Observational Studies

Teppei Yamamoto

Keio University

Introduction to Causal Inference
Spring 2016

1. Introduction
2. Identification under Conditional Ignorability
3. Back-Door Criterion
4. Estimation under Conditional Ignorability
   - Subclassification
   - Matching
   - Weighting
5. Inference without Conditional Ignorability
   - Nonparametric Bounds
   - Sensitivity Analysis
     - Imbens' Approach
     - Rosenbaum's Approach

1. Introduction
2. Identification under Conditional Ignorability
3. Back-Door Criterion
4. Estimation under Conditional Ignorability
   - Subclassification
   - Matching
   - Weighting
5. Inference without Conditional Ignorability
   - Nonparametric Bounds
   - Sensitivity Analysis
     - Imbens' Approach
     - Rosenbaum's Approach

Essential Role of Research Design

- Causal inference requires good identification strategy
- Treatment assignment mechanism determines whether average causal effects are identifiable

1. Treatment is randomized by the researcher ☺☺☺
   - Laboratory experiments
   - Survey experiments
   - Field experiments
2. Treatment is haphazard (natural experiment) ☺☺
   - Birthdays
   - Weather
   - Close elections
   - Arbitrary administrative rules
3. Treatment is "as-if" random after statistical control ☺
   - Regression
   - Matching
4. Treatment is self-selected and no plausible control is available ☹

Causal Inference in Observational Studies

- Threat to internal validity of causal inference $=$ **confounding**
- Randomization balances both observed and unobserved confounders between the treated and untreated

Causal Inference in Observational Studies

- Threat to internal validity of causal inference $=$ **confounding**
- Randomization balances both observed and unobserved confounders between the treated and untreated
- In practice, randomized experiment is often infeasible
- And natural experiments are difficult to find

Causal Inference in Observational Studies

- Threat to internal validity of causal inference $=$ **confounding**
- Randomization balances both observed and unobserved confounders between the treated and untreated
- In practice, randomized experiment is often infeasible
- And natural experiments are difficult to find
- Practical solution: Adjust for the observed covariates and hope that unobservables are balanced

Causal Inference in Observational Studies

- Threat to internal validity of causal inference = **confounding**
- Randomization balances both observed and unobserved confounders between the treated and untreated
- In practice, randomized experiment is often infeasible
- And natural experiments are difficult to find
- Practical solution: Adjust for the observed covariates and hope that unobservables are balanced
- Better than just hoping: Do your best to design an observational study to approximate an experiment

> "The planner of an observational study should always ask himself: How would the study be conducted if it were possible to do it by controlled experimentation." (Cochran 1965)

What Makes a Good Observational Study?

Treatments, Covariates, Outcomes

- Randomized Experiment:
  - well-defined treatment
  - clear distinction between covariates and outcomes
- Good Observational Study:
  - well-defined treatment
  - clear distinction between covariates and outcomes
- Bad Observational Study:
  - hard to say when treatment began or what the treatment really is
  - distinction between covariates and outcomes is blurred
  - problems that arise in experiments seem to be avoided but are in fact just ignored

What Makes a Good Observational Study?

How were treatments assigned?

- Randomized Experiment:
  - random assignment
- Good Observational Study:
  - circumstances for the study were chosen so that treatment seems haphazard, or at least not obviously related to potential outcomes (i.e. natural or quasi-experiments)
  - there is objective evidence that treatment assignment was a function of known observed pre-treatment covariates (e.g. administrative rules)
- Bad Observational Study:
  - no attention given to assignment process
  - units self-select into treatment based on potential outcomes

What Makes a Good Observational Study?

Were treated and controls comparable?

- Randomized Experiment:
  - balance table for observables
- Good Observational Study:
  - balance table for observables
  - sensitivity analysis for unobservables
- Bad Observational Study:
  - no direct assessment of comparability is presented

What Makes a Good Observational Study?

Eliminating plausible alternatives to treatment effects?

- Randomized Experiment:
  - list plausible alternatives
  - experimental design includes features that shed light on these alternatives (e.g. placebos)
  - report on potential attrition and non-compliance
- Good Observational Study:
  - list plausible alternatives
  - study design includes features that shed light on these alternatives (e.g. multiple control groups, longitudinal covariate data, etc.)
  - requires more work than in experiment since there are usually many more alternatives
- Bad Observational Study:
  - alternatives are mentioned in discussion section of the paper

Pre-treatment Covariates

- Units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$
- Potential outcomes: $Y_i(d)$, where $d = 0, 1$

Pre-treatment Covariates

- Units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$
- Potential outcomes: $Y_i(d)$, where $d = 0, 1$
- Quantities of interest:

$$\begin{aligned} \text{ATE:} \quad \tau_{ATE} &\equiv \mathbb{E}[Y_i(1) - Y_i(0)] \\ \text{ATT:} \quad \tau_{ATT} &\equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \end{aligned}$$

Question: Can we identify τ_{ATE} and τ_{ATT} when D_i is not randomized?

Pre-treatment Covariates

- Units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$
- Potential outcomes: $Y_i(d)$, where $d = 0, 1$
- Quantities of interest:

$$\begin{aligned} \text{ATE:} \quad \tau_{ATE} &\equiv \mathbb{E}[Y_i(1) - Y_i(0)] \\ \text{ATT:} \quad \tau_{ATT} &\equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \end{aligned}$$

  Question: Can we identify τ_{ATE} and τ_{ATT} when D_i is not randomized?

- Pre-treatment covariates: $X_i = [X_{i1}, ..., X_{iK}]^\top \in \mathcal{X}$
  - Predetermined and causally precedent with respect to D_i
  - Examples: Sex, race, age, etc.
  - X_i may be correlated with both D_i and $Y_i(d)$, thereby confounding the causal relationship

Pre-treatment Covariates

- Units: $i = 1, ..., n$
- Treatment: $D_i \in \{0, 1\}$
- Potential outcomes: $Y_i(d)$, where $d = 0, 1$
- Quantities of interest:

$$\begin{aligned} \text{ATE:} \quad \tau_{ATE} &\equiv \mathbb{E}[Y_i(1) - Y_i(0)] \\ \text{ATT:} \quad \tau_{ATT} &\equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \end{aligned}$$

  Question: Can we identify τ_{ATE} and τ_{ATT} when D_i is not randomized?

- Pre-treatment covariates: $X_i = [X_{i1}, ..., X_{iK}]^\top \in \mathcal{X}$
  - Predetermined and causally precedent with respect to D_i
  - Examples: Sex, race, age, etc.
  - X_i may be correlated with both D_i and $Y_i(d)$, thereby confounding the causal relationship
  - Excludes correlates that are potentially affected by D_i (post-treatment covariates)

Conditional Ignorability

- In randomized experiments, D_i satisfies:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i$$

Conditional Ignorability

- In randomized experiments, D_i satisfies:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i$$

- Instead, we make the conditional ignorability (CI) assumption:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid X_i = x \quad \text{for any} \quad x \in \mathcal{X}$$

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variable, etc.)

Read: Among units with identical values of X_i, D_i is "as-if" randomly assigned.

Conditional Ignorability

- In randomized experiments, D_i satisfies:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i$$

- Instead, we make the conditional ignorability (CI) assumption:

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid X_i = x \quad \text{for any} \quad x \in \mathcal{X}$$

(a.k.a. exogeneity, unconfoundedness, selection on observables, no omitted variable, etc.)

Read: Among units with identical values of X_i, D_i is "as-if" randomly assigned.

- We also need the common support (a.k.a. positivity) assumption:

$$0 < \Pr(D_i = 1 \mid X_i = x) < 1 \quad \text{for any} \quad x \in \mathcal{X}$$

Read: With any value of X_i, unit could have received either treatment or control.

Identification of ATE with Regression Function

- In randomized experiments, we considered identification with population difference in means:

$$\hat{\tau} = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

Identification of ATE with Regression Function

- In randomized experiments, we considered identification with population difference in means:

$$\hat{\tau} \;=\; \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

- Here, we use difference in the population regression functions:

$$\hat{\tau}(x) \;=\; \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

Identification of ATE with Regression Function

- In randomized experiments, we considered identification with population difference in means:

$$\hat{\tau} \;=\; \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

- Here, we use difference in the population regression functions:

$$\hat{\tau}(x) \;=\; \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

- Result: Under the conditional ignorability and common support assumptions, τ is nonparametrically identified as

$$\begin{aligned} \tau_{ATE} &= \mathbb{E}[\hat{\tau}(X_i)] \\ &= \int \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} f(x) dx \end{aligned}$$

where the first $\mathbb{E}$ is taken with respect to the distribution of X_i, $f(x)$.

Proof of Identification Formula

Given the conditional ignorability assumption, we have

$$\begin{aligned}\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i] &= \mathbb{E}[Y_i(1) \mid X_i, D_i = 1] - \mathbb{E}[Y_i(0) \mid X_i, D_i = 0] \\ &= \mathbb{E}[Y_i \mid X_i, D_i = 1] - \mathbb{E}[Y_i \mid X_i, D_i = 0]\end{aligned}$$

Therefore, under the common support assumption,

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i]] \quad \text{(law of iterated expectation)} \\ &= \int \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] f(x) dx \quad \text{(definition of } \mathbb{E}\text{)} \\ &= \int \{\mathbb{E}[Y_i \mid X_i = x, D_i = 1] - \mathbb{E}[Y_i \mid X_i = x, D_i = 0]\} f(x) dx \\ &= \mathbb{E}[\hat{\tau}(x)].\end{aligned}$$

Proof of Identification Formula

Given the conditional ignorability assumption, we have

$$\begin{aligned}\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i] &= \mathbb{E}[Y_i(1) \mid X_i, D_i = 1] - \mathbb{E}[Y_i(0) \mid X_i, D_i = 0] \\ &= \mathbb{E}[Y_i \mid X_i, D_i = 1] - \mathbb{E}[Y_i \mid X_i, D_i = 0]\end{aligned}$$

Therefore, under the common support assumption,

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i]] \quad \text{(law of iterated expectation)} \\ &= \int \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] f(x) dx \quad \text{(definition of } \mathbb{E}\text{)} \\ &= \int \{\mathbb{E}[Y_i \mid X_i = x, D_i = 1] - \mathbb{E}[Y_i \mid X_i = x, D_i = 0]\} f(x) dx \\ &= \mathbb{E}[\hat{\tau}(x)].\end{aligned}$$

N.B.: This result holds regardless of the true form of the population regression function (i.e. nonparametric identification).

Identification of ATT with Regression Function

By the similar logic, τ_{ATT} is also nonparametrically identified under the conditional ignorability and common support assumptions as:

$$\tau_{ATT} = \mathbb{E}[\hat{\tau}(X_i) \mid D_i = 1]$$

where $\mathbb{E}$ is taken with respect to the distribution of X_i given $D_i = 1$.

Identification of ATT with Regression Function

By the similar logic, τ_{ATT} is also nonparametrically identified under the conditional ignorability and common support assumptions as:

$$\tau_{ATT} \quad = \quad \mathbb{E}[\hat{\tau}(X_i) \mid D_i = 1]$$

where $\mathbb{E}$ is taken with respect to the distribution of X_i given $D_i = 1$.

Proof also proceeds the same way:

$$\begin{aligned}
\tau_{ATT} &= \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i, D_i = 1] \mid D_i = 1] \quad \text{(LIE)} \\
&= \int \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x, D_i = 1] f(x \mid D_i = 1) dx \quad \text{(def. of } \mathbb{E}) \\
&= \int \{\mathbb{E}[Y_i \mid X_i = x, D_i = 1] - \mathbb{E}[Y_i \mid X_i = x, D_i = 0]\} f(x \mid D_i = 1) dx \\
&= \mathbb{E}[\hat{\tau}(x) \mid D_i = 1].
\end{aligned}$$

Identification of ATT with Regression Function

By the similar logic, τ_{ATT} is also nonparametrically identified under the conditional ignorability and common support assumptions as:

$$\tau_{ATT} = \mathbb{E}[\hat{\tau}(X_i) \mid D_i = 1]$$

where $\mathbb{E}$ is taken with respect to the distribution of X_i given $D_i = 1$.

Proof also proceeds the same way:

$$\begin{aligned}
\tau_{ATT} &= \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i, D_i = 1] \mid D_i = 1] \quad \text{(LIE)} \\
&= \int \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x, D_i = 1] f(x \mid D_i = 1) dx \quad \text{(def. of } \mathbb{E}) \\
&= \int \{\mathbb{E}[Y_i \mid X_i = x, D_i = 1] - \mathbb{E}[Y_i \mid X_i = x, D_i = 0]\} f(x \mid D_i = 1) dx \\
&= \mathbb{E}[\hat{\tau}(x) \mid D_i = 1].
\end{aligned}$$

Is $\tau_{ATE} = \tau_{ATT}$ when CI holds?

Identification of ATT with Regression Function

By the similar logic, τ_{ATT} is also nonparametrically identified under the conditional ignorability and common support assumptions as:

$$\tau_{ATT} = \mathbb{E}[\hat{\tau}(X_i) \mid D_i = 1]$$

where $\mathbb{E}$ is taken with respect to the distribution of X_i given $D_i = 1$.

Proof also proceeds the same way:

$$\begin{aligned}
\tau_{ATT} &= \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \\
&= \mathbb{E}[\mathbb{E}[Y_i(1) - Y_i(0) \mid X_i, D_i = 1] \mid D_i = 1] \quad \text{(LIE)} \\
&= \int \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x, D_i = 1] f(x \mid D_i = 1) dx \quad \text{(def. of } \mathbb{E}) \\
&= \int \{\mathbb{E}[Y_i \mid X_i = x, D_i = 1] - \mathbb{E}[Y_i \mid X_i = x, D_i = 0]\} f(x \mid D_i = 1) dx \\
&= \mathbb{E}[\hat{\tau}(x) \mid D_i = 1].
\end{aligned}$$

Is $\tau_{ATE} = \tau_{ATT}$ when CI holds? No, because $Y_i(d)$ and D_i are correlated without conditioning on X_i.

Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, S_i?

Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, S_i?
- Consider our formula for ATE, except we condition on S_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{S_i}[\mathbb{E}[Y_i \mid D_i = 1, S_i] - \mathbb{E}[Y_i \mid D_i = 0, S_i]]$$

Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, S_i?
- Consider our formula for ATE, except we condition on S_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{S_i}[\mathbb{E}[Y_i \mid D_i = 1, S_i] - \mathbb{E}[Y_i \mid D_i = 0, S_i]]$$

- Because S_i is potentially affected by the treatment, the *observed* post-treatment covariate only equals one of its *potential* value:

$$S_i = D_i S_i(1) + (1 - D_i) S_i(0)$$

Therefore, we have a mismatch problem:

$$\tilde{\tau} = \mathbb{E}_{S_i}[\mathbb{E}[Y_i(1) \mid D_i = 1, S_i(1)]] - \mathbb{E}_{S_i}[\mathbb{E}[Y_i(0) \mid D_i = 0, S_i(0)]] \neq \tau_{ATE}$$

Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, S_i?
- Consider our formula for ATE, except we condition on S_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{S_i}[\mathbb{E}[Y_i \mid D_i = 1, S_i] - \mathbb{E}[Y_i \mid D_i = 0, S_i]]$$

- Because S_i is potentially affected by the treatment, the *observed* post-treatment covariate only equals one of its *potential* value:

$$S_i = D_i S_i(1) + (1 - D_i) S_i(0)$$

  Therefore, we have a mismatch problem:

$$\tilde{\tau} = \mathbb{E}_{S_i}[\mathbb{E}[Y_i(1) \mid D_i = 1, S_i(1)]] - \mathbb{E}_{S_i}[\mathbb{E}[Y_i(0) \mid D_i = 0, S_i(0)]] \neq \tau_{ATE}$$

- This implies $\tilde{\tau} = \tau_{ATE}$ if $f(S_i) = f(S_i(1)) = f(S_i(0))$.
  This would be true if

Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, S_i?
- Consider our formula for ATE, except we condition on S_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{S_i}[\mathbb{E}[Y_i \mid D_i = 1, S_i] - \mathbb{E}[Y_i \mid D_i = 0, S_i]]$$

- Because S_i is potentially affected by the treatment, the *observed* post-treatment covariate only equals one of its *potential* value:

$$S_i = D_i S_i(1) + (1 - D_i) S_i(0)$$

  Therefore, we have a mismatch problem:

$$\tilde{\tau} = \mathbb{E}_{S_i}[\mathbb{E}[Y_i(1) \mid D_i = 1, S_i(1)]] - \mathbb{E}_{S_i}[\mathbb{E}[Y_i(0) \mid D_i = 0, S_i(0)]] \neq \tau_{ATE}$$

- This implies $\tilde{\tau} = \tau_{ATE}$ if $f(S_i) = f(S_i(1)) = f(S_i(0))$.
  This would be true if D_i has no effect on S_i, but then we would never think of controlling for S_i in the first place!

Post-Treatment Bias

- Why should we **not** condition on post-treatment covariates, S_i?
- Consider our formula for ATE, except we condition on S_i instead of X_i:

$$\tilde{\tau} \equiv \mathbb{E}_{S_i}[\mathbb{E}[Y_i \mid D_i = 1, S_i] - \mathbb{E}[Y_i \mid D_i = 0, S_i]]$$

- Because S_i is potentially affected by the treatment, the *observed* post-treatment covariate only equals one of its *potential* value:

$$S_i = D_i S_i(1) + (1 - D_i) S_i(0)$$

  Therefore, we have a mismatch problem:

$$\tilde{\tau} = \mathbb{E}_{S_i}[\mathbb{E}[Y_i(1) \mid D_i = 1, S_i(1)]] - \mathbb{E}_{S_i}[\mathbb{E}[Y_i(0) \mid D_i = 0, S_i(0)]] \neq \tau_{ATE}$$

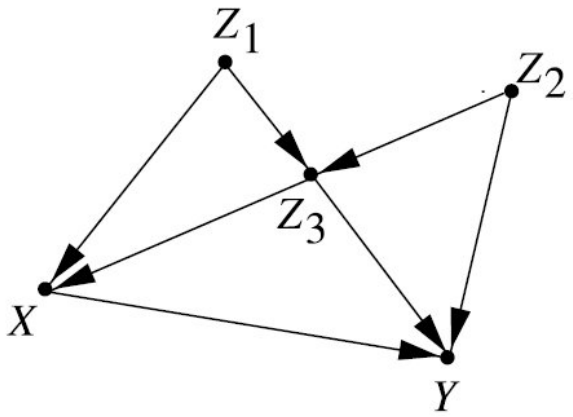
- This implies $\tilde{\tau} = \tau_{ATE}$ if $f(S_i) = f(S_i(1)) = f(S_i(0))$.
  This would be true if D_i has no effect on S_i, but then we would never think of controlling for S_i in the first place!

