - X_j)}.$$

- Downside:
  - Variables with larger numeric scales dominate the distance.
  - Example: income measured in dollars swamps age measured in years.
  - Therefore Euclidean distance is rarely appropriate without standardization.
- So which variables have more influence?
  - Those with larger variance or measured in larger units.

Choice of Distance Metric [Part 2]

- Common improvements standardize by variances or covariance structure:
  - **Abadie–Imbens (2006) metric**:

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)' \mathrm{diag}(\Sigma_X^{-1})(X_i - X_j)},$$

    which rescales each variable by its variance.
  - **Mahalanobis distance** (most widely used):

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)' \Sigma_X^{-1} (X_i - X_j)}.$$

  - Mahalanobis accounts for both scale differences and covariances between covariates.
- Σ_X^{-1}: inverse of covariance matrix of covariates.

Choice of Matching Approach

- Should you match directly on covariates or on the propensity score?
  - Covariate matching preserves "local" similarity in each dimension.
  - Propensity score matching collapses all covariates into one probability.
  - But propensity scores require assuming a model for treatment.
- No single best choice:
  - Both have strengths and weaknesses depending on overlap, dimensionality, and sample size.
  - Best practice: check that results are robust across multiple matching strategies.

And How Many Matches? [Part 1]

- No universal rule; this is a classic **bias–variance tradeoff**:
  - **1 nearest neighbor** $\rightarrow$ minimal bias, but high variance.
  - **More neighbors** $\rightarrow$ more stable (lower variance), but possibly more biased.
- Why bias increases with more matches?
  - Additional neighbors are usually further away in X-space.

And How Many Matches? [Part 2]

- Two main matching rules:
  - **Nearest-neighbor**: pick the M closest matches regardless of distance.
  - **Caliper matching**: pick all matches within a distance threshold.
- Example of caliper:
  - Using a propensity score caliper of 0.01 means only matches with scores within 1 percentage point are allowed.
- Question: What is advantage of caliper matching?
  - It avoids "bad matches" even if they are the nearest neighbors.
  - Ensures matching only when overlap is meaningful.

And How Many Matches? [Part 3]

- Practical guidance:
  - Try several numbers of matches.
  - If results change substantially as radius or number of matches increases:
    - Matching quality is poor $\rightarrow$ bias risk high.
  - If estimates remain stable but precision improves:
    - Using more matches is acceptable.

With or Without Replacement? [Part 1]

- **With replacement**:
  - A control unit can be used as a match more than once.
  - Produces the best (closest) possible matches $\rightarrow$ lower bias.
  - But reusing units reduces precision and increases variance.
- **Without replacement**:
  - Each unit can be used only once.
  - Ensures more diverse comparison units.
  - But potentially increases bias when the best control is "used up."

With or Without Replacement? [Part 2]

- Roberts–Whited recommend:
  - Use **matching with replacement**.
  - Bias is the primary concern; precision can be improved later.
  - Matching without replacement depends on order of matching, which introduces randomness and instability.

Which Covariates?

- Include all covariates that:
  - affect the outcome y, **and**
  - are correlated with treatment d.
- Why?
  - Omitting an X that drives both treatment and outcome $\rightarrow$ omitted variable bias.
- But avoid covariates affected by the treatment:
  - These are "bad controls" and would block part of the treatment effect.
  - Use lagged covariates instead when possible.

Matches for Whom?

- **ATE**: match both treated and untreated units, imputing both counterfactuals.
- **ATT**: match only for treated units.
  - Appropriate when you care about the treatment effect for the treated population.
- **ATU**: match only for untreated units.
  - Rarely used but conceptually symmetric.
- Thus: your matching strategy determines the estimand.

