Module 5: Observational Studies

Fall 2021

Matthew Blackwell

Gov 2003 (Harvard)

Where are we? Where are we going?

- Up to now: experiments where design makes everything easier.

Where are we? Where are we going?

- Up to now: experiments where design makes everything easier.
- Now: what happens when do observational studies?

Where are we? Where are we going?

- Up to now: experiments where design makes everything easier.
- Now: what happens when do observational studies?
    - Start with identification, selection on observables, and DAGs.

Where are we? Where are we going?

- Up to now: experiments where design makes everything easier.
- Now: what happens when do observational studies?
    - Start with identification, selection on observables, and DAGs.
    - Rest of the course will cover different designs for observational studies.

1/ Identification in observational studies

Randomized experiment review

- **Experiment**: when the researcher controls the treatment assignment.

Randomized experiment review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.

Randomized experiment review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.

Randomized experiment review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- **Randomized experiment** is an experiment with two properties:

Randomized experiment review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- **Randomized experiment** is an experiment with two properties:

1. **Positivity**: assignment is probabilistic: $0 < \mathbb{P}(D_i = 1) < 1$

Randomized experiment review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- **Randomized experiment** is an experiment with two properties:

1. **Positivity**: assignment is probabilistic: $0 < \mathbb{P}(D_i = 1) < 1$
   - No deterministic assignment.

Randomized experiment review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- **Randomized experiment** is an experiment with two properties:

1. **Positivity**: assignment is probabilistic: $0 < \mathbb{P}(D_i = 1) < 1$
   - No deterministic assignment.
2. **Unconfoundedness**: $\mathbb{P}[D_i = 1 | \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1]$

Randomized experiment review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- **Randomized experiment** is an experiment with two properties:

1. **Positivity**: assignment is probabilistic: $0 < \mathbb{P}(D_i = 1) < 1$
   - No deterministic assignment.
2. **Unconfoundedness**: $\mathbb{P}[D_i = 1|\mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1]$
   - Treatment assignment does not depend on any potential outcomes.

Randomized experiment review

- **Experiment**: when the researcher controls the treatment assignment.
  - $p_i = \mathbb{P}[D_i = 1]$ be the probability of treatment assignment probability.
  - p_i is controlled and known by researcher in an experiment.
- **Randomized experiment** is an experiment with two properties:

1. **Positivity**: assignment is probabilistic: $0 < \mathbb{P}(D_i = 1) < 1$
   - No deterministic assignment.
2. **Unconfoundedness**: $\mathbb{P}[D_i = 1 | \mathbf{Y}(1), \mathbf{Y}(0)] = \mathbb{P}[D_i = 1]$
   - Treatment assignment does not depend on any potential outcomes.
   - Sometimes written as $D_i \perp\!\!\!\perp (\mathbf{Y}(1), \mathbf{Y}(0))$

What is the selection problem?

- What if we **observe** a non-randomized treatment?

What is the selection problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i related to POs

What is the selection problem?

- What if we **observe** a non-randomized treatment?
    - Maybe treatment assignment is **confounded** so D_i related to POs
- What can we learn about the ATE here? Look at the difference-in-means:

What is the selection problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i related to POs
- What can we learn about the ATE here? Look at the difference-in-means:

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0]$$

What is the selection problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i related to POs
- What can we learn about the ATE here? Look at the difference-in-means:

$$
\begin{aligned}
&\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
=&\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \qquad \text{(consistency)}
\end{aligned}
$$

What is the selection problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i related to POs
- What can we learn about the ATE here? Look at the difference-in-means:

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

What is the selection problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i related to POs
- What can we learn about the ATE here? Look at the difference-in-means:

$$
\begin{aligned}
& \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
=& \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
=& \mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1] + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
=& \underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}
\end{aligned}
$$

What is the selection problem?

- What if we **observe** a non-randomized treatment?
  - Maybe treatment assignment is **confounded** so D_i related to POs
- What can we learn about the ATE here? Look at the difference-in-means:

$$
\begin{aligned}
&\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
=&\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
=&\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1] + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
=&\underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}
\end{aligned}
$$

- Without unconfoundedness: Naive diff-in-means = PATT + selection bias.

What is the selection problem?

- What if we **observe** a non-randomized treatment?
    - Maybe treatment assignment is **confounded** so D_i related to POs
- What can we learn about the ATE here? Look at the difference-in-means:

$$
\begin{aligned}
&\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
=&\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
=&\mathbb{E}[Y_i(1)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 1] + \mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0] \\
=&\underbrace{\mathbb{E}[Y_i(1) - Y_i(0)|D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}
\end{aligned}
$$

- Without unconfoundedness: Naive diff-in-means = PATT + selection bias.
- **Selection bias**: how different the treated and control groups are in terms of their potential outcome under control.