- A better way to think of a post-treatment covariate is mediation, a topic we will come back to later.

1. Introduction
2. Identification under Conditional Ignorability
3. Back-Door Criterion
4. Estimation under Conditional Ignorability
   - Subclassification
   - Matching
   - Weighting
5. Inference without Conditional Ignorability
   - Nonparametric Bounds
   - Sensitivity Analysis
     - Imbens' Approach
     - Rosenbaum's Approach

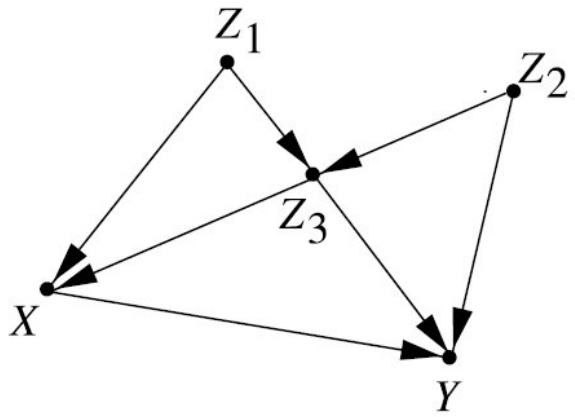
Identification Analysis with Causal Graphs

- An alternative, perhaps more intuitive, way to think about confounding is in terms of DAGs
- Suppose we want to estimate the ATE of X on Y; which covariates do we need to measure?



Identification Analysis with Causal Graphs

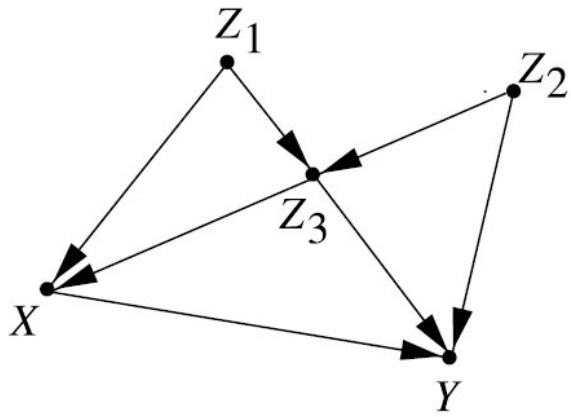
- An alternative, perhaps more intuitive, way to think about confounding is in terms of DAGs
- Suppose we want to estimate the ATE of X on Y; which covariates do we need to measure?



- Pearl develops criteria, which can be directly read off of the graph alone.
- Before studying the criteria we need to define some new concepts.

Identification Analysis with Causal Graphs

- An alternative, perhaps more intuitive, way to think about confounding is in terms of DAGs
- Suppose we want to estimate the ATE of X on Y; which covariates do we need to measure?



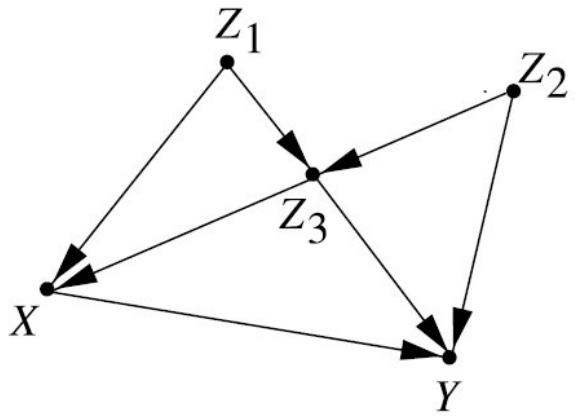
- Pearl develops criteria, which can be directly read off of the graph alone.
- Before studying the criteria we need to define some new concepts.

Review of concepts:

- Nodes: X, Y, Z_1, Z_2 and Z_3.
- Paths: "$X \to Y$", "$X \leftarrow Z_3 \to Y$", "$X \leftarrow Z_1 \to Z_3 \leftarrow Z_2 \to Y$", etc.

Identification Analysis with Causal Graphs

- An alternative, perhaps more intuitive, way to think about confounding is in terms of DAGs
- Suppose we want to estimate the ATE of X on Y; which covariates do we need to measure?



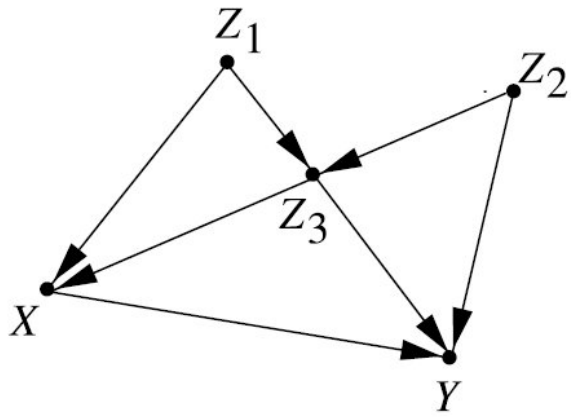
- Pearl develops criteria, which can be directly read off of the graph alone.
- Before studying the criteria we need to define some new concepts.

Review of concepts:

- Nodes: X, Y, Z_1, Z_2 and Z_3.
- Paths: "$X \to Y$", "$X \leftarrow Z_3 \to Y$", "$X \leftarrow Z_1 \to Z_3 \leftarrow Z_2 \to Y$", etc.
- Z_1 is a parent of

Identification Analysis with Causal Graphs

- An alternative, perhaps more intuitive, way to think about confounding is in terms of DAGs
- Suppose we want to estimate the ATE of X on Y; which covariates do we need to measure?



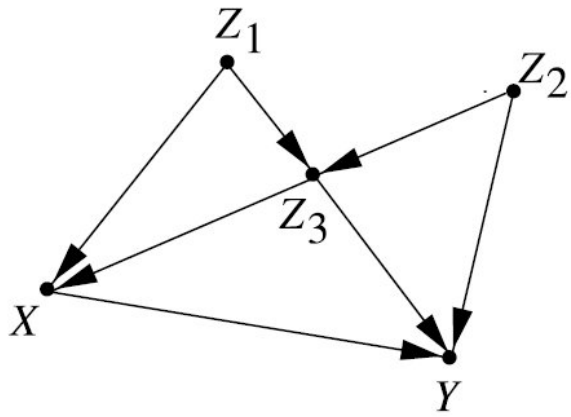
- Pearl develops criteria, which can be directly read off of the graph alone.
- Before studying the criteria we need to define some new concepts.

Review of concepts:

- Nodes: X, Y, Z_1, Z_2 and Z_3.
- Paths: "$X \to Y$", "$X \leftarrow Z_3 \to Y$", "$X \leftarrow Z_1 \to Z_3 \leftarrow Z_2 \to Y$", etc.
- Z_1 is a parent of X and Z_3. X and Z_3 are children of Z_1.
- Z_1 is an ancestor of

Identification Analysis with Causal Graphs

- An alternative, perhaps more intuitive, way to think about confounding is in terms of DAGs
- Suppose we want to estimate the ATE of X on Y; which covariates do we need to measure?



- Pearl develops criteria, which can be directly read off of the graph alone.
- Before studying the criteria we need to define some new concepts.

Review of concepts:

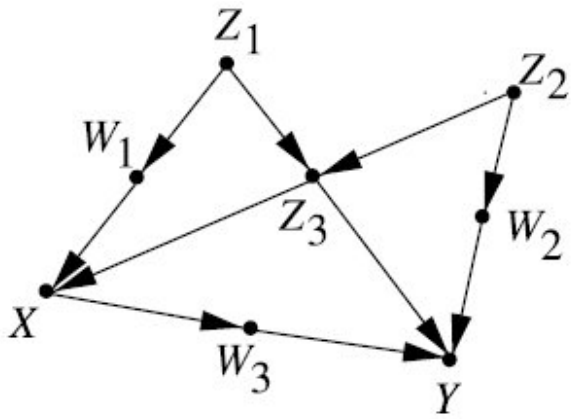
- Nodes: X, Y, Z_1, Z_2 and Z_3.
- Paths: "$X \to Y$", "$X \leftarrow Z_3 \to Y$", "$X \leftarrow Z_1 \to Z_3 \leftarrow Z_2 \to Y$", etc.
- Z_1 is a parent of X and Z_3. X and Z_3 are children of Z_1.
- Z_1 is an ancestor of Y. Y is a descendant of Z_1.

Blocked Paths and *d*-separation

Definition (blocked paths)

A set of nodes S blocks a path p if either

1. p contains at least one *arrow-emitting node* in S, or
2. p contains at least one *collision node* that is outside S and has no descendant in S.

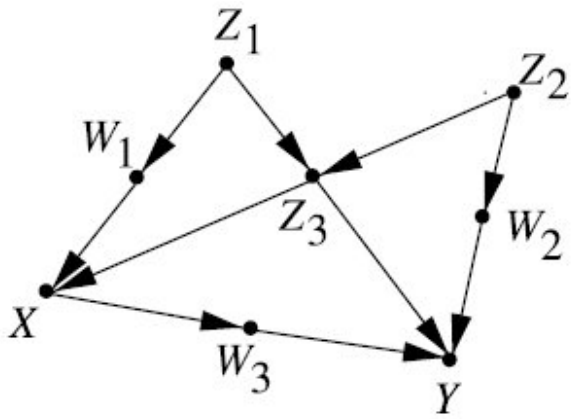


Blocked Paths and *d*-separation

Definition (blocked paths)

A set of nodes S blocks a path p if either

1. p contains at least one *arrow-emitting node* in S, or
2. p contains at least one *collision node* that is outside S and has no descendant in S.



Examples:

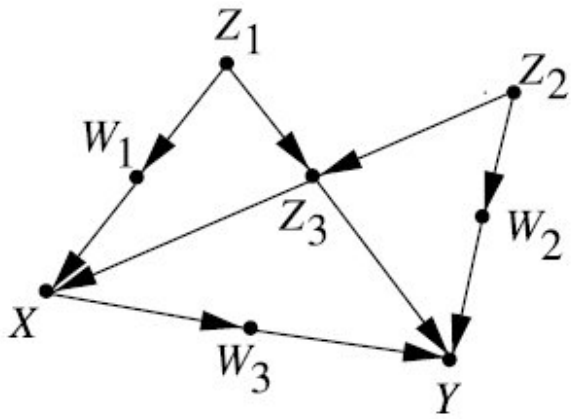
"$X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$" is blocked by

Blocked Paths and *d*-separation

Definition (blocked paths)

A set of nodes S blocks a path p if either

1. p contains at least one *arrow-emitting node* in S, or
2. p contains at least one *collision node* that is outside S and has no descendant in S.



Examples:

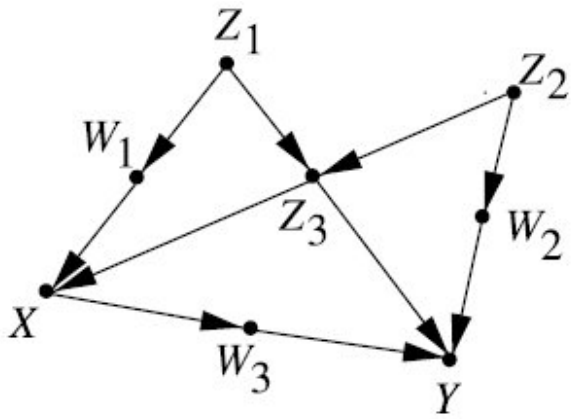
"$X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$" is blocked by $\{W_1\}$, $\{Z_1\}$, $\{Z_1, Z_3\}$, etc.

Blocked Paths and *d*-separation

Definition (blocked paths)

A set of nodes S blocks a path p if either

1. p contains at least one *arrow-emitting node* in S, or
2. p contains at least one *collision node* that is outside S and has no descendant in S.



Examples:

"$X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$" is blocked by $\{W_1\}$, $\{Z_1\}$, $\{Z_1, Z_3\}$, etc.

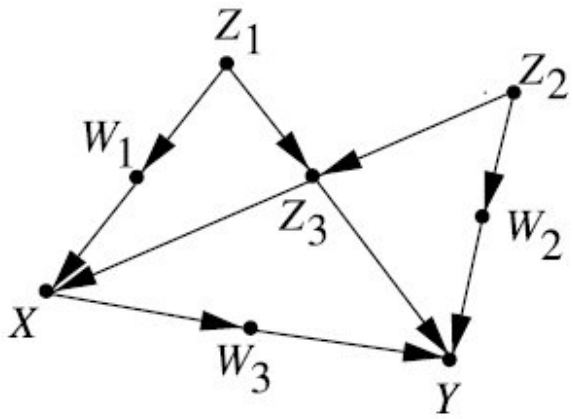
"$X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$" is blocked by

Blocked Paths and *d*-separation

Definition (blocked paths)

A set of nodes S blocks a path p if either

1. p contains at least one *arrow-emitting node* in S, or
2. p contains at least one *collision node* that is outside S and has no descendant in S.



Examples:

"$X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$" is blocked by $\{W_1\}$, $\{Z_1\}$, $\{Z_1, Z_3\}$, etc.

"$X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$" is blocked by $\{\emptyset\}$, an empty set.

Definition (*d*-separation)

If S blocks all paths from X to Y, then S *d*-separates X and Y.

If S *d*-separates X and Y, then $X \perp\!\!\!\perp Y \mid S$.

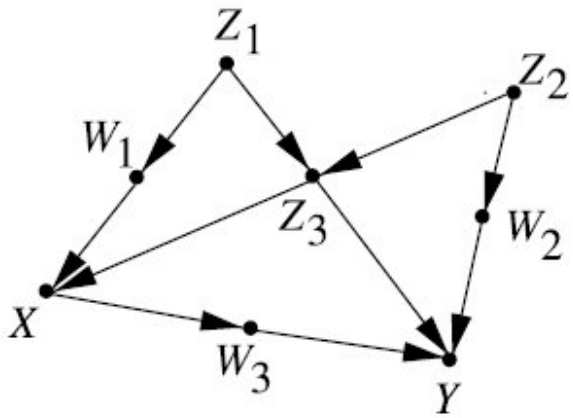
Example: W_1 and Z_3 are *d*-separated by set $S =$

Blocked Paths and *d*-separation

Definition (blocked paths)

A set of nodes S blocks a path p if either

1. p contains at least one *arrow-emitting node* in S, or
2. p contains at least one *collision node* that is outside S and has no descendant in S.



Examples:

"$X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$" is blocked by $\{W_1\}$, $\{Z_1\}$, $\{Z_1, Z_3\}$, etc.

"$X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$" is blocked by $\{\emptyset\}$, an empty set.

Definition (*d*-separation)

If S blocks all paths from X to Y, then S *d*-separates X and Y.

If S *d*-separates X and Y, then $X \perp\!\!\!\perp Y \mid S$.

Example: W_1 and Z_3 are *d*-separated by set $S = \{Z_1\}$.

The Back-Door Criterion for Causal Identification

The correspondence between d-separation and conditional independence leads to the following powerful theorem:

Theorem (the back-door criterion)

A set S is sufficient for adjustment to identify the causal effect of X on Y if:

1. *No element of S is a descendant of X,* **and**
2. *The elements of S block all back-door paths from X to Y*

The back-door criterion tells you which covariates to condition on in order to identify a causal effect, given a hypothesized DAG.

The Back-Door Criterion for Causal Identification

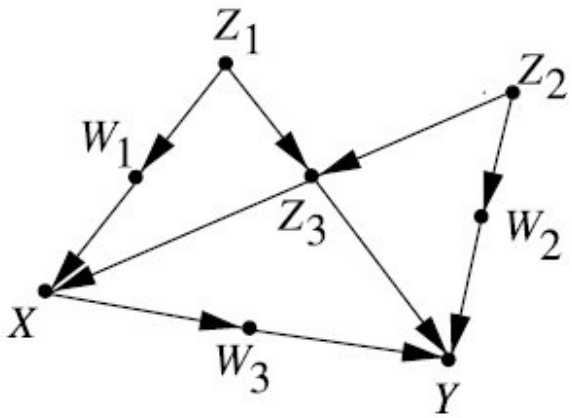
The correspondence between d-separation and conditional independence leads to the following powerful theorem:

Theorem (the back-door criterion)

A set S is sufficient for adjustment to identify the causal effect of X on Y if:

1. *No element of S is a descendant of X*, **and**
2. *The elements of S block all back-door paths from X to Y*

The back-door criterion tells you which covariates to condition on in order to identify a causal effect, given a hypothesized DAG.



Example: Which variables should we control for to identify the effect of X on Y?

- $S = \{W_1, W_2\}$?

The Back-Door Criterion for Causal Identification

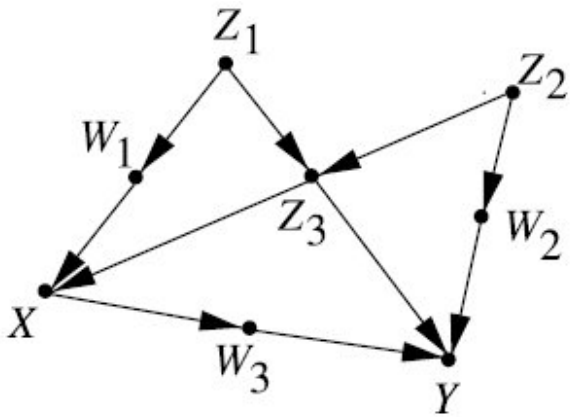
The correspondence between d-separation and conditional independence leads to the following powerful theorem:

Theorem (the back-door criterion)

A set S is sufficient for adjustment to identify the causal effect of X on Y if:

1. *No element of S is a descendant of X*, **and**
2. *The elements of S block all back-door paths from X to Y*

The back-door criterion tells you which covariates to condition on in order to identify a causal effect, given a hypothesized DAG.



Example: Which variables should we control for to identify the effect of X on Y?

- $S = \{W_1, W_2\}$? No.
- $S = \{Z_1, Z_3\}$?