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Testing the "Overlap" Assumption

- Overlap (common support) requires:

$$0 < P(d = 1 \mid X = x) < 1 \quad \forall x.$$

- **If matching on a single covariate or on the propensity score:**
  - Plot the distribution of X (or $ps(X)$) for treated vs. control units.
  - Look for regions where one group has no support.
- **If matching on multiple covariates:**
  - Examine how "far apart" matched pairs are along each covariate.
  - For each covariate x, identify the worst matches:

$$\frac{|x_i - x_{l(i)}|}{\mathrm{sd}(x)}.$$

  - Large standardized differences $\rightarrow$ poor overlap for that variable.
- Goal:
  - Ensure there exist plausible counterparts for each treated unit.
  - Detect whether the matching estimator is extrapolating.

If There Is Lack of "Overlap"

- Lack of overlap means some treated observations have no comparable controls (or vice-versa).
- Remedies are somewhat subjective but widely used:
  - **Trim or discard** units with no good match ("common support trimming").
  - Switch to **caliper matching** to avoid poor matches.
  - Use **propensity score matching** or **IPW** to reduce dimensionality.
  - Consider redefining the estimand: perhaps the ATE is not identified, but the ATT is.
- Key principle:
  - Better to estimate a credible effect for a smaller population than a biased ATE for the full sample.

Testing the "Unconfoundedness" Assumption

- Unconfoundedness requires:

$$(y(1), y(0)) \perp\!\!\!\perp d \mid X.$$

- **This is fundamentally untestable** because:
  - We never observe both potential outcomes for any unit.
  - We never observe the error term u.
  - Therefore, we cannot check whether treatment is independent of unobserved determinants of outcomes.
- Thus:
  - Any matching or propensity-score estimate is causal only **if the assumption is believed to hold**.
  - No statistical test can confirm this.

But There Are Other Things to Try...

- Although unconfoundedness is untestable, we can perform robustness and falsification checks analogous to natural experiments:

- **1. Timing tests**
  - Effects should appear only after treatment, not before.
  - A pre-trend or placebo effect suggests selection bias.

- **2. "Placebo" outcomes**
  - Test whether treatment affects variables that should not change.
  - Significant effects imply omitted variables or selection.

- **3. Heterogeneity tests**
  - Look at subsamples where theory predicts stronger or weaker effects.
  - If patterns are inconsistent with theory, matching may be invalid.

- These tests do *not prove* unconfoundedness, but they increase confidence in the credibility of the design.

Outline

Introduction to matching
- Comparison to OLS regression
- Key Limitation

How to do matching
- Notation & assumptions
- Matching on Covariates
- Matching on Propensity Score

Practical considerations

Testing the assumptions

Key weaknesses and uses of matching

Weaknesses Reiterated [Part 1]

- Matching requires many subjective researcher choices:
  - choice of distance metric,
  - number of matches,
  - calipers,
  - replacement vs. no replacement,
  - covariate set,
  - choice of propensity score model (Logit? Probit? ML?).
- Different seemingly "reasonable" choices may yield different matches $\rightarrow$ **and therefore different treatment effect estimates**.
- In practice, matching methods lack the clear empirical guidance provided by approaches based on clean quasi-experimental variation (RD, IV, DiD).

Weaknesses Reiterated [Part 2]

- Matching does **not** solve the fundamental identification problem:
  - It does **not** solve simultaneity bias.
  - It does **not** eliminate omitted variable bias from unobservables.
  - It does **not** correct for measurement error.
- Matching is simply a **control strategy**—just like OLS:
  - OLS: controls through functional form.
  - Matching: controls by local comparisons.
- OLS and matching differ mainly in how they weight comparisons, but **both require unconfoundedness** to identify causal effects.

Tangent – Related Problem

- Researchers sometimes estimate:

$$y = \beta_0 + \beta_1 d + \beta_2 ps(X) + u,$$

  where:
  - d: treatment indicator,
  - $ps(X)$: estimated probability of treatment.

- They then claim:
  "Including the propensity score controls for selection bias, so β_1 is a causal effect."