Selection bias = unidentified ATT

$$\mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0] = \underbrace{\tau_t}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0) | D_i = 1] - \mathbb{E}[Y_i(0) | D_i = 0]}_{\text{selection bias}}$$

- Difference in means: combination of two unknown quantities.

Selection bias = unidentified ATT

$$\mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0] = \underbrace{\tau_t}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0) | D_i = 1] - \mathbb{E}[Y_i(0) | D_i = 0]}_{\text{selection bias}}$$

- Difference in means: combination of two unknown quantities.
    - Can't distinguish if a diff-in-means is the ATT or selection bias.

Selection bias = unidentified ATT

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \underbrace{\tau_t}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

- Difference in means: combination of two unknown quantities.
  - Can't distinguish if a diff-in-means is the ATT or selection bias.
- Example: effect of negative ads on vote shares.

Selection bias = unidentified ATT

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \underbrace{\tau_t}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

- Difference in means: combination of two unknown quantities.
    - Can't distinguish if a diff-in-means is the ATT or selection bias.
- Example: effect of negative ads on vote shares.
    - Naive estimate: negative candidates do worse than positive candidates.

Selection bias = unidentified ATT

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \underbrace{\tau_t}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

- Difference in means: combination of two unknown quantities.
  - Can't distinguish if a diff-in-means is the ATT or selection bias.
- Example: effect of negative ads on vote shares.
  - Naive estimate: negative candidates do worse than positive candidates.
  - $\rightsquigarrow$ negative ATT **OR** positive ATT with large negative selection bias.

Selection bias = unidentified ATT

$$\mathbb{E}[Y_i|D_i=1] - \mathbb{E}[Y_i|D_i=0] = \underbrace{\tau_t}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i=1] - \mathbb{E}[Y_i(0)|D_i=0]}_{\text{selection bias}}$$

- Difference in means: combination of two unknown quantities.
  - Can't distinguish if a diff-in-means is the ATT or selection bias.
- Example: effect of negative ads on vote shares.
  - Naive estimate: negative candidates do worse than positive candidates.
  - $\rightsquigarrow$ negative ATT **OR** positive ATT with large negative selection bias.
  - SB = candidates that go negative are worse than those who stay positive, even if they ran the same campaigns.

Selection bias = unidentified ATT

$$\mathbb{E}[Y_i|D_i=1]-\mathbb{E}[Y_i|D_i=0]=\underbrace{\tau_t}_{\text{ATT}}+\underbrace{\mathbb{E}[Y_i(0)|D_i=1]-\mathbb{E}[Y_i(0)|D_i=0]}_{\text{selection bias}}$$

- Difference in means: combination of two unknown quantities.
  - Can't distinguish if a diff-in-means is the ATT or selection bias.
- Example: effect of negative ads on vote shares.
  - Naive estimate: negative candidates do worse than positive candidates.
  - $\rightsquigarrow$ negative ATT **OR** positive ATT with large negative selection bias.
  - SB = candidates that go negative are worse than those who stay positive, even if they ran the same campaigns.
- With an unbounded Y_i, we can't even bound the ATT because, in principle, SB could be anywhere from $-\infty$ to ∞.

Selection bias = unidentified ATT

$$\mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] = \underbrace{\tau_t}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_i(0)|D_i = 1] - \mathbb{E}[Y_i(0)|D_i = 0]}_{\text{selection bias}}$$

- Difference in means: combination of two unknown quantities.
    - Can't distinguish if a diff-in-means is the ATT or selection bias.
- Example: effect of negative ads on vote shares.
    - Naive estimate: negative candidates do worse than positive candidates.
    - $\rightsquigarrow$ negative ATT **OR** positive ATT with large negative selection bias.
    - SB = candidates that go negative are worse than those who stay positive, even if they ran the same campaigns.
- With an unbounded Y_i, we can't even bound the ATT because, in principle, SB could be anywhere from $-\infty$ to ∞.
- We say ATT (and ATE) are **unidentified** without further assumptions.

What is identification?

- **Identification** connects the counterfactual to the observed.

What is identification?

- **Identification** connects the counterfactual to the observed.
    - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.

What is identification?

- **Identification** connects the counterfactual to the observed.
    - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
    - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.

What is identification?

- **Identification** connects the counterfactual to the observed.
  - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
  - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
  - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$

What is identification?

- **Identification** connects the counterfactual to the observed.
  - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
  - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
  - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$
  - We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!

What is identification?

- **Identification** connects the counterfactual to the observed.
    - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
    - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
    - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$
    - We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!
- Quantity ψ is **identified** if we can write it as function of $\mathbb{P}$.

What is identification?