The Back-Door Criterion for Causal Identification

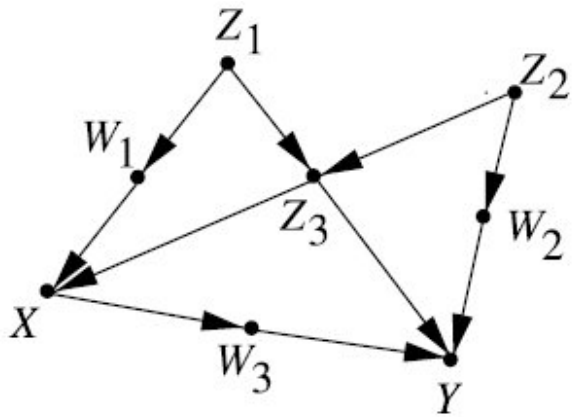
The correspondence between d-separation and conditional independence leads to the following powerful theorem:

Theorem (the back-door criterion)

A set S is sufficient for adjustment to identify the causal effect of X on Y if:

1. *No element of S is a descendant of X*, **and**
2. *The elements of S block all back-door paths from X to Y*

The back-door criterion tells you which covariates to condition on in order to identify a causal effect, given a hypothesized DAG.



Example: Which variables should we control for to identify the effect of X on Y?

- $S = \{W_1, W_2\}$? No.
- $S = \{Z_1, Z_3\}$? Yes!
- $S = \{Z_3\}$?

The Back-Door Criterion for Causal Identification

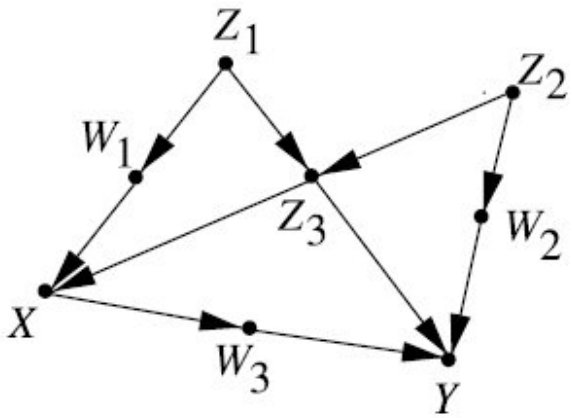
The correspondence between d-separation and conditional independence leads to the following powerful theorem:

Theorem (the back-door criterion)

A set S is sufficient for adjustment to identify the causal effect of X on Y if:

1. *No element of S is a descendant of X*, **and**
2. *The elements of S block all back-door paths from X to Y*

The back-door criterion tells you which covariates to condition on in order to identify a causal effect, given a hypothesized DAG.



Example: Which variables should we control for to identify the effect of X on Y?

- $S = \{W_1, W_2\}$? No.
- $S = \{Z_1, Z_3\}$? Yes!
- $S = \{Z_3\}$? No, because it unblocks $X \leftarrow W_1 \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow W_2 \rightarrow Y$.

The *M*-bias

- The graphical approach often reveals nonintuitive sources of bias, including the infamous *M*-bias.

The *M*-bias

- The graphical approach often reveals nonintuitive sources of bias, including the infamous *M*-bias.

Example:

- Suppose we only observe seat-belt usage. Should we control for it?

The *M*-bias

- The graphical approach often reveals nonintuitive sources of bias, including the infamous *M*-bias.

Example:

- Suppose we only observe seat-belt usage. Should we control for it? No!

The *M*-bias

- The graphical approach often reveals nonintuitive sources of bias, including the infamous *M*-bias.

Example:

- Suppose we only observe seat-belt usage. Should we control for it? No!
- Suppose we don't observe any variable other than enrollment and smoking. Are we screwed?

The *M*-bias

- The graphical approach often reveals nonintuitive sources of bias, including the infamous *M*-bias.

Example:

- Suppose we only observe seat-belt usage. Should we control for it? No!
- Suppose we don't observe any variable other than enrollment and smoking. Are we screwed? No!

The *M*-bias

- The graphical approach often reveals nonintuitive sources of bias, including the infamous *M*-bias.

Example:

- Suppose we only observe seat-belt usage. Should we control for it? No!
- Suppose we don't observe any variable other than enrollment and smoking. Are we screwed? No!
- In practice, missing arrows in this DAG encode strong assumptions that are probably not true (e.g. no common cause of Risk Aversion and X).
- What to do in practice (where we are usually uncertain about a DAG itself) is an open question.
- A standard recommendation (based on a lot of anecdotal evidence) is still to control for every observed pre-treatment covariate available.

1. Introduction
2. Identification under Conditional Ignorability
3. Back-Door Criterion
4. Estimation under Conditional Ignorability
   - Subclassification
   - Matching
   - Weighting
5. Inference without Conditional Ignorability
   - Nonparametric Bounds
   - Sensitivity Analysis
     - Imbens' Approach
     - Rosenbaum's Approach

Identification Results for Discrete Covariates

- If X_i is all discrete, the identification results can be rewritten as:

$$\tau_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

$$\tau_{ATT} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x \mid D_i = 1)$$

Identification Results for Discrete Covariates

- If X_i is all discrete, the identification results can be rewritten as:

$$\tau_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

$$\tau_{ATT} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x \mid D_i = 1)$$

- That is, τ_{ATE} can be calculated by:
  (1) Group units into strata (or cells) defined by the values of X_i.
  (2) For each stratum, calculate difference in means of Y_i between the treated and untreated.
  (3) Calculate the weighted average of (2), with weights equal to the proportions of units in the strata.

Identification Results for Discrete Covariates

- If X_i is all discrete, the identification results can be rewritten as:

$$\tau_{ATE} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x)$$

$$\tau_{ATT} = \sum_{x \in \mathcal{X}} \{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]\} \Pr(X_i = x \mid D_i = 1)$$

- That is, τ_{ATE} can be calculated by:
  (1) Group units into strata (or cells) defined by the values of X_i.
  (2) For each stratum, calculate difference in means of Y_i between the treated and untreated.
  (3) Calculate the weighted average of (2), with weights equal to the proportions of units in the strata.
- τ_{ATT} can be calculated similarly:
  (1) $\cdot$ (2) Same as (1) $\cdot$ (2) for ATE.
  (3) Calculate the weighted average of (2), with weights equal to the proportions of units in the strata within the treatment group.

Subclassification Estimators

- The sample analogues of the formulas on the previous slide are the subclassification estimators:

$$\hat{\tau}_{ATE} = \sum_{j=1}^{M} \left\{ \overline{Y}_{1j} - \overline{Y}_{0j} \right\} \frac{n_j}{n}$$

$$\hat{\tau}_{ATT} = \sum_{j=1}^{M} \left\{ \overline{Y}_{1j} - \overline{Y}_{0j} \right\} \frac{n_{1j}}{n_1}$$

where
$$\begin{cases} M = \text{\# of strata} \\ n_j = \text{\# of units in cell } j \\ n_{1j} = \text{\# of treated units in cell } j \\ \overline{Y}_{dj} = \text{mean outcome for units with } D_i = d \text{ in cell } j \end{cases}$$

Example: Smoking and Mortality (Cochran 1968)

TABLE 1
DEATH RATES PER 1,000 PERSON-YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Example: Smoking and Mortality (Cochran 1968)

TABLE 2

MEAN AGES, YEARS

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Subclassification by Age ($M = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

What is the subclassification estimator for the ATE of smoking on death rate?

Subclassification by Age ($M = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

What is the subclassification estimator for the ATE of smoking on death rate?

$$\hat{\tau}_{ATE} = (28 - 24) \cdot \frac{10}{20} + (22 - 16) \cdot \frac{10}{20} = 5$$

Subclassification by Age ($M = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

What is the subclassification estimator for the ATT of smoking on death rate?

Subclassification by Age ($M = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

What is the subclassification estimator for the ATT of smoking on death rate?

$$\hat{\tau}_{ATT} = (28 - 24) \cdot \frac{3}{10} + (22 - 16) \cdot \frac{7}{10} = 5.4$$

Subclassification by Age and Gender ($M = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate for the ATE of smoking on death rate?

Subclassification by Age and Gender ($M = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate for the ATE of smoking on death rate?

Not identified! (because of the lack of common support)

Subclassification by Age and Gender ($M = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate for the ATT of smoking on death rate?

Subclassification by Age and Gender ($M = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate for the ATT of smoking on death rate?

$$\begin{aligned} \hat{\tau}_{ATT} &= (28 - 22) \cdot \frac{3}{10} + (21 - 16) \cdot \frac{3}{10} + (23 - 17) \cdot \frac{4}{10} \\ &= 5.1 \end{aligned}$$

Summary

- Causal inference in observational studies often rests on the conditional ignorability assumption
- Goal is to approximate a randomized experiment within subgroups
- Better to have a design-based justification for conditional ignorability
- Do not control for post-treatment covariates
- A DAG can tell you what specific variables to control for, if you can draw one
- If you have a small number of discrete covariates, ATE can be estimated completely nonparametrically via subclassification

Matching

- Subclassification only works when all covariates in X_i are discrete
- An alternative when some/all of X_i are continuous: matching

Matching

- Subclassification only works when all covariates in X_i are discrete
- An alternative when some/all of X_i are continuous: matching
- Basic idea: Impute missing potential outcomes using observed outcomes of "closest" units (a.k.a. nearest neighbors)

Matching

- Subclassification only works when all covariates in X_i are discrete
- An alternative when some/all of X_i are continuous: matching
- Basic idea: Impute missing potential outcomes using observed outcomes of "closest" units (a.k.a. nearest neighbors)

1. For each observation in the treated group i, find an observation in the untreated group with the most similar values of X
2. Estimate ATT by the average difference between the pairs:

$$\hat{\tau}_{ATT} = \frac{1}{n_1} \sum_{i:D_i=1} (Y_i - \tilde{Y}_i) \simeq \frac{1}{n_1} \sum_{i:D_i=1} (Y_i(1) - Y_i(0)) = \tau_{SATT}$$

where $\tilde{Y}_i$ is the observed outcome of i's untreated "buddy"

Matching

- Subclassification only works when all covariates in X_i are discrete
- An alternative when some/all of X_i are continuous: matching
- Basic idea: Impute missing potential outcomes using observed outcomes of "closest" units (a.k.a. nearest neighbors)

1. For each observation in the treated group i, find an observation in the untreated group with the most similar values of X
2. Estimate ATT by the average difference between the pairs:

$$\hat{\tau}_{ATT} = \frac{1}{n_1} \sum_{i:D_i=1} (Y_i - \tilde{Y}_i) \simeq \frac{1}{n_1} \sum_{i:D_i=1} (Y_i(1) - Y_i(0)) = \tau_{SATT}$$

where $\tilde{Y}_i$ is the observed outcome of i's untreated "buddy"

When there are multiple (M_i) "close" units, their average can be used:

$$\hat{\tau}_{ATT} = \frac{1}{n_1} \sum_{i:D_i=1} \left\{ Y_i - \left(\frac{1}{M_i} \sum_{m=1}^{M_i} \tilde{Y}_{i_m}, \right) \right\}$$

where $\tilde{Y}_{i_m}$ is i's mth untreated buddy

Example with Single Pre-treatment Covariate

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	D_i	X_i
1	6	?	1	3
2	1	?	1	1
3	0	?	1	4
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Example with Single Pre-treatment Covariate

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	4
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Match and plug in:

Example with Single Pre-treatment Covariate

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	$Y_i(1)$	$Y_i(0)$	D_i	X_i
1	6	9	1	3
2	1	0	1	1
3	0	9	1	4
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Match and plug in:

$$\hat{\tau}_{ATT} = \frac{1}{3}\{(6-9)+(1-0)+(0-9)\} = -3.7$$

The Curse of Dimensionality

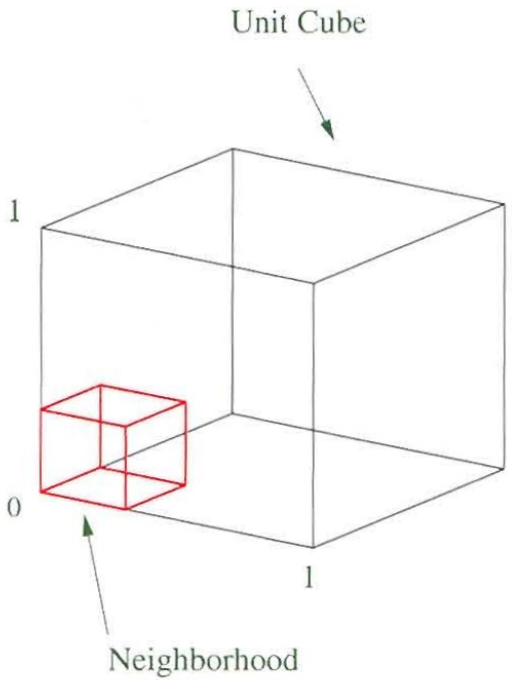
- How do we define the "closest" when X_i contains > 1 variable?

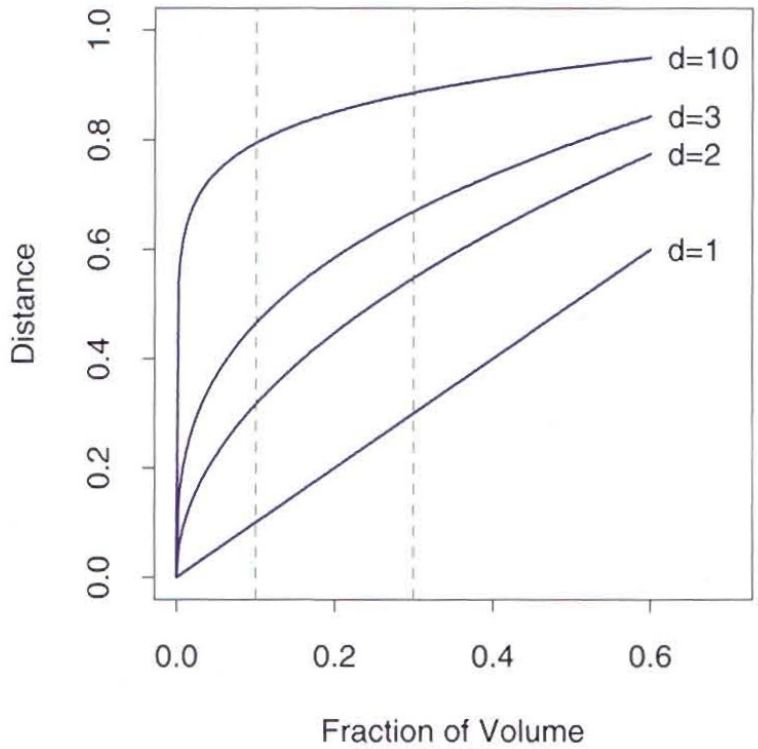
The Curse of Dimensionality

- How do we define the "closest" when X_i contains > 1 variable?
- Can we hope to exactly match on every X_{ik} if we have large n?

The Curse of Dimensionality

- How do we define the "closest" when X_i contains > 1 variable?
- Can we hope to exactly match on every X_{ik} if we have large n?
  $\Longrightarrow$ No! because of curse of dimensionality





As number of dimensions (d) in the covariate space increases, data sparsity exponentially increases for a given sample size.

Distance Metrics for Matching

- With many covariates, we can use a low-dimensional (usually scalar) distance metric:
  1. Mahalanobis distance:

$$D_M(X_i, X_j) = \sqrt{(X_i - X_j)^\top \Sigma_X^{-1} (X_i - X_j)}$$

where Σ_X is the (sample) variance-covariance matrix of X_i

Distance Metrics for Matching

- With many covariates, we can use a low-dimensional (usually scalar) distance metric:
  1. Mahalanobis distance:

$$D_M(X_i, X_j) = \sqrt{(X_i - X_j)^\top \Sigma_X^{-1} (X_i - X_j)}$$

     where Σ_X is the (sample) variance-covariance matrix of X_i

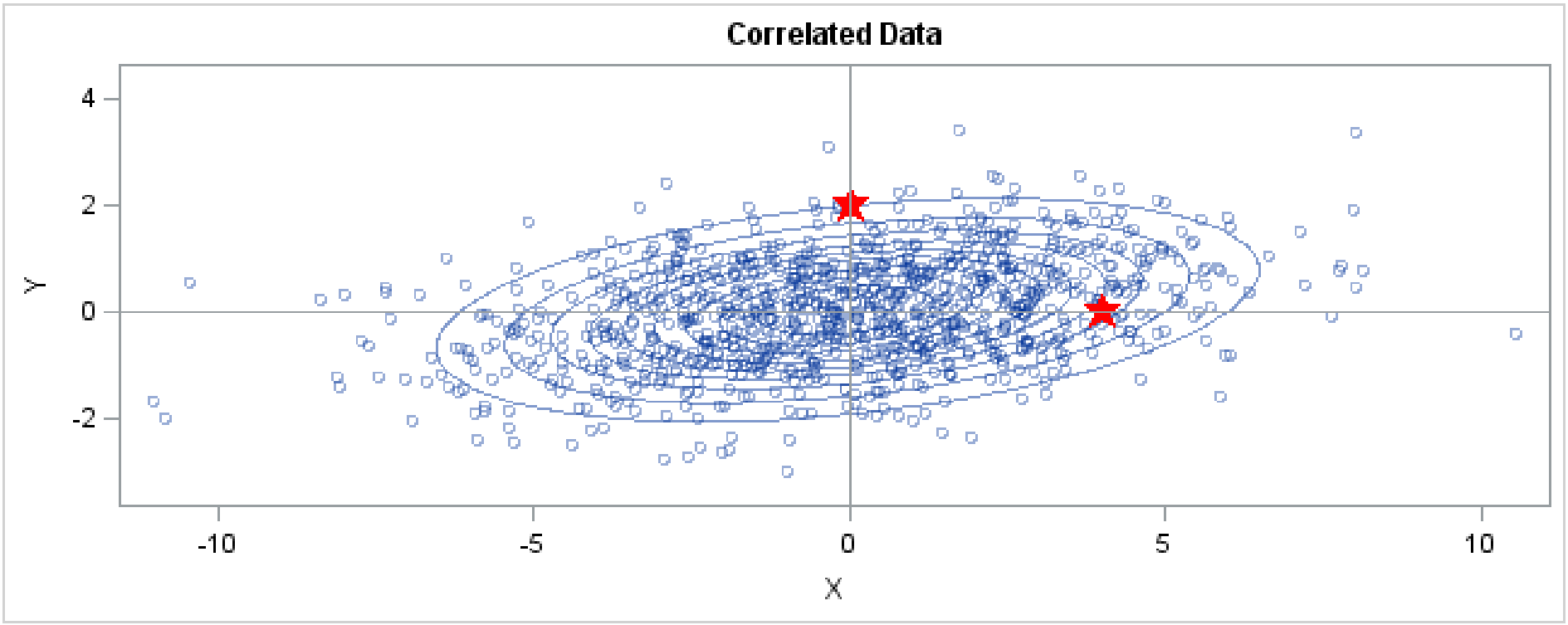
  2. Genetic matching (Diamond and Sekhon 2005; `GenMatch` in R):

$$D_{gen}(X_i, X_j) = \sqrt{(X_i - X_j)^\top \left(\Sigma_X^{-1/2}\right)^\top W \left(\Sigma_X^{-1/2}\right) (X_i - X_j)},$$

     where W is a weight matrix chosen via an optimization algorithm

  etc. (many other variants)

Mahalanobis Distance: Graphical Illustration



Which observations are closer to the origin?