- **This claim is incorrect.** Why?

Tangent – Related Problem [Part 2]

- The assumption behind propensity scores is:

$$(y(1), y(0)) \perp\!\!\!\perp d \mid X.$$

- But the researcher regresses on **only** $ps(X)$, not on the full covariate vector X.
- Problem:
  - Controlling for $ps(X)$ in *regression* is not equivalent to matching or reweighting.
  - It does **not** eliminate bias from unobserved confounders.
  - It assumes X includes every variable that jointly affects d and y.
- Bottom line:
  - A complicated Logit does not magically create exogeneity.
  - Without unconfoundedness, neither propensity score matching nor regressions using $ps(X)$ identify causal effects.

Another Weakness – Inference

- Matching estimators often have complicated sampling distributions.
- Why inference is difficult:
  - Matching induces dependence between observations (matched units reused).
  - The matching algorithm itself creates additional randomness.
  - Analytic formulas for standard errors exist but are complex (Abadie & Imbens 2006, 2008).
- Bootstrapping often fails because matching is a non-smooth estimator.
- This makes inference less straightforward than regression or DiD.

Use as a Robustness Check

- Matching provides a **nonparametric** benchmark for OLS results:
  - It removes functional-form assumptions.
  - It forces comparisons only among similar units.
- If matching and OLS agree:
  - The OLS estimate is more credible.
- If they differ sharply:
  - OLS may be relying heavily on extrapolation or model assumptions.
- But Angrist–Pischke note:
  - With good covariates and flexible controls, differences are often small.

Use as Precursor to Regression [Part 1]

- Matching can help define a sample where **overlap holds**.
- Example:
  - Estimate the propensity score first.
  - Restrict sample to observations with:

$$0.10 < ps(X) < 0.90.$$

  - Then run OLS, DiD, or panel regressions on the trimmed sample.
- Purpose:
  - Avoids using observations with almost no comparable counterparts.
  - Ensures treatment and control units come from similar regions of the covariate space.

Use as Precursor to Regression [Part 2]

- Another example:
  - Suppose firms in Industry X experience a shock.
  - Build a control group by matching only firms with similar size, leverage, profitability, etc.
  - Then estimate treatment effects on the matched sample.
- Matching provides a principled way to construct an economically comparable control group before running the main regression.

Matching – Practical Advice

- Stata's `psmatch2` (Leuven & Sianesi) is widely used:
  - Supports nearest-neighbor matching, kernel matching, radius (caliper) matching.
  - Provides matching diagnostics and estimates standard errors using Abadie–Imbens formulas.
- In R:
  - `MatchIt`, `Matching`, `twang`, `rbounds`.
- In Python:
  - `econml`, `causalml`, `DoWhy`.
- Best practice:
  - Try multiple matching strategies.
  - Report diagnostics and balance metrics.
  - Emphasize robustness rather than "one true" estimate.

Summary of Matching

- **Matching is a control strategy.**
  - It estimates treatment effects when treatment is *as good as random* after conditioning on observable covariates X.
  - Conceptually similar to OLS with controls—but avoids imposing functional-form assumptions about how X affects y.
- **What matching does <u>not</u> do:**
  - It does **not** fix identification problems caused by:
    - simultaneity,
    - omitted unobservables,
    - measurement error,
    - reverse causality.
  - Matching only controls for the variables you actually observe (X).

Summary of Matching

- **Many ways to implement matching:**
  - Match on covariates or on the propensity score.
  - Nearest-neighbor vs. caliper/radius matching.
  - With or without replacement; different distance metrics.
  - Because choices are subjective, different methods may yield different estimates.
- **Practical value: robustness and diagnostics.**
  - Matching provides a nonparametric check on OLS estimates.
  - When covariates are rich and OLS uses flexible controls, matching and OLS typically produce similar ATE estimates.
  - Large discrepancies usually indicate model misspecification or poor overlap.