- **Identification** connects the counterfactual to the observed.
  - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
  - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
  - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$
  - We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!
- Quantity ψ is **identified** if we can write it as function of $\mathbb{P}$.
  - Would we know this quantity if we had access to unlimited data?

What is identification?

- **Identification** connects the counterfactual to the observed.
  - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
  - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
  - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$
  - We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!
- Quantity ψ is **identified** if we can write it as function of $\mathbb{P}$.
  - Would we know this quantity if we had access to unlimited data?
  - $\rightsquigarrow$ no worrying about estimation uncertainty here.

What is identification?

- **Identification** connects the counterfactual to the observed.
  - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
  - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
  - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$
  - We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!
- Quantity ψ is **identified** if we can write it as function of $\mathbb{P}$.
  - Would we know this quantity if we had access to unlimited data?
  - $\rightsquigarrow$ no worrying about estimation uncertainty here.
- Connecting counterfactual to the observational requires **assumptions**.

What is identification?

- **Identification** connects the counterfactual to the observed.
    - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
    - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
    - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$
    - We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!
- Quantity ψ is **identified** if we can write it as function of $\mathbb{P}$.
    - Would we know this quantity if we had access to unlimited data?
    - $\leadsto$ no worrying about estimation uncertainty here.
- Connecting counterfactual to the observational requires **assumptions**.
    - **"What's your identification strategy?"** = what are the assumptions that allow you to claim you've estimated a causal effect?

What is identification?

- **Identification** connects the counterfactual to the observed.
  - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
  - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
  - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$
  - We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!
- Quantity ψ is **identified** if we can write it as function of $\mathbb{P}$.
  - Would we know this quantity if we had access to unlimited data?
  - $\rightsquigarrow$ no worrying about estimation uncertainty here.
- Connecting counterfactual to the observational requires **assumptions**.
  - **"What's your identification strategy?"** = what are the assumptions that allow you to claim you've estimated a causal effect?
  - Research design can help justify assumptions (experiments, RDD, etc)

What is identification?

- **Identification** connects the counterfactual to the observed.
  - **Counterfactual distribution** $\mathbb{P}^*$ of $\{Y_i(1), Y_i(0), D_i, \mathbf{X}_i\}$.
  - **Observational distribution** $\mathbb{P}$ of $\{Y_i, D_i, \mathbf{X}_i\}$.
  - Causal quantities are functions of $\mathbb{P}^*$, but we get samples from $\mathbb{P}$
  - We can only learn about $\mathbb{P}^*$ through $\mathbb{P}$!
- Quantity ψ is **identified** if we can write it as function of $\mathbb{P}$.
  - Would we know this quantity if we had access to unlimited data?
  - $\leadsto$ no worrying about estimation uncertainty here.
- Connecting counterfactual to the observational requires **assumptions**.
  - **"What's your identification strategy?"** = what are the assumptions that allow you to claim you've estimated a causal effect?
  - Research design can help justify assumptions (experiments, RDD, etc)
  - Or you will have justify them through argument.

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
    - If identified, we know our causal parameter is some function of $\mathbb{P}$.

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, we worked with the **population** diff-in-means:

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

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, we worked with the **population** diff-in-means:

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

  - But $\mathbb{P}$ is not directly observable! It's a population distribution!

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, we worked with the **population** diff-in-means:

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

  - But $\mathbb{P}$ is not directly observable! It's a population distribution!
- Once identified, we need to actually **estimate** functions of $\mathbb{P}$.

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, we worked with the **population** diff-in-means:

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

  - But $\mathbb{P}$ is not directly observable! It's a population distribution!
- Once identified, we need to actually **estimate** functions of $\mathbb{P}$.
  - $\widehat{\tau}_{\text{diff}}$ is an estimator for population diff-in-means

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, we worked with the **population** diff-in-means:

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

  - But $\mathbb{P}$ is not directly observable! It's a population distribution!
- Once identified, we need to actually **estimate** functions of $\mathbb{P}$.
  - $\widehat{\tau}_{\text{diff}}$ is an estimator for population diff-in-means
  - Now just estimating conditional expectations, etc

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, we worked with the **population** diff-in-means:

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

  - But $\mathbb{P}$ is not directly observable! It's a population distribution!
- Once identified, we need to actually **estimate** functions of $\mathbb{P}$.
  - $\widehat{\tau}_{\text{diff}}$ is an estimator for population diff-in-means
  - Now just estimating conditional expectations, etc
  - $\rightsquigarrow$ **after identification, causal inference part done**

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, we worked with the **population** diff-in-means:

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

  - But $\mathbb{P}$ is not directly observable! It's a population distribution!
- Once identified, we need to actually **estimate** functions of $\mathbb{P}$.
  - $\widehat{\tau}_{\text{diff}}$ is an