Mahalanobis Distance: Numeric Example

	index	X_1	X_2
Treated	i	0	0
Control A	A	5	5
Control B	B	4	0

where $\Sigma_X = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$

Which control is closer?

Mahalanobis Distance: Numeric Example

	index	X_1	X_2
Treated	i	0	0
Control A	A	5	5
Control B	B	4	0

where $\Sigma_X = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$

Which control is closer?

$$D_M(X_i, X_A) = \sqrt{(X_i - X_j)^\top \Sigma^{-1} (X_i - X_j)}$$

$$=$$

Mahalanobis Distance: Numeric Example

	index	X_1	X_2
Treated	i	0	0
Control A	A	5	5
Control B	B	4	0

where $\Sigma_X = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$

Which control is closer?

$$
\begin{aligned}
D_M(X_i, X_A) &= \sqrt{(X_i - X_j)^\top \Sigma^{-1} (X_i - X_j)} \\
&= \sqrt{((\begin{array}{cc} 0 & 0 \end{array}) - (\begin{array}{cc} 5 & 5 \end{array})) \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}^{-1} ((\begin{array}{cc} 0 & 0 \end{array}) - (\begin{array}{cc} 5 & 5 \end{array}))^\top} \\
&= \sqrt{(\begin{array}{cc} -5 & -5 \end{array})^\top \begin{pmatrix} 5.2 & -4.7 \\ -4.7 & 5.2 \end{pmatrix} (\begin{array}{cc} -5 & -5 \end{array})} = 26
\end{aligned}
$$

$$
D_M(X_i, X_B) =
$$

Mahalanobis Distance: Numeric Example

	index	X_1	X_2
Treated	i	0	0
Control A	A	5	5
Control B	B	4	0

where $\Sigma_X = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$

Which control is closer?

$$
\begin{aligned}
D_M(X_i, X_A) &= \sqrt{(X_i - X_j)^\top \Sigma^{-1} (X_i - X_j)} \\
&= \sqrt{((\begin{array}{cc} 0 & 0 \end{array}) - (\begin{array}{cc} 5 & 5 \end{array})) \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}^{-1} ((\begin{array}{cc} 0 & 0 \end{array}) - (\begin{array}{cc} 5 & 5 \end{array}))^\top} \\
&= \sqrt{(\begin{array}{cc} -5 & -5 \end{array})^\top \begin{pmatrix} 5.2 & -4.7 \\ -4.7 & 5.2 \end{pmatrix} (\begin{array}{cc} -5 & -5 \end{array})} = 26 \\
D_M(X_i, X_B) &= \sqrt{(\begin{array}{cc} -4 & 0 \end{array}) \begin{pmatrix} 5.2 & -4.7 \\ -4.7 & 5.2 \end{pmatrix} (\begin{array}{cc} -4 & 0 \end{array})^\top} = 84
\end{aligned}
$$

Propensity Score and the Balancing Property

- Another important metric: propensity score
- Definition: Probability of receiving the treatment given X_i

$$\pi(X_i) \equiv \Pr(D_i = 1 \mid X_i)$$

Propensity Score and the Balancing Property

- Another important metric: propensity score
- Definition: Probability of receiving the treatment given X_i

$$\pi(X_i) \equiv \Pr(D_i = 1 \mid X_i)$$

- **Result**: Suppose the following assumptions hold:
  1. $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i \mid X_i$ (conditional ignorability)
  2. $0 < \Pr(D_i = 1 \mid X_i = x) < 1$ for any x (common support)

  Then, the propensity score has the balancing property:

$$D_i \perp\!\!\!\perp X_i \mid \pi(X_i)$$

  Read: Among those units with the same propensity score, X_i is identically distributed between the treated and untreated.

Proof of the Balancing Property

Recall: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

Proof of the Balancing Property

Recall: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$\begin{aligned} \Pr(D_i = 1 | \pi(X_i), X_i) &= \mathbb{E}[D_i | \pi(X_i), X_i] \\ &= \mathbb{E}[D_i | X_i] \quad (\because \end{aligned}$$

Proof of the Balancing Property

Recall: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i), X_i) &= \mathbb{E}[D_i | \pi(X_i), X_i] \\
&= \mathbb{E}[D_i | X_i] \quad (\because X_i \text{ contains all information in } \pi(X_i)) \\
&=
\end{aligned}$$

Proof of the Balancing Property

Recall: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$
\begin{aligned}
\Pr(D_i = 1 | \pi(X_i), X_i) &= \mathbb{E}[D_i | \pi(X_i), X_i] \\
&= \mathbb{E}[D_i | X_i] \quad (\because X_i \text{ contains all information in } \pi(X_i)) \\
&= \Pr(D_i = 1 \mid X_i) \; =
\end{aligned}
$$

Proof of the Balancing Property

Recall: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i), X_i) &= \mathbb{E}[D_i | \pi(X_i), X_i] \\
&= \mathbb{E}[D_i | X_i] \quad (\because X_i \text{ contains all information in } \pi(X_i)) \\
&= \Pr(D_i = 1 \mid X_i) = \pi(X_i) \text{ (by definition!)}
\end{aligned}$$

And we can also show:

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i)) &= \mathbb{E}[D_i | \pi(X_i)] \\
&= \mathbb{E}[\mathbb{E}[D_i | X_i] \mid \pi(X_i)] \quad (\because
\end{aligned}$$

Proof of the Balancing Property

Recall: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i), X_i) &= \mathbb{E}[D_i | \pi(X_i), X_i] \\
&= \mathbb{E}[D_i | X_i] \quad (\because X_i \text{ contains all information in } \pi(X_i)) \\
&= \Pr(D_i = 1 \mid X_i) \; = \; \pi(X_i) \text{ (by definition!)}
\end{aligned}$$

And we can also show:

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i)) &= \mathbb{E}[D_i | \pi(X_i)] \\
&= \mathbb{E}[\mathbb{E}[D_i | X_i] \mid \pi(X_i)] \quad (\because \text{ L. I. E.}) \\
&=
\end{aligned}$$

Proof of the Balancing Property

Recall: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i), X_i) &= \mathbb{E}[D_i | \pi(X_i), X_i] \\
&= \mathbb{E}[D_i | X_i] \quad (\because X_i \text{ contains all information in } \pi(X_i)) \\
&= \Pr(D_i = 1 \mid X_i) = \pi(X_i) \text{ (by definition!)}
\end{aligned}$$

And we can also show:

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i)) &= \mathbb{E}[D_i | \pi(X_i)] \\
&= \mathbb{E}[\mathbb{E}[D_i | X_i] \mid \pi(X_i)] \quad (\because \text{L. I. E.}) \\
&= \mathbb{E}[\pi(X_i) \mid \pi(X_i)] =
\end{aligned}$$

Proof of the Balancing Property

Recall: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i), X_i) &= \mathbb{E}[D_i | \pi(X_i), X_i] \\
&= \mathbb{E}[D_i | X_i] \quad (\because X_i \text{ contains all information in } \pi(X_i)) \\
&= \Pr(D_i = 1 \mid X_i) = \pi(X_i) \text{ (by definition!)}
\end{aligned}$$

And we can also show:

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i)) &= \mathbb{E}[D_i | \pi(X_i)] \\
&= \mathbb{E}[\mathbb{E}[D_i | X_i] \mid \pi(X_i)] \quad (\because \text{L. I. E.}) \\
&= \mathbb{E}[\pi(X_i) \mid \pi(X_i)] = \pi(X_i)
\end{aligned}$$

Proof of the Balancing Property

Recall: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i), X_i) &= \mathbb{E}[D_i | \pi(X_i), X_i] \\
&= \mathbb{E}[D_i | X_i] \quad (\because X_i \text{ contains all information in } \pi(X_i)) \\
&= \Pr(D_i = 1 \mid X_i) \; = \; \pi(X_i) \text{ (by definition!)}
\end{aligned}$$

And we can also show:

$$\begin{aligned}
\Pr(D_i = 1 | \pi(X_i)) &= \mathbb{E}[D_i | \pi(X_i)] \\
&= \mathbb{E}[\mathbb{E}[D_i | X_i] \mid \pi(X_i)] \quad (\because \text{ L. I. E.}) \\
&= \mathbb{E}[\pi(X_i) \mid \pi(X_i)] \; = \; \pi(X_i)
\end{aligned}$$

Therefore, $\Pr(D_i = 1 | \pi(X_i), X_i) = \Pr(D_i = 1 | \pi(X_i))$, which implies $D_i \perp\!\!\!\perp X_i \mid \pi(X_i)$, the balancing property. □

Identification with the Propensity Score

- The balancing property implies that conditional ignorability holds given just the propensity score:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \pi(X_i)$$

Identification with the Propensity Score

- The balancing property implies that conditional ignorability holds given just the propensity score:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \pi(X_i)$$

- Implication:

  It is sufficient to just condition on $\pi(X_i)$, instead of whole X_i!

Identification with the Propensity Score

- The balancing property implies that conditional ignorability holds given just the propensity score:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \pi(X_i)$$

- Implication:

  It is sufficient to just condition on $\pi(X_i)$, instead of whole X_i!

- Doesn't that sound awesome? Yes, but there is a catch: $\pi(X_i)$ itself needs to be estimated!

Identification with the Propensity Score

- The balancing property implies that conditional ignorability holds given just the propensity score:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \pi(X_i)$$

- Implication:

  It is sufficient to just condition on $\pi(X_i)$, instead of whole X_i!

- Doesn't that sound awesome? Yes, but there is a catch: $\pi(X_i)$ itself needs to be estimated!

- Two-step procedure to estimate causal effects:

  (1) Estimate $\pi(X_i)$ with a model for a binary response (e.g. logit, probit — details in Quant III)

  (2) Do nearest neighbor matching on $\pi(X_i)$

Proof of the Identification Result

Again: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

Proof of the Identification Result

Again: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$
\begin{aligned}
&\Pr(D_i = 1 | Y_i(1), Y_i(0), \pi(X_i)) \\
&\quad = \mathbb{E}[D_i | Y_i(1), Y_i(0), \pi(X_i)] \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | Y_i(1), Y_i(0), X_i] \mid Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because
\end{aligned}
$$

Proof of the Identification Result

Again: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$
\begin{aligned}
&\Pr(D_i = 1 | Y_i(1), Y_i(0), \pi(X_i)) \\
&\quad = \mathbb{E}[D_i | Y_i(1), Y_i(0), \pi(X_i)] \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | Y_i(1), Y_i(0), X_i] \mid Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ L. I. E.}) \\
&\quad =
\end{aligned}
$$

Proof of the Identification Result

Again: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$
\begin{aligned}
&\Pr(D_i = 1 | Y_i(1), Y_i(0), \pi(X_i)) \\
&\quad = \mathbb{E}[D_i | Y_i(1), Y_i(0), \pi(X_i)] \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | Y_i(1), Y_i(0), X_i] \mid Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ L. I. E.}) \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | X_i] | Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because
\end{aligned}
$$

Proof of the Identification Result

Again: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$
\begin{aligned}
&\Pr(D_i = 1 | Y_i(1), Y_i(0), \pi(X_i)) \\
&\quad = \mathbb{E}[D_i | Y_i(1), Y_i(0), \pi(X_i)] \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | Y_i(1), Y_i(0), X_i] \mid Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ L. I. E.}) \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | X_i] | Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ conditional ignorability}) \\
&\quad = \mathbb{E}\left[\pi(X_i) | Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because
\end{aligned}
$$

Proof of the Identification Result

Again: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$
\begin{aligned}
&\Pr(D_i = 1 | Y_i(1), Y_i(0), \pi(X_i)) \\
&\quad = \mathbb{E}[D_i | Y_i(1), Y_i(0), \pi(X_i)] \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | Y_i(1), Y_i(0), X_i] \mid Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ L. I. E.}) \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | X_i] | Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ conditional ignorability}) \\
&\quad = \mathbb{E}\left[\pi(X_i) | Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ definition of } \pi(X_i)) \\
&\quad = \pi(X_i)
\end{aligned}
$$

And in the previous proof, we have already shown:

$$\Pr(D_i = 1 | \pi(X_i)) \quad =$$

Proof of the Identification Result

Again: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$
\begin{aligned}
&\Pr(D_i = 1 | Y_i(1), Y_i(0), \pi(X_i)) \\
&\quad = \mathbb{E}[D_i | Y_i(1), Y_i(0), \pi(X_i)] \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | Y_i(1), Y_i(0), X_i] \mid Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ L. I. E.}) \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | X_i] | Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ conditional ignorability}) \\
&\quad = \mathbb{E}\left[\pi(X_i) | Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ definition of } \pi(X_i)) \\
&\quad = \pi(X_i)
\end{aligned}
$$

And in the previous proof, we have already shown:

$$
\Pr(D_i = 1 | \pi(X_i)) \quad = \quad \pi(X_i)
$$

Proof of the Identification Result

Again: To prove independence between two random variables A and B, all you need is to show that $\Pr(A \mid B) = \Pr(A)$.

$$
\begin{aligned}
&\Pr(D_i = 1 | Y_i(1), Y_i(0), \pi(X_i)) \\
&\quad = \mathbb{E}[D_i | Y_i(1), Y_i(0), \pi(X_i)] \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | Y_i(1), Y_i(0), X_i] \mid Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ L. I. E.}) \\
&\quad = \mathbb{E}\left[\mathbb{E}[D_i | X_i] | Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ conditional ignorability}) \\
&\quad = \mathbb{E}\left[\pi(X_i) | Y_i(1), Y_i(0), \pi(X_i)\right] \quad (\because \text{ definition of } \pi(X_i)) \\
&\quad = \pi(X_i)
\end{aligned}
$$

And in the previous proof, we have already shown:

$$\Pr(D_i = 1 | \pi(X_i)) \quad = \quad \pi(X_i)$$

Therefore, $\Pr(D_i = 1 | Y_i(1), Y_i(0), \pi(X_i)) = \Pr(D_i = 1 | \pi(X_i))$, which implies $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \pi(X_i)$, conditional ignorability just given $\pi(X_i)$. $\square$

Estimating the Propensity Score

- Estimation of propensity scores requires a correct specification of $\pi(X_i)$ (functional form, etc.) Any guidance?

Estimating the Propensity Score

- Estimation of propensity scores requires a correct specification of $\pi(X_i)$ (functional form, etc.) Any guidance?
- Solution: Check balance
  - Ideally, compare the joint distribution of all X_i between the treated and untreated in the matched sample
  - In practice, check various low-dimensional summaries of $F(x)$ (mean difference, variance ratio, etc.)

Estimating the Propensity Score

- Estimation of propensity scores requires a correct specification of $\pi(X_i)$ (functional form, etc.) Any guidance?
- Solution: Check balance
  - Ideally, compare the joint distribution of all X_i between the treated and untreated in the matched sample
  - In practice, check various low-dimensional summaries of $F(x)$ (mean difference, variance ratio, etc.)
  - Balance tests are often used (e.g. t-test, F-test, KS test) like the ones we saw for randomized experiments.
  - Note that balance tests can be misleading in a matching context
  - Reason: you can make everything insignificant by simply dropping lots of observations — make sure this is not happening

Estimating the Propensity Score

- Estimation of propensity scores requires a correct specification of $\pi(X_i)$ (functional form, etc.) Any guidance?
- Solution: Check balance
  - Ideally, compare the joint distribution of all X_i between the treated and untreated in the matched sample
  - In practice, check various low-dimensional summaries of $F(x)$ (mean difference, variance ratio, etc.)
  - Balance tests are often used (e.g. t-test, F-test, KS test) like the ones we saw for randomized experiments.
  - Note that balance tests can be misleading in a matching context
  - Reason: you can make everything insignificant by simply dropping lots of observations — make sure this is not happening
- Estimate $\rightarrow$ Check Balance $\rightarrow$ Re-estimate $\rightarrow$ Check Balance $\rightarrow \cdots$ (ad infinitum until you get a good balance)
- Is this data snooping? No, because inference remains blind to Y

Kolmogorov-Smirnov (KS) Test

- The KS test is used to test whether two random variables are sampled from the same distribution
- The test is nonparametric, meaning that it works (asymptotically) without assumptions about the form of the underlying distribution

Kolmogorov-Smirnov (KS) Test

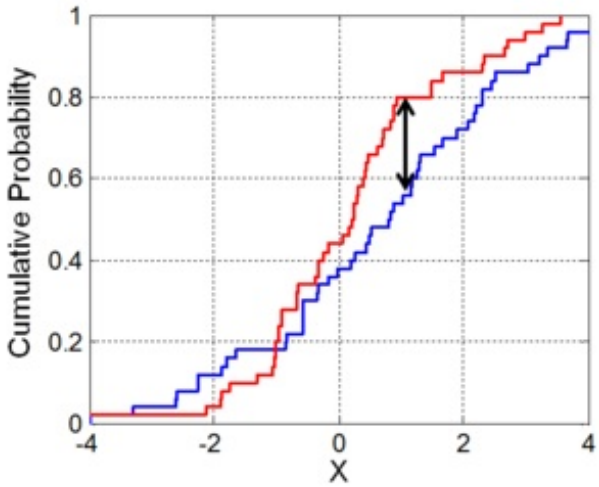
- The KS test is used to test whether two random variables are sampled from the same distribution
- The test is nonparametric, meaning that it works (asymptotically) without assumptions about the form of the underlying distribution

Consider n observations of two random variables, X_0 and X_1.

The (two-sample) KS statistic:

$$D = \sup_x \left| \widehat{F}_1(x) - \widehat{F}_0(x) \right|,$$

where $\widehat{F}_0(x)$, $\widehat{F}_1(x)$ is the empirical CDF of X_0, X_1.



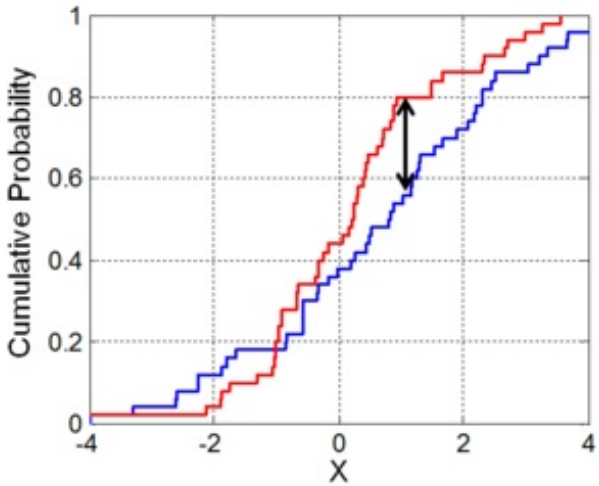
Kolmogorov-Smirnov (KS) Test

- The KS test is used to test whether two random variables are sampled from the same distribution
- The test is nonparametric, meaning that it works (asymptotically) without assumptions about the form of the underlying distribution

Consider n observations of two random variables, X_0 and X_1.

The (two-sample) KS statistic:

$$D = \sup_x \left| \widehat{F}_1(x) - \widehat{F}_0(x) \right|,$$



where $\widehat{F}_0(x)$, $\widehat{F}_1(x)$ is the empirical CDF of X_0, X_1.

The KS null hypothesis: $F_1(x) = F_0(x)$ (no difference in true distributions)

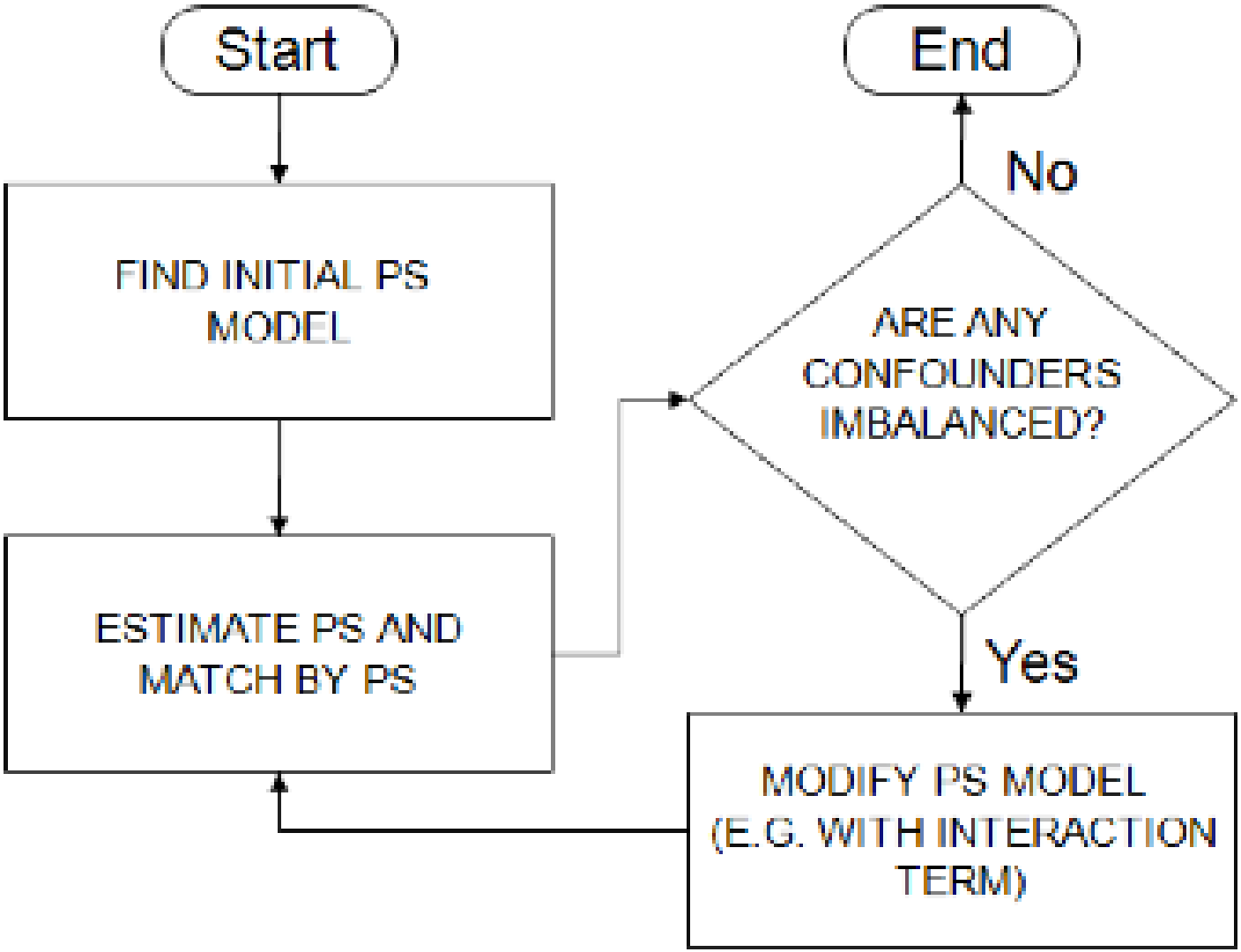
Under the null, D has the Kolmogorov distribution as $n \to \infty$.

Reject the null at level α if

$$D > c_\alpha \sqrt{(n_1 + n_0)/n_1 n_0}$$

level (α)	.1	.05	.01
critical value (c_α)	1.22	1.36	1.63

Matching Workflow



Are Coethnics More Effective Counterinsurgents?

TABLE 2. Balance Summary Statistics and Tests: Russian and Chechen Sweeps

Pretreatment Covariates	Mean Treated	Mean Control	Mean Difference	Std. Bias	Rank Sum Test	K-S Test
Demographics						
Population	8.657	8.606	0.049	0.033	0.708	0.454
Tariqa	0.076	0.048	0.028	0.104	0.331	—
Poverty	1.917	1.931	−0.016	−0.024	0.792	1.000
Spatial						
Elevation	5.078	5.233	−0.155	−0.135	0.140	0.228
Isolation	1.007	1.070	−0.063	−0.096	0.343	0.851
Groznyy	0.131	0.138	−0.007	−0.018	0.864	—
War Dynamics						
TAC	0.241	0.282	−0.041	−0.095	0.424	—
Garrison	0.379	0.414	−0.035	−0.072	0.549	—
Rebel	0.510	0.441	0.070	0.139	0.240	—
Selection						
Presweep violence	3.083	3.117	−0.034	0.009	0.454	0.292
Large-scale theft	0.034	0.055	−0.021	−0.115	0.395	—
Killing	0.117	0.090	0.027	0.084	0.443	—
Violence Inflicted						
Total abuse	0.970	0.833	0.137	0.124	0.131	0.454
Prior sweeps	1.729	1.812	−0.090	−0.089	0.394	0.367
Other						
Month	7.428	6.986	0.442	0.130	0.260	0.292
Year	2004.159	2004.110	0.049	0.043	0.889	1.000

Note: 145 matched pairs. Matching with replacement.

Lyall (2010), *American Political Science Review*.

Is SAT Coaching Effective?

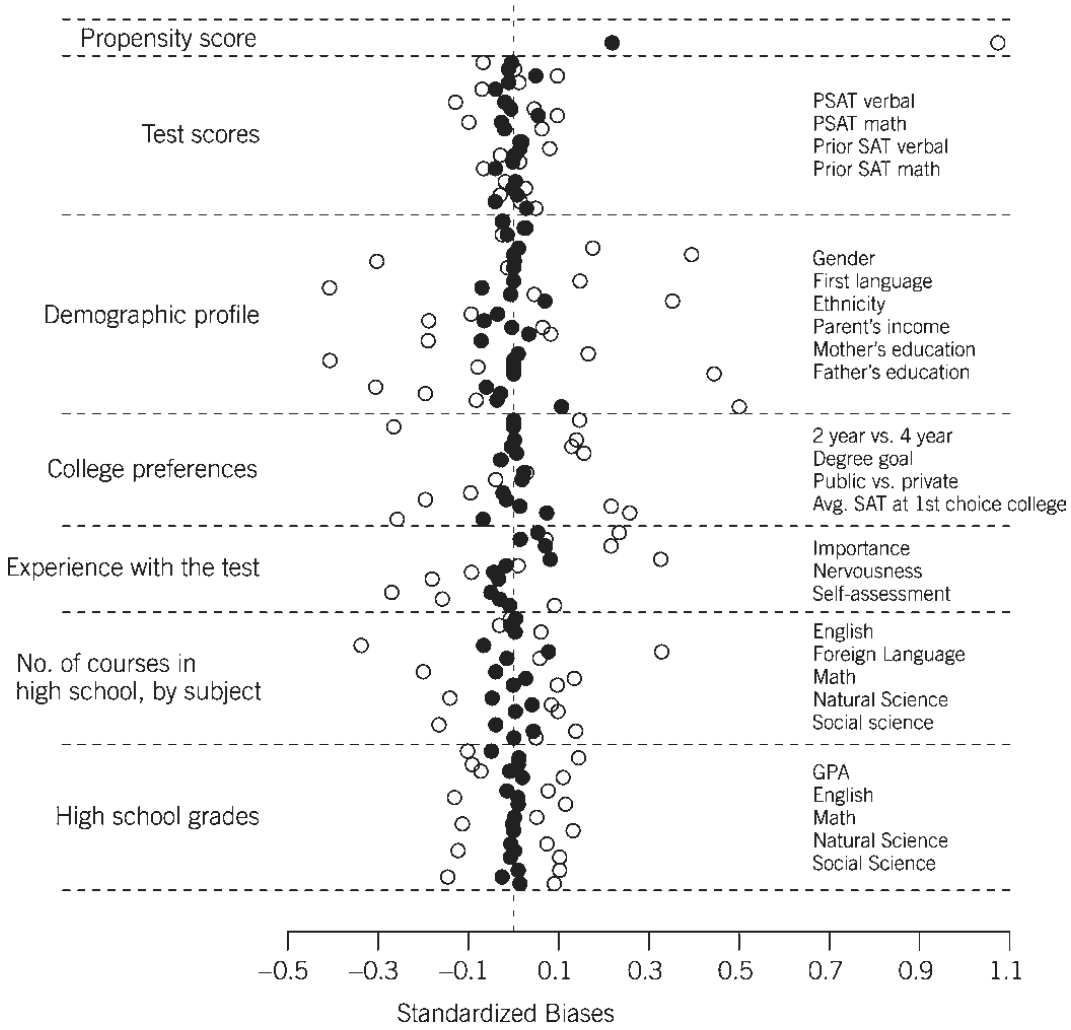


Figure 3. Standardized Biases Without Stratification or Matching, Open Circles, and Under the Optimal [.5, 2] Full Match, Shaded Circles.

Hansen (2004), *Journal of the American Statistical Association*.

A Plethora of Matching Methods

- One-to-one or Many-to-one matching
- Matching with or without replacement
- Calipar matching
- Doubly robust estimation
- Genetic matching
- Optimal matching
- Coarsened exact matching
- Covariate balancing propensity scores

...and many more in the pipeline!

A Plethora of Matching Methods

- One-to-one or Many-to-one matching
- Matching with or without replacement
- Calipar matching
- Doubly robust estimation
- Genetic matching
- Optimal matching
- Coarsened exact matching
- Covariate balancing propensity scores

  ...and many more in the pipeline!

Q: Oh my. Which matching method should I use?

A Plethora of Matching Methods

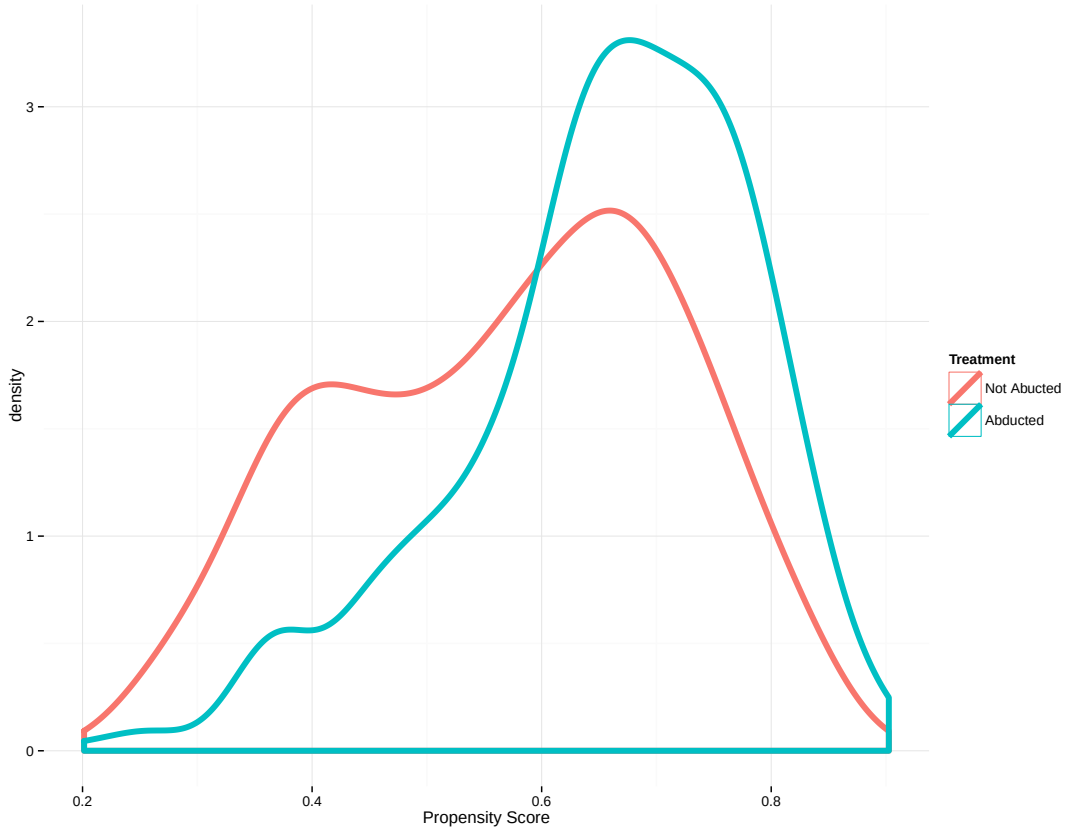
- One-to-one or Many-to-one matching
- Matching with or without replacement
- Calipar matching
- Doubly robust estimation
- Genetic matching
- Optimal matching
- Coarsened exact matching
- Covariate balancing propensity scores

  ...and many more in the pipeline!

Q: Oh my. Which matching method should I use?
A: Whichever gives you the best balance!

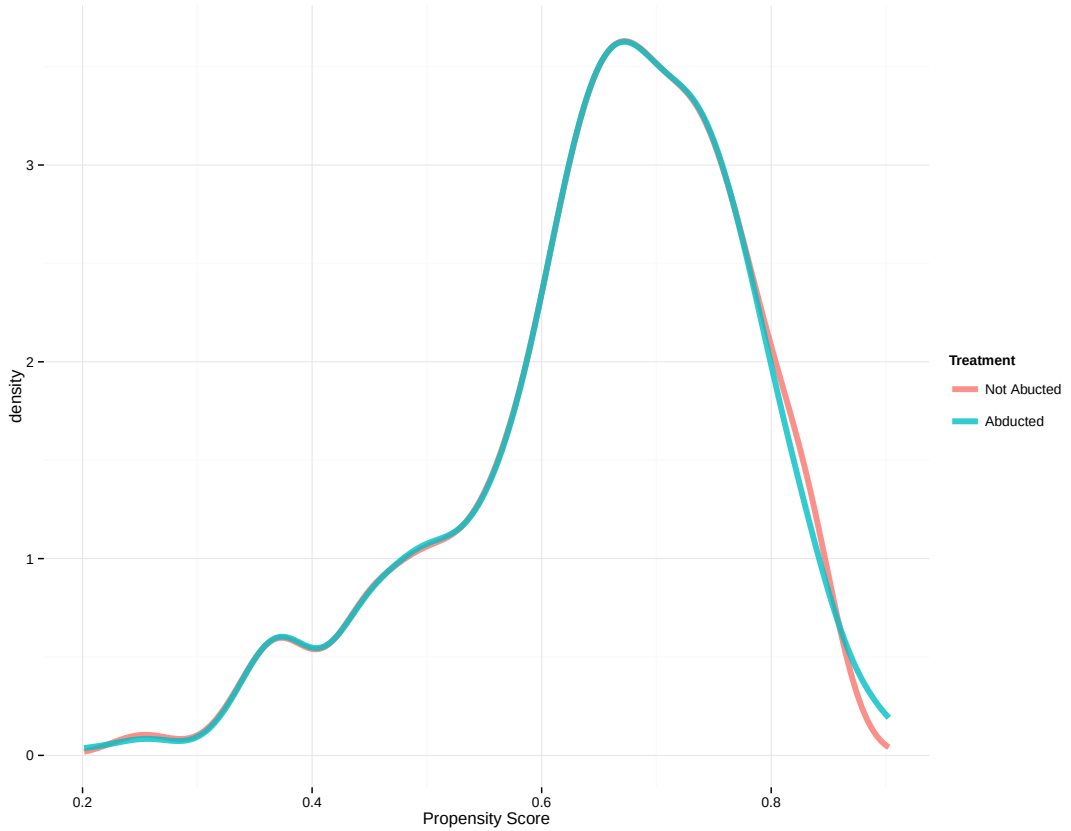
Blattman (2010): Before Matching

```
pscore.fmla <- as.formula(paste("abd~",paste(names(covar),
                                             collapse="+")))
abd <- data$abd
pscore_model <- glm(pscore.fmla, data = data,
                    family = binomial(link = logit))
pscore <- predict(pscore_model, type = "response")
```



Blattman (2010): Propensity Score Matching

```
library(Matching)
match.pscore <- Match(Tr=abd, X=pscore, M=1, estimand="ATT")
```



Blattman (2010): Check Balance

```
MatchBalance(abd ~ age, data=data, match.out = match.pscore)
```

```
***** (V1) age *****
                          Before Matching     After Matching
mean treatment........     21.366              21.366
mean control..........     20.151              20.515
std mean diff.........     24.242              16.976

var ratio (Tr/Co).....     1.0428              0.98412
T-test p-value........  0.0012663           0.0034409
KS Bootstrap p-value..      0.016               0.034
KS Naive p-value......   0.024912            0.070191
KS Statistic..........    0.11227            0.077899
```

Blattman (2010): Mahalanobis Distance Matchng

```
match.mah <- Match(Tr=abd, X=covar, M=1, estimand="ATT",
                   Weight = 2)
MatchBalance(abd ~ age, data=data, match.out = match.mah)

***** (V1) age *****
                          Before Matching      After Matching
mean treatment........     21.366              21.366
mean control..........     20.151              21.154
std mean diff.........     24.242              4.2314

var ratio (Tr/Co).....     1.0428              1.0336
T-test p-value........  0.0012663          3.0386e-05
KS Bootstrap p-value..       0.008               0.798
KS Naive p-value......    0.024912             0.94687
KS Statistic..........     0.11227             0.034261
```

Blattman (2010): Genetic Matching

```
genout <- GenMatch(Tr=abd,X=covar,BalanceMatrix=covar,
                   estimand="ATT",pop.size=1000)
match.gen <- Match(Tr=abd,X=covar,M=1,estimand="ATT",
                    Weight.matrix=genout)
MatchBalance(abd~age,match.out=match.gen,data=covar)

***** (V1) age *****
                          Before Matching      After Matching
mean treatment........     21.366               21.366
mean control..........     20.151               21.225
std mean diff.........     24.242               2.8065

var ratio (Tr/Co).....     1.0428               1.1337
T-test p-value........  0.0012663            0.21628
KS Bootstrap p-value..      0.008                0.454
KS Naive p-value......   0.024912             0.68567
KS Statistic..........    0.11227             0.046512
```

Weighting on the Propensity Score

- An alternative way to achieve balance is weighting
- Can be seen as a continuous version of matching

Weighting on the Propensity Score

- An alternative way to achieve balance is weighting
- Can be seen as a continuous version of matching
- **Result:** Under the conditional ignorability and common support assumptions, we can identify the ATE and ATT as:

$$\tau_{ATE} = \mathbb{E}\left[Y_i \cdot \frac{D_i - \pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))}\right]$$

$$\tau_{ATT} = \frac{1}{\Pr(D_i = 1)} \cdot \mathbb{E}\left[Y_i \cdot \frac{D_i - \pi(X_i)}{1 - \pi(X_i)}\right]$$

Weighting on the Propensity Score

- An alternative way to achieve balance is weighting
- Can be seen as a continuous version of matching
- **Result:** Under the conditional ignorability and common support assumptions, we can identify the ATE and ATT as:

$$\begin{aligned}
\tau_{ATE} &= \mathbb{E}\left[Y_i \cdot \frac{D_i - \pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \right] \\
\tau_{ATT} &= \frac{1}{\Pr(D_i = 1)} \cdot \mathbb{E}\left[Y_i \cdot \frac{D_i - \pi(X_i)}{1 - \pi(X_i)} \right]
\end{aligned}$$

- These can be estimated using sample analogues:

$$\begin{aligned}
\widehat{\tau}_{ATE} &= \frac{1}{N} \sum_{i=1}^{N} \left\{ Y_i \cdot \frac{D_i - \widehat{\pi}(X_i)}{\widehat{\pi}(X_i) \cdot (1 - \widehat{\pi}(X_i))} \right\} = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{D_i Y_i}{\widehat{\pi}(X_i)} - \frac{(1 - D_i) Y_i}{1 - \widehat{\pi}(X_i)} \right\} \\
\widehat{\tau}_{ATT} &= \frac{1}{N_1} \sum_{i=1}^{N} \left\{ Y_i \cdot \frac{D_i - \widehat{\pi}(X_i)}{1 - \widehat{\pi}(X_i)} \right\} = \frac{1}{N_1} \sum_{i=1}^{N} \left\{ D_i Y_i - (1 - D_i) Y_i \frac{\widehat{\pi}(X_i)}{1 - \widehat{\pi}(X_i)} \right\}
\end{aligned}$$

These inverse PS weighting (IPW) estimators are consistent, but not unbiased.

Proof of Identification with PS Weighting

We begin with the estimator conditional on a specific covariate value x:

$$
\begin{aligned}
\tilde{\tau}(x) &\equiv \mathbb{E}\left[Y_i \cdot \frac{D_i - \pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \,\middle|\, X_i = x \right] \\
&= \mathbb{E}\left[Y_i \cdot \frac{1 - \pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \,\middle|\, X_i = x, D_i = 1 \right] \Pr(D_i = 1 \mid X_i = x) \\
&\quad -\mathbb{E}\left[Y_i \cdot \frac{\pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \,\middle|\, X_i = x, D_i = 0 \right] \Pr(D_i = 0 \mid X_i = x) \\
&\quad (\because
\end{aligned}
$$

Proof of Identification with PS Weighting

We begin with the estimator conditional on a specific covariate value x:

$$
\begin{aligned}
\tilde{\tau}(x) &\equiv \mathbb{E}\left[Y_i \cdot \frac{D_i - \pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \middle| X_i = x \right] \\
&= \mathbb{E}\left[Y_i \cdot \frac{1 - \pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \middle| X_i = x, D_i = 1 \right] \Pr(D_i = 1 \mid X_i = x) \\
&\quad -\mathbb{E}\left[Y_i \cdot \frac{\pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \middle| X_i = x, D_i = 0 \right] \Pr(D_i = 0 \mid X_i = x) \\
&\qquad\qquad (\because \text{ Law of Total Expectation}) \\
&= \mathbb{E}\left[\frac{Y_i}{\pi(X_i)} \middle| X_i = x, D_i = 1 \right] \pi(x) - \mathbb{E}\left[\frac{Y_i}{1 - \pi(X_i)} \middle| X_i = x, D_i = 0 \right] (1 - \pi(x)) \\
&=
\end{aligned}
$$

Proof of Identification with PS Weighting

We begin with the estimator conditional on a specific covariate value x:

$$
\begin{aligned}
\tilde{\tau}(x) &\equiv \mathbb{E}\left[Y_i \cdot \frac{D_i - \pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \,\middle|\, X_i = x \right] \\
&= \mathbb{E}\left[Y_i \cdot \frac{1 - \pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \,\middle|\, X_i = x, D_i = 1 \right] \Pr(D_i = 1 \mid X_i = x) \\
&\quad - \mathbb{E}\left[Y_i \cdot \frac{\pi(X_i)}{\pi(X_i) \cdot (1 - \pi(X_i))} \,\middle|\, X_i = x, D_i = 0 \right] \Pr(D_i = 0 \mid X_i = x) \\
&\qquad\qquad (\because \text{ Law of Total Expectation}) \\
&= \mathbb{E}\left[\frac{Y_i}{\pi(X_i)} \,\middle|\, X_i = x, D_i = 1 \right] \pi(x) - \mathbb{E}\left[\frac{Y_i}{1 - \pi(X_i)} \,\middle|\, X_i = x, D_i = 0 \right] (1 - \pi(x)) \\
&= \mathbb{E}[Y_i | X_i = x, D_i = 1] - \mathbb{E}[Y_i | X_i = x, D_i = 0] \\
&= \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x] \;=\; \tau(x) \quad (\because \text{ conditional ignorability})
\end{aligned}
$$

Averaging $\tau(x)$ over the distribution of x, $f(x)$, yields the expression of τ_{ATE} on the previous slide.
The same argument also shows the result for τ_{ATT} if we now average over $f(x \mid D_i = 1)$. □

Blattman (2010): Balance after PS Weighting Score

```
# Balance in age before weighting
mean(covar$age[abd == 1]) - mean(covar$age[abd == 0])
[1] 1.215263

# Balance in age after weighting
sum(covar$age * (abd - pscore) / (1 - pscore)) / length(abd)
[1] 0.007663878
```

Performance of the IPW estimators

- Recall that IPW estimators are consistent but biased in small samples. How bad are their small sample biases?
- It turns out the bias is substantial when some weights are extremely large or small.
- Weights tend to be extreme when there is a lack of overlap.

Performance of the IPW estimators

- Recall that IPW estimators are consistent but biased in small samples. How bad are their small sample biases?
- It turns out the bias is substantial when some weights are extremely large or small.
- Weights tend to be extreme when there is a lack of overlap.
- A simple workaround is trim units with extreme weights.
- Problem: Trimming changes the estimand to a quantity that is still causal yet difficult to interpret.

Performance of the IPW estimators

- Recall that IPW estimators are consistent but biased in small samples. How bad are their small sample biases?
- It turns out the bias is substantial when some weights are extremely large or small.
- Weights tend to be extreme when there is a lack of overlap.
- A simple workaround is trim units with extreme weights.
- Problem: Trimming changes the estimand to a quantity that is still causal yet difficult to interpret.
- Alternative weighting methods for better balance includes:
  - Entropy balancing (Hainmueller 2012, `ebal`): Choose weights that directly optimize covariate balance.
  - Covariate balancing propensity scores (Imai and Ratkovic 2014, `CBPS`)

  etc. (again, many more are in the pipeline)

Summary: Estimation under Conditional Ignorability

- Matching and weighting are main methods to estimate average causal effects when one can assume conditional ignorability
- Many alternative methods are available and easily implemented in R, Stata, SAS, etc., and no single method is dominant
- Key is to balance treatment and control groups and avoid extrapolation
- Use whichever method to achieve good balance, then estimate causal effects

Summary: Estimation under Conditional Ignorability

- Matching and weighting are main methods to estimate average causal effects when one can assume conditional ignorability
- Many alternative methods are available and easily implemented in R, Stata, SAS, etc., and no single method is dominant
- Key is to balance treatment and control groups and avoid extrapolation
- Use whichever method to achieve good balance, then estimate causal effects

- Note: You could also use regression to control for the observed pretreatment covariates – a model-based approach to conditioning on the covariates
- Estimates will be close to unbiased for ATE if (1) linear approximation is good or (2) causal effects are highly homogeneous across units

1. Introduction
2. Identification under Conditional Ignorability
3. Back-Door Criterion
4. Estimation under Conditional Ignorability
   - Subclassification
   - Matching
   - Weighting
5. Inference without Conditional Ignorability
   - Nonparametric Bounds
   - Sensitivity Analysis
     - Imbens' Approach
     - Rosenbaum's Approach

Motivation

- Causal quantities of interest are often not identifiable from observed data without invoking strong assuptions that cannot be well justified
- How can we make causal inference when we are not willing to fully endorse assumptions such as conditional ignorability?
- Manski's approach: Partial identification
  - Assume only what is credible
  - Derive bounds (preferrably nonparametric sharp bounds) on the QoI, i.e., the set of possible values that it can *logically* take
  - This establishes a "domain of consensus" among researchers who might disagree on what assumptions they are willing to make

Motivation

- Causal quantities of interest are often not identifiable from observed data without invoking strong assuptions that cannot be well justified
- How can we make causal inference when we are not willing to fully endorse assumptions such as conditional ignorability?
- Manski's approach: Partial identification
  - Assume only what is credible
  - Derive bounds (preferrably nonparametric sharp bounds) on the QoI, i.e., the set of possible values that it can *logically* take
  - This establishes a "domain of consensus" among researchers who might disagree on what assumptions they are willing to make
  - Identification analysis precedes statistical inference

Motivation

- Causal quantities of interest are often not identifiable from observed data without invoking strong assuptions that cannot be well justified
- How can we make causal inference when we are not willing to fully endorse assumptions such as conditional ignorability?
- Manski's approach: Partial identification
  - Assume only what is credible
  - Derive bounds (preferrably nonparametric sharp bounds) on the QoI, i.e., the set of possible values that it can *logically* take
  - This establishes a "domain of consensus" among researchers who might disagree on what assumptions they are willing to make
  - Identification analysis precedes statistical inference
- Key principle: Law of Decreasing Credibility
  - The stronger your assumptions, the less credible your inference.

Motivation

- Causal quantities of interest are often not identifiable from observed data without invoking strong assuptions that cannot be well justified
- How can we make causal inference when we are not willing to fully endorse assumptions such as conditional ignorability?
- Manski's approach: Partial identification
  - Assume only what is credible
  - Derive bounds (preferrably nonparametric sharp bounds) on the QoI, i.e., the set of possible values that it can *logically* take
  - This establishes a "domain of consensus" among researchers who might disagree on what assumptions they are willing to make
  - Identification analysis precedes statistical inference
- Key principle: Law of Decreasing Credibility
  - The stronger your assumptions, the less credible your inference.
  - Add an assumption, rederive the bounds and see exactly the contribution of the assumption

Self-Selection into Treatment

Notation:

- Treatment (binary): $D_i \in \{0, 1\}$
- Potential outcomes: Y_{di}
- ATE is the quantity of interest:

$$\tau \equiv \mathbb{E}[Y_{1i} - Y_{0i}]$$

Self-Selection into Treatment

Notation:

- Treatment (binary): $D_i \in \{0, 1\}$
- Potential outcomes: Y_{di}
- ATE is the quantity of interest:

$$\tau \equiv \mathbb{E}[Y_{1i} - Y_{0i}]$$

How much can we learn about τ from data, without making any assumption?

Self-Selection into Treatment

Notation:

- Treatment (binary): $D_i \in \{0, 1\}$
- Potential outcomes: Y_{di}
- ATE is the quantity of interest:

$$\tau \equiv \mathbb{E}[Y_{1i} - Y_{0i}]$$

How much can we learn about τ from data, without making any assumption?

$$\begin{aligned}
\tau &= \mathbb{E}[Y_{1i} - Y_{0i}] \\
&= \mathbb{E}[Y_{1i} \mid D_i = 1] \Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0] \Pr(D_i = 0) \\
&\quad - \mathbb{E}[Y_{0i} \mid D_i = 1] \Pr(D_i = 1) - \mathbb{E}[Y_{0i} \mid D_i = 0] \Pr(D_i = 0)
\end{aligned}$$

Self-Selection into Treatment

Notation:

- Treatment (binary): $D_i \in \{0, 1\}$
- Potential outcomes: Y_{di}
- ATE is the quantity of interest:

$$\tau \equiv \mathbb{E}[Y_{1i} - Y_{0i}]$$

How much can we learn about τ from data, without making any assumption?

$$\begin{aligned}
\tau &= \mathbb{E}[Y_{1i} - Y_{0i}] \\
&= \mathbb{E}[Y_{1i} \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]\Pr(D_i = 0) \\
&\quad - \mathbb{E}[Y_{0i} \mid D_i = 1]\Pr(D_i = 1) - \mathbb{E}[Y_{0i} \mid D_i = 0]\Pr(D_i = 0)
\end{aligned}$$

Self-Selection into Treatment

Notation:

- Treatment (binary): $D_i \in \{0, 1\}$
- Potential outcomes: Y_{di}
- ATE is the quantity of interest:

$$\tau \equiv \mathbb{E}[Y_{1i} - Y_{0i}]$$

How much can we learn about τ from data, without making any assumption?

$$\begin{aligned}
\tau &= \mathbb{E}[Y_{1i} - Y_{0i}] \\
&= \mathbb{E}[Y_{1i} \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]\Pr(D_i = 0) \\
&\quad - \mathbb{E}[Y_{0i} \mid D_i = 1]\Pr(D_i = 1) - \mathbb{E}[Y_{0i} \mid D_i = 0]\Pr(D_i = 0) \\
&= \mathbb{E}[Y_i \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]\Pr(D_i = 0) \\
&\quad - \mathbb{E}[Y_{0i} \mid D_i = 1]\Pr(D_i = 1) - \mathbb{E}[Y_i \mid D_i = 0]\Pr(D_i = 0)
\end{aligned}$$

Quantities in red are unobserved, so data tell us nothing about them unless we make some assumptions.

Constructing Nonparametric Bounds

	D_i	Y_{0i}	Y_{1i}
$\Pr(D_i = 0)$	0	$\mathbb{E}[Y_{0i}\|D_i = 0]$	?
$\Pr(D_i = 1)$	1	?	$\mathbb{E}[Y_{1i}\|D_i = 1]$

Constructing Nonparametric Bounds

	D_i	Y_{0i}	Y_{1i}
$\Pr(D_i = 0)$	0	$\mathbb{E}[Y_{0i}\|D_i = 0]$	$\mathbb{E}[Y_{1i}\|D_i = 1]$
$\Pr(D_i = 1)$	1	$\mathbb{E}[Y_{0i}\|D_i = 0]$	$\mathbb{E}[Y_{1i}\|D_i = 1]$

- Randomized experiments let us impute missing PO directly
- Treatment and control groups are identical in expectation

  $\therefore$ PO of treatment and control groups identical in expectation

Constructing Nonparametric Bounds

	D_i	Y_{0i}	Y_{1i}
$\Pr(D_i = 0)$	0	$\mathbb{E}[Y_{0i} \mid D_i = 0]$	$\underline{Y}$
$\Pr(D_i = 1)$	1	$\overline{Y}$	$\mathbb{E}[Y_{1i} \mid D_i = 1]$

- Nuclear option: assume the worst possible outcome
- Treated units would have best possible outcome ($\overline{Y}$) if untreated
- Control units would have had worst possible outcome ($\underline{Y}$) if treated

Constructing Nonparametric Bounds

	D_i	Y_{0i}	Y_{1i}
$\Pr(D_i = 0)$	0	$\mathbb{E}[Y_{0i} \mid D_i = 0]$	$\underline{Y}$
$\Pr(D_i = 1)$	1	$\overline{Y}$	$\mathbb{E}[Y_{1i} \mid D_i = 1]$

- Nuclear option: assume the worst possible outcome
- Treated units would have best possible outcome ($\overline{Y}$) if untreated
- Control units would have had worst possible outcome ($\underline{Y}$) if treated

This gives the sharp lower bound on τ:

$$\underline{\tau} = \left(\mathbb{E}[Y_i \mid D_i = 1] - \overline{Y}\right) \Pr(D_i = 1) + \left(\underline{Y} - \mathbb{E}[Y_i \mid D_i = 0]\right) \Pr(D_i = 0)$$

Constructing Nonparametric Bounds

	D_i	Y_{0i}	Y_{1i}
$\Pr(D_i = 0)$	0	$\mathbb{E}[Y_{0i} \mid D_i = 0]$	$\overline{Y}$
$\Pr(D_i = 1)$	1	$\underline{Y}$	$\mathbb{E}[Y_{1i} \mid D_i = 1]$

- Conversely, consider the best possible scenario
- Control units would have had best possible outcome ($\overline{Y}$) if treated
- Treated units would have worst possible outcome ($\underline{Y}$) if untreated

This yields the sharp upper bound on τ:

$$\overline{\tau} = (\mathbb{E}[Y_i \mid D_i = 1] - \underline{Y}) \Pr(D_i = 1) + \left(\overline{Y} - \mathbb{E}[Y_i \mid D_i = 0]\right) \Pr(D_i = 0)$$

Adding Assumptions

$$\begin{aligned}\tau &= \mathbb{E}[Y_i \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]\Pr(D_i = 0) \\ &\quad -\mathbb{E}[Y_{0i} \mid D_i = 1]\Pr(D_i = 1) - \mathbb{E}[Y_i \mid D_i = 0]\Pr(D_i = 0)\end{aligned}$$

- The no-assumption sharp bounds are often too wide to be useful (e.g. If $Y_i \in \{0, 1\}$, the bounds always include zero)
- The next step is to add an assumption and see how bounds change

Example: Monotone treatment selection (MTS) assumption:

$$\begin{aligned}\mathbb{E}[Y_{0i} \mid D_i = 0] &\leq \mathbb{E}[Y_{0i} \mid D_i = 1] \\ \mathbb{E}[Y_{1i} \mid D_i = 0] &\leq \mathbb{E}[Y_{1i} \mid D_i = 1]\end{aligned}$$

- In words, MTS assumes that

Adding Assumptions

$$\begin{aligned}\tau &= \mathbb{E}[Y_i \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]\Pr(D_i = 0) \\ &\quad -\mathbb{E}[Y_{0i} \mid D_i = 1]\Pr(D_i = 1) - \mathbb{E}[Y_i \mid D_i = 0]\Pr(D_i = 0)\end{aligned}$$

- The no-assumption sharp bounds are often too wide to be useful (e.g. If $Y_i \in \{0, 1\}$, the bounds always include zero)
- The next step is to add an assumption and see how bounds change

Example: Monotone treatment selection (MTS) assumption:

$$\begin{aligned}\mathbb{E}[Y_{0i} \mid D_i = 0] &\leq \mathbb{E}[Y_{0i} \mid D_i = 1] \\ \mathbb{E}[Y_{1i} \mid D_i = 0] &\leq \mathbb{E}[Y_{1i} \mid D_i = 1]\end{aligned}$$

- In words, MTS assumes that units who select the treatment have higher expectation of outcome under either condition on average (e.g. sicker)

Adding Assumptions

$$\begin{aligned}\tau &= \mathbb{E}[Y_i \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]\Pr(D_i = 0) \\ &\quad -\mathbb{E}[Y_{0i} \mid D_i = 1]\Pr(D_i = 1) - \mathbb{E}[Y_i \mid D_i = 0]\Pr(D_i = 0)\end{aligned}$$

- The no-assumption sharp bounds are often too wide to be useful (e.g. If $Y_i \in \{0, 1\}$, the bounds always include zero)
- The next step is to add an assumption and see how bounds change

Example: Monotone treatment selection (MTS) assumption:

$$\begin{aligned}\mathbb{E}[Y_{0i} \mid D_i = 0] &\leq \mathbb{E}[Y_{0i} \mid D_i = 1] \\ \mathbb{E}[Y_{1i} \mid D_i = 0] &\leq \mathbb{E}[Y_{1i} \mid D_i = 1]\end{aligned}$$

- In words, MTS assumes that units who select the treatment have higher expectation of outcome under either condition on average (e.g. sicker)
- This implies a tighter upper bound on τ:

Adding Assumptions

$$\begin{aligned}\tau \quad = \quad & \mathbb{E}[Y_i \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]\Pr(D_i = 0) \\ & -\mathbb{E}[Y_{0i} \mid D_i = 1]\Pr(D_i = 1) - \mathbb{E}[Y_i \mid D_i = 0]\Pr(D_i = 0)\end{aligned}$$

- The no-assumption sharp bounds are often too wide to be useful (e.g. If $Y_i \in \{0, 1\}$, the bounds always include zero)
- The next step is to add an assumption and see how bounds change

Example: Monotone treatment selection (MTS) assumption:

$$\begin{aligned}\mathbb{E}[Y_{0i} \mid D_i = 0] \quad & \leq \quad \mathbb{E}[Y_{0i} \mid D_i = 1] \\ \mathbb{E}[Y_{1i} \mid D_i = 0] \quad & \leq \quad \mathbb{E}[Y_{1i} \mid D_i = 1]\end{aligned}$$

- In words, MTS assumes that units who select the treatment have higher expectation of outcome under either condition on average (e.g. sicker)
- This implies a tighter upper bound on τ:

$$\begin{aligned}\tau \quad \leq \quad & \mathbb{E}[Y_i \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 1]\Pr(D_i = 0) \\ & -\mathbb{E}[Y_{0i} \mid D_i = 0]\Pr(D_i = 1) - \mathbb{E}[Y_i \mid D_i = 0]\Pr(D_i = 0)\end{aligned}$$

Adding Assumptions

$$\begin{aligned}\tau &= \mathbb{E}[Y_i \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 0]\Pr(D_i = 0) \\ &\quad -\mathbb{E}[Y_{0i} \mid D_i = 1]\Pr(D_i = 1) - \mathbb{E}[Y_i \mid D_i = 0]\Pr(D_i = 0)\end{aligned}$$

- The no-assumption sharp bounds are often too wide to be useful (e.g. If $Y_i \in \{0, 1\}$, the bounds always include zero)
- The next step is to add an assumption and see how bounds change

Example: Monotone treatment selection (MTS) assumption:

$$\begin{aligned}\mathbb{E}[Y_{0i} \mid D_i = 0] &\leq \mathbb{E}[Y_{0i} \mid D_i = 1] \\ \mathbb{E}[Y_{1i} \mid D_i = 0] &\leq \mathbb{E}[Y_{1i} \mid D_i = 1]\end{aligned}$$

- In words, MTS assumes that units who select the treatment have higher expectation of outcome under either condition on average (e.g. sicker)
- This implies a tighter upper bound on τ:

$$\begin{aligned}\tau &\leq \mathbb{E}[Y_i \mid D_i = 1]\Pr(D_i = 1) + \mathbb{E}[Y_{1i} \mid D_i = 1]\Pr(D_i = 0) \\ &\quad -\mathbb{E}[Y_{0i} \mid D_i = 0]\Pr(D_i = 1) - \mathbb{E}[Y_i \mid D_i = 0]\Pr(D_i = 0) \\ &= \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \quad (= \text{naive diff. in means})\end{aligned}$$

Example: Sentencing and Recidivism

- Manski (2007, Ch.7.2) uses an example on how the type of sentence for juvenile offenders affect recidivism:
  - $D_i = 1$ if sentence involves confinement in residential facilities; 0 if not
  - $Y_{di} = 1$ if commits a crime again given sentence type d; 0 if not

Example: Sentencing and Recidivism

- Manski (2007, Ch.7.2) uses an example on how the type of sentence for juvenile offenders affect recidivism:
  - $D_i = 1$ if sentence involves confinement in residential facilities; 0 if not
  - $Y_{di} = 1$ if commits a crime again given sentence type d; 0 if not
- Observed strata:

	Y	
D	0	1
0	.36	.53
1	.03	.08

Example: Sentencing and Recidivism

- Manski (2007, Ch.7.2) uses an example on how the type of sentence for juvenile offenders affect recidivism:
  - $D_i = 1$ if sentence involves confinement in residential facilities; 0 if not
  - $Y_{di} = 1$ if commits a crime again given sentence type d; 0 if not
- Observed strata:

	Y	
D	0	1
0	.36	.53
1	.03	.08

- Nonparametric bounds:
  - Random assignment: $.08/(.03 + .08) - .53/(.36 + .53) = .13$

Example: Sentencing and Recidivism

- Manski (2007, Ch.7.2) uses an example on how the type of sentence for juvenile offenders affect recidivism:
  - $D_i = 1$ if sentence involves confinement in residential facilities; 0 if not
  - $Y_{di} = 1$ if commits a crime again given sentence type d; 0 if not
- Observed strata:

	Y	
D	0	1
0	.36	.53
1	.03	.08

- Nonparametric bounds:
  - Random assignment: $.08/(.03 + .08) - .53/(.36 + .53) = .13$
  - No assumption: $[-.53 - .03,\ .36 + .08] = [-.56,\ .44]$
  - MTS: $[-.56,\ .13]$

Sensitivity Analysis for Conditional Ignorability

- An alternative approach: Sensitivity analysis

Sensitivity Analysis for Conditional Ignorability

- An alternative approach: Sensitivity analysis
- Sensitivity analysis takes the following general form:
  1. Quantify the degree of violation of the key assumption by a sensitivity parameter (σ)
  2. Set σ to various values and derive what the true value of the quantity of interest would be
  3. See at what point the effect would go away completely (or become statistically insignificant)

Sensitivity Analysis for Conditional Ignorability

- An alternative approach: Sensitivity analysis
- Sensitivity analysis takes the following general form:
  1. Quantify the degree of violation of the key assumption by a sensitivity parameter (σ)
  2. Set σ to various values and derive what the true value of the quantity of interest would be
  3. See at what point the effect would go away completely (or become statistically insignificant)
- In the current context, we ask:

  How substantial would the unobserved confouding have to be in order for the estimated treatment effect to completely go away?

Sensitivity Analysis for ATE

- Assume that conditional ignorability would hold if you could control for unobserved U_i, i.e.,

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i \mid U_i$$

but

$$(Y_{1i}, Y_{0i}) \not\perp\!\!\!\perp D_i$$

Sensitivity Analysis for ATE

- Assume that conditional ignorability would hold if you could control for unobserved U_i, i.e.,

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i \mid U_i$$

but

$$(Y_{1i}, Y_{0i}) \not\perp\!\!\!\perp D_i$$

- If U_i were observed (and discrete), the true ATE could be estimated by subclassification:

$$\tau = \sum_u \{\mathbb{E}[Y_i \mid D_i = 1, U_i = u] - \mathbb{E}[Y_i \mid D_i = 0, U_i = u]\} \Pr(U_i = u)$$

- But you don't observe U_i, so you can only estimate

$$\hat{\tau} = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

Sensitivity Analysis for ATE

For simplicy assume $U_i \in \{0, 1\}$. Then the bias is

$$\begin{aligned}\mathbb{E}[\hat{\tau}] - \tau &= \mathbb{E}\left\{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]\right\} \\ &\quad - \sum_{u=0,1} \left\{\mathbb{E}[Y_i \mid D_i = 1, U_i = u] - \mathbb{E}[Y_i \mid D_i = 0, U_i = u]\right\} \Pr(U_i = u)\end{aligned}$$

Sensitivity Analysis for ATE

For simplicy assume $U_i \in \{0, 1\}$. Then the bias is

$$\begin{aligned}
\mathbb{E}[\hat{\tau}] - \tau &= \mathbb{E}\left\{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]\right\} \\
&\quad - \sum_{u=0,1} \left\{\mathbb{E}[Y_i \mid D_i = 1, U_i = u] - \mathbb{E}[Y_i \mid D_i = 0, U_i = u]\right\} \Pr(U_i = u) \\
&= \text{(some algebra)}
\end{aligned}$$

Sensitivity Analysis for ATE

For simplicy assume $U_i \in \{0, 1\}$. Then the bias is

$$
\begin{aligned}
\mathbb{E}[\hat{\tau}] - \tau &= \mathbb{E}\left\{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]\right\} \\
&\quad - \sum_{u=0,1} \left\{\mathbb{E}[Y_i \mid D_i = 1, U_i = u] - \mathbb{E}[Y_i \mid D_i = 0, U_i = u]\right\} \Pr(U_i = u) \\
&= \text{(some algebra)} \\
&= \left\{\mathbb{E}[Y_i \mid D_i = 1, U_i = 1] - \mathbb{E}[Y_i \mid D_i = 1, U_i = 0]\right\} \\
&\qquad \cdot \left\{\Pr(U_i = 1 \mid D_i = 1) - \Pr(U_i = 1)\right\} \\
&\quad - \left\{\mathbb{E}[Y_i \mid D_i = 0, U_i = 1] - \mathbb{E}[Y_i \mid D_i = 0, U_i = 0]\right\} \\
&\qquad \cdot \left\{\Pr(U_i = 1 \mid D_i = 0) - \Pr(U_i = 1)\right\}
\end{aligned}
$$

Sensitivity Analysis for ATE

For simplicy assume $U_i \in \{0, 1\}$. Then the bias is

$$
\begin{aligned}
\mathbb{E}[\hat{\tau}] - \tau &= \mathbb{E}\left\{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]\right\} \\
&\quad - \sum_{u=0,1} \left\{\mathbb{E}[Y_i \mid D_i = 1, U_i = u] - \mathbb{E}[Y_i \mid D_i = 0, U_i = u]\right\} \Pr(U_i = u) \\
&= \text{(some algebra)} \\
&= \left\{\mathbb{E}[Y_i \mid D_i = 1, U_i = 1] - \mathbb{E}[Y_i \mid D_i = 1, U_i = 0]\right\} \\
&\qquad \cdot \left\{\Pr(U_i = 1 \mid D_i = 1) - \Pr(U_i = 1)\right\} \\
&\quad - \left\{\mathbb{E}[Y_i \mid D_i = 0, U_i = 1] - \mathbb{E}[Y_i \mid D_i = 0, U_i = 0]\right\} \\
&\qquad \cdot \left\{\Pr(U_i = 1 \mid D_i = 0) - \Pr(U_i = 1)\right\}
\end{aligned}
$$

So the bias can be characterized by:

- The relationship between Y_i and U_i in each treatment group
- The relationship between U_i and D_i

Sensitivity Analysis for ATE

Now to further simplify things, assume that D_i and U_i do not interact (i.e. the average effect of U_i on Y_i is constant between treatment groups)

Sensitivity Analysis for ATE

Now to further simplify things, assume that D_i and U_i do not interact (i.e. the average effect of U_i on Y_i is constant between treatment groups)

Under this asumption, the bias is:

$$\begin{aligned}\mathbb{E}[\hat{\tau}] - \tau &= \{\Pr(U_i = 1|D_i = 1) - \Pr(U_i = 1|D_i = 0)\} \\ &\quad \cdot \{\mathbb{E}[Y_i|U_i = 1] - \mathbb{E}[Y_i|U_i = 0]\} \\ &\equiv \delta\,\gamma,\end{aligned}$$

where $\begin{cases} \delta &= \text{difference in average } U_i \text{ between treatment conditions} \\ \gamma &= \text{effect of } U_i \text{ on } Y_i \end{cases}$

Sensitivity Analysis for ATE

Now to further simplify things, assume that D_i and U_i do not interact (i.e. the average effect of U_i on Y_i is constant between treatment groups)

Under this asumption, the bias is:

$$\begin{aligned}\mathbb{E}[\hat{\tau}] - \tau &= \{\Pr(U_i = 1|D_i = 1) - \Pr(U_i = 1|D_i = 0)\} \\ &\quad \cdot \{\mathbb{E}[Y_i|U_i = 1] - \mathbb{E}[Y_i|U_i = 0]\} \\ &\equiv \delta\,\gamma,\end{aligned}$$

where $\begin{cases} \delta &= \text{difference in average } U_i \text{ between treatment conditions} \\ \gamma &= \text{effect of } U_i \text{ on } Y_i \end{cases}$

Imbens-style sensitivity analysis proceeds by setting δ and γ (sensitivity parameters) to different values and see what the true τ would be.

Sensitivity Analysis for ATE

Now to further simplify things, assume that D_i and U_i do not interact (i.e. the average effect of U_i on Y_i is constant between treatment groups)

Under this asumption, the bias is:

$$\begin{aligned}\mathbb{E}[\hat{\tau}] - \tau &= \{\Pr(U_i = 1|D_i = 1) - \Pr(U_i = 1|D_i = 0)\} \\ &\quad \cdot \{\mathbb{E}[Y_i|U_i = 1] - \mathbb{E}[Y_i|U_i = 0]\} \\ &\equiv \delta\,\gamma,\end{aligned}$$

where $\begin{cases} \delta &= \text{difference in average } U_i \text{ between treatment conditions} \\ \gamma &= \text{effect of } U_i \text{ on } Y_i \end{cases}$

Imbens-style sensitivity analysis proceeds by setting δ and γ (sensitivity parameters) to different values and see what the true τ would be.

Notes:

- Observed covariates (X_i) can be incorporated with minor extension.
- The framework is fully nonparametric so far, but we need additional parametric assumptions to accommodate X, non-binary U or D, etc.

Sensitivity Analysis for ATE

- A parametric example with continuous treatment X, with the following true model

$$\begin{aligned} X &= U\delta + \eta \\ Y &= X\beta + U\gamma + \varepsilon \end{aligned}$$

  where η and ε are independent error terms with $\mathbb{E}[\eta \mid U] = \mathbb{E}[\varepsilon \mid X, U] = 0$

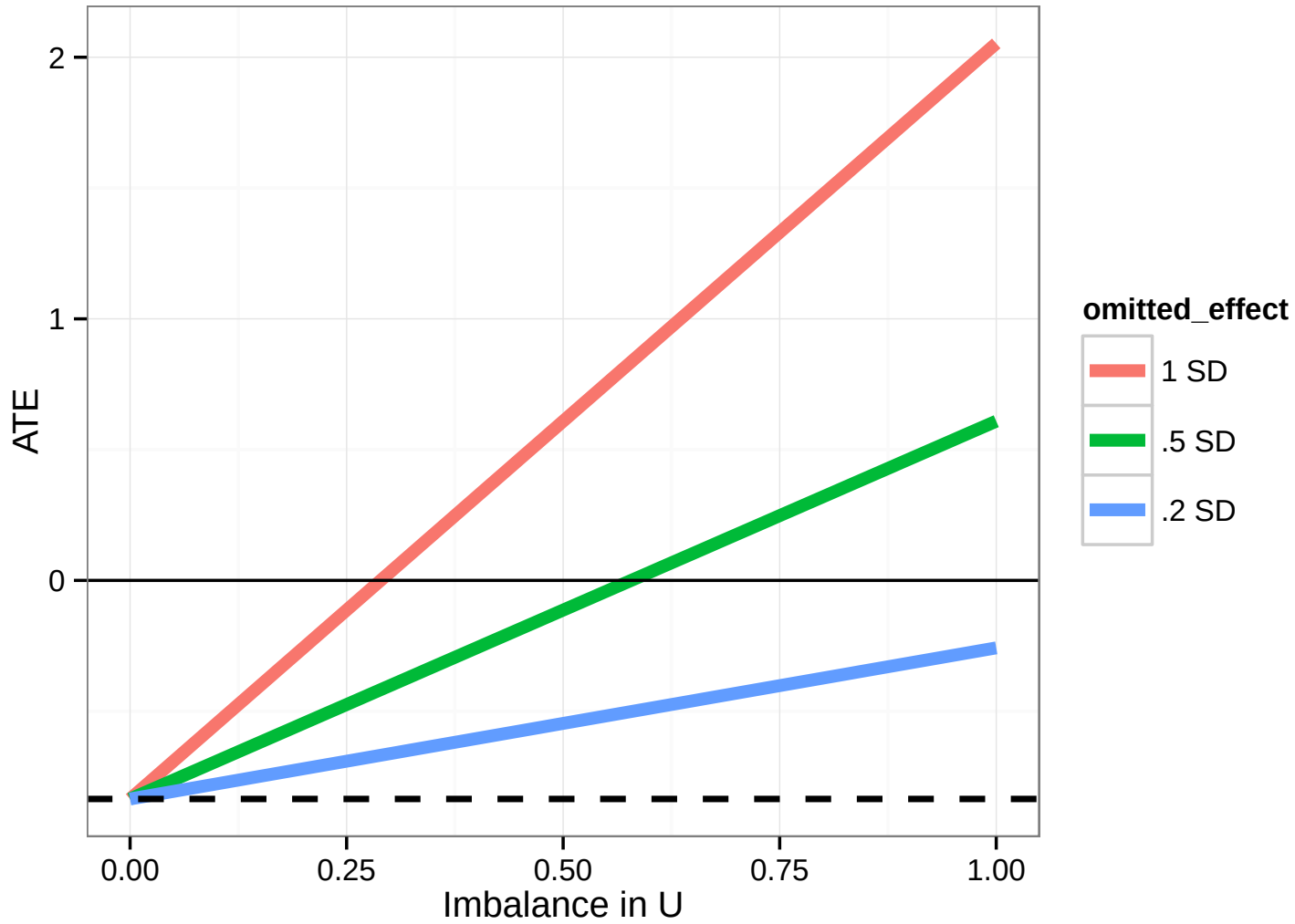
- With U unobserved, we run the linear regression

$$\begin{aligned} \hat{\beta} &= (X^\top X)^{-1} X^\top Y \\ &= (X^\top X)^{-1} X^\top (X\beta + U\gamma + \varepsilon) \\ \mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^\top X)^{-1} X^\top (X\beta + U\gamma + \varepsilon)] \\ &= \mathbb{E}[\cancel{(X^\top X)^{-1} X^\top X}\beta] + \mathbb{E}[(X^\top X)^{-1} X^\top U\gamma] + \cancel{\mathbb{E}[(X^\top X)^{-1} X^\top X\varepsilon)]} \\ &= \beta + \delta\gamma \end{aligned}$$

- Note this is identical to the "omitted variables bias formula" which you might remember from a regression class

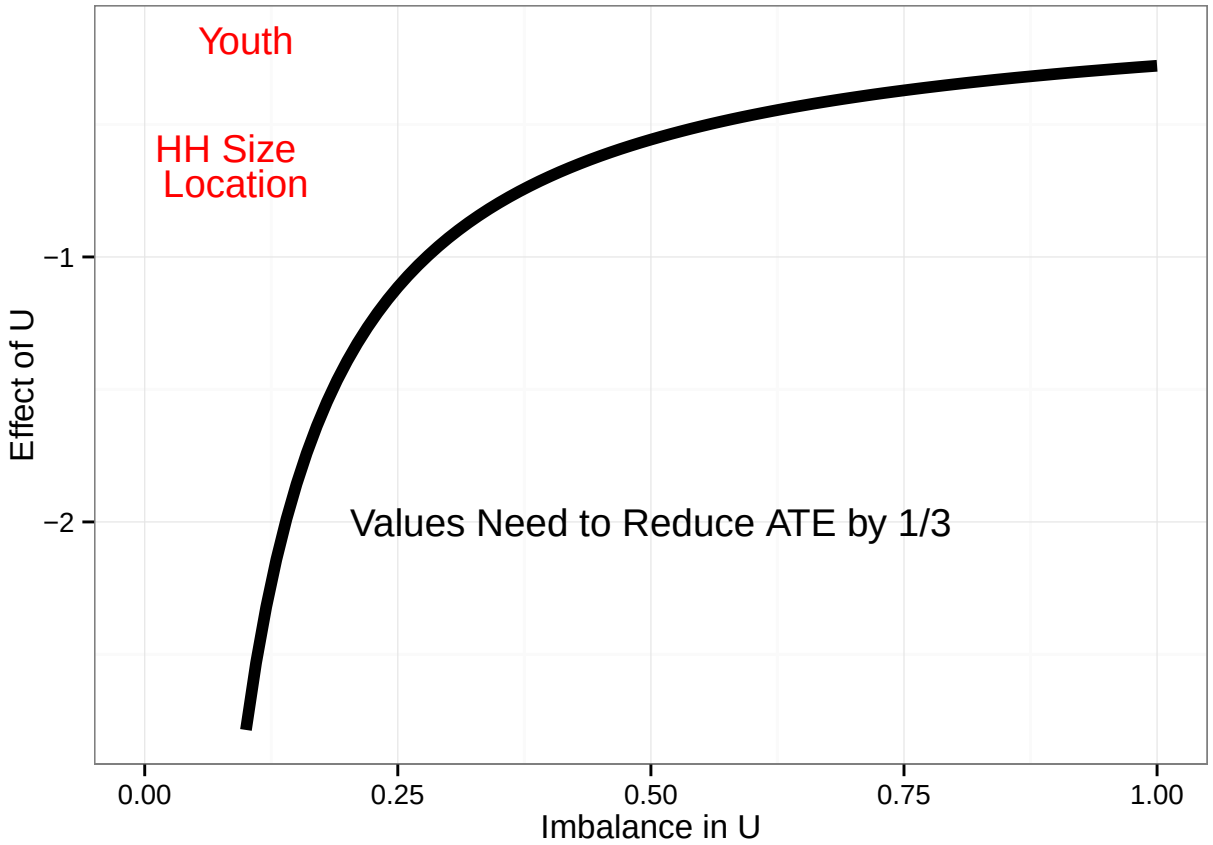
Example: Blattman and Annan on Child Soldiers

Plotting implied true ATE as function of assumed δ and γ:



Example: Blattman and Annan on Child Soldiers

Another way of plotting:

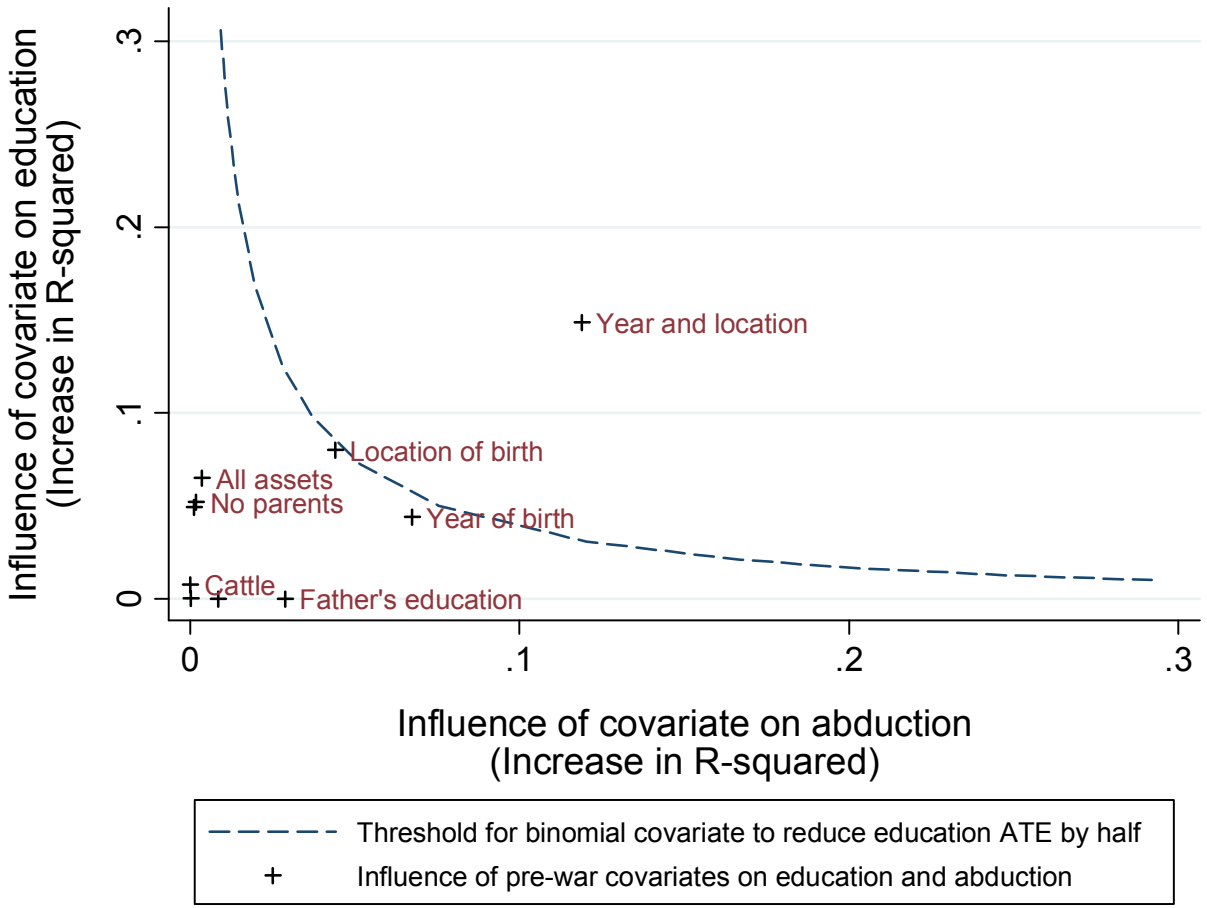


Note how we can use observed covariates as benchmarks.

Example: Blattman and Annan on Child Soldiers

A common approach is to standardize the scale by converting δ and γ to partial R^2s:

Figure 5: Impact of relaxing the assumption of unconfoundedness



Blattman and Annan (2010, ReStat)

Sensitivity Analysis for Randomization Inference

- Another sensitivity approach is developed in Rosenbaum (2002):
  - Uses a single sensitivity parameter $\Gamma \geq 1$ representing departure from unconfoundedness
  - Unlike Imbens' approach which targets ATE, Rosenbaum considers sharp null tests and p-values from randomization inference

Sensitivity Analysis for Randomization Inference

- Another sensitivity approach is developed in Rosenbaum (2002):
  - Uses a single sensitivity parameter $\Gamma \geq 1$ representing departure from unconfoundedness
  - Unlike Imbens' approach which targets ATE, Rosenbaum considers sharp null tests and p-values from randomization inference

Example: One-to-one exact matching without replacement

- Consider two matched units i and j with $X_i = X_j$
- Under conditional ignorability:
  - Both units must have the same treatment probability: $\pi(X_i) = \pi(X_j)$
  - Within the pair the treatment is as-if randomized

Sensitivity Analysis for Randomization Inference

- Another sensitivity approach is developed in Rosenbaum (2002):
  - Uses a single sensitivity parameter $\Gamma \geq 1$ representing departure from unconfoundedness
  - Unlike Imbens' approach which targets ATE, Rosenbaum considers sharp null tests and p-values from randomization inference

Example: One-to-one exact matching without replacement

- Consider two matched units i and j with $X_i = X_j$
- Under conditional ignorability:
  - Both units must have the same treatment probability: $\pi(X_i) = \pi(X_j)$
  - Within the pair the treatment is as-if randomized
- Without conditional ignorability:
  - The true treatment probability is a function of both X and unobserved confounders
  - That is, $\pi(X_i) > \pi(X_j)$ or $\pi(X_i) < \pi(X_j)$ even if $X_i = X_j$

Rosenbaum's Γ

- Quantify the degree of confounding by bounding the odds ratio by Γ:

$$\frac{1}{\Gamma} \leq \frac{\pi(X_i)/(1-\pi(X_i))}{\pi(X_j)/(1-\pi(X_j))} \leq \Gamma$$

  $\Gamma = 1$ no hidden bias, but if $\Gamma = 2$ unit i can be up to twice/half as likely to be treated than unit j (despite identical X)

Rosenbaum's Γ

- Quantify the degree of confounding by bounding the odds ratio by Γ:

$$\frac{1}{\Gamma} \leq \frac{\pi(X_i)/(1-\pi(X_i))}{\pi(X_j)/(1-\pi(X_j))} \leq \Gamma$$

  $\Gamma = 1$ no hidden bias, but if $\Gamma = 2$ unit i can be up to twice/half as likely to be treated than unit j (despite identical X)

Sensitivity analysis procedure for pair matching:

(1) Set Γ to a certain level

(2) Calculate the max/min treatment assignment probabilities for the Γ:

$$\frac{1}{1+\Gamma} \leq \pi(X_i) \leq \frac{\Gamma}{1+\Gamma}$$

(3) With $\pi(X_i)$ set to values most in favor of the null for each i, do a randomization test and record the p-value
   - For one-to-one match w/o replacement with a continuous outcome, we typically use Wilcoxon's signed rank test

(4) Iterate through (1) - (3) with different Γ values

- Various versions exist for different matching methods, statistics, etc.

Wilcoxon's Signed Rank Test

A test of the difference in medians for matched data:

1. Calculate the absolute difference $|\Delta_i|$ between Y_i and matched pair Y_j.
2. Rank the pairs in ascending order of absolute difference, $R_i = 1, 2, ..., N_R$.

   Drop pairs with $\Delta_i = 0$.

   Break ties by assigning the average of the pairs' ranks if not tied.
3. Sign the ranks with the sign of $Y_i - Y_j$, or $\text{sgn}(\Delta_i)R_i$
4. Calculate the sum of the positive signed ranks as a test statistic W

   $W = \sum_{i=1}^{N_{R^+}} R_i \quad \forall R_i > 0.$
5. Compare W to a critical value

Example: Exact Pair Matching w/o Replacement

Under conditional ignorability: $\Gamma = 1$, $\max \pi(X_i) = \min \pi(X_i) = 0.5$

i	Y_i	Y_j	Δ_i	$\lvert\Delta_i\rvert$	R_i	$\text{sgn}(\Delta_i)R_i$	Γ	worst $\pi(X_i)$
1	13	-3	16	16	4	4	1	.5
2	15	7	8	8	3	3	1	.5
3	-1	-4	3	3	2	2	1	.5
4	5	7	-2	2	1	-1	1	.5

Example: Exact Pair Matching w/o Replacement

Under conditional ignorability: $\Gamma = 1$, $\max \pi(X_i) = \min \pi(X_i) = 0.5$

i	Y_i	Y_j	Δ_i	$\lvert\Delta_i\rvert$	R_i	$\text{sgn}(\Delta_i)R_i$	Γ	worst $\pi(X_i)$
1	13	-3	16	16	4	4	1	.5
2	15	7	8	8	3	3	1	.5
3	-1	-4	3	3	2	2	1	.5
4	5	7	-2	2	1	-1	1	.5

- Wilcoxon statistic: $W = 4 + 3 + 2 = 9$
- Randomization distribution of W:

$$W \in \{0, 1, 2, ...9, 10\} \text{ with probability } \frac{1}{16} \text{ for each event}$$

- p-value for the sharp null is: $p = \Pr(W \geq 9 \mid H_0) = 0.125$

Example: Exact Pair Matching w/o Replacement

With unobserved confounding: $\Gamma = 2$, $\max \pi(X_i) = 0.67$, $\min \pi(X_i) = 0.33$

i	Y_i	Y_j	Δ_i	$\lvert\Delta_i\rvert$	R_i	$\text{sgn}(\Delta_i)R_i$	Γ	worst $\pi(X_i)$
1	13	-3	16	16	4	4	2	.67
2	15	7	8	8	3	3	2	.67
3	-1	-4	3	3	2	2	2	.67
4	5	7	-2	2	1	-1	2	.33

Example: Exact Pair Matching w/o Replacement

With unobserved confounding: $\Gamma = 2$, $\max \pi(X_i) = 0.67$, $\min \pi(X_i) = 0.33$

i	Y_i	Y_j	Δ_i	$\lvert\Delta_i\rvert$	R_i	$\text{sgn}(\Delta_i)R_i$	Γ	worst $\pi(X_i)$
1	13	-3	16	16	4	4	2	.67
2	15	7	8	8	3	3	2	.67
3	-1	-4	3	3	2	2	2	.67
4	5	7	-2	2	1	-1	2	.33

- Wilcoxon statistic: $W = 4 + 3 + 2 = 9$
- Randomization distribution of W:

$$W \in \{0, 1, 2, ...9, 10\}$$

with probabilities

$$\left(\frac{1}{3}\right)^4, \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^3, ..., \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^3, \left(\frac{2}{3}\right)^4 = \frac{1}{81}, \frac{2}{81}, ..., \frac{8}{81}, \frac{16}{81}$$

- max p-value for the sharp null is: $p = \Pr(W \geq 9 \mid H_0) = 0.296$

Example: Exact Pair Matching w/o Replacement

With unobserved confounding: $\Gamma = 2$, $\max \pi(X_i) = 0.67$, $\min \pi(X_i) = 0.33$

i	Y_i	Y_j	Δ_i	$\lvert\Delta_i\rvert$	R_i	$\text{sgn}(\Delta_i)R_i$	Γ	best $\pi(X_i)$
1	13	-3	16	16	4	4	2	.33
2	15	7	8	8	3	3	2	.33
3	-1	-4	3	3	2	2	2	.33
4	5	7	-2	2	1	-1	2	.67

Example: Exact Pair Matching w/o Replacement

With unobserved confounding: $\Gamma = 2$, $\max \pi(X_i) = 0.67$, $\min \pi(X_i) = 0.33$

i	Y_i	Y_j	Δ_i	$\lvert\Delta_i\rvert$	R_i	$\text{sgn}(\Delta_i)R_i$	Γ	best $\pi(X_i)$
1	13	-3	16	16	4	4	2	.33
2	15	7	8	8	3	3	2	.33
3	-1	-4	3	3	2	2	2	.33
4	5	7	-2	2	1	-1	2	.67

- Wilcoxon statistic: $W = 4 + 3 + 2 = 9$
- Randomization distribution of W:

$$W \in \{0, 1, 2, ...9, 10\}$$

with probabilities

$$\left(\frac{2}{3}\right)^4, \left(\frac{1}{3}\right)\left(\frac{2}{3}\right)^3, ..., \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)^3, \left(\frac{1}{3}\right)^4 = \frac{16}{81}, \frac{8}{81}, ..., \frac{2}{81}, \frac{1}{81}$$

- min p-value for the sharp null is: $p = \Pr(W \geq 9 \mid H_0) = 0.037$

Example: Blattman Data

```
matched.data<-Match(Y=Y, Tr=Treat, X=X,replace=F)
psens(matched.data, Gamma=2, GammaInc=.1)

Rosenbaum Sensitivity Test for Wilcoxon Signed Rank P-
 Gamma Lower bound Upper bound
   1.0           0      0.0000
   1.1           0      0.0002
   1.2           0      0.0027
   1.3           0      0.0183
   1.4           0      0.0725
   1.5           0      0.1924
   1.6           0      0.3744
   1.7           0      0.5774
   1.8           0      0.7522
   1.9           0      0.8732
   2.0           0      0.9429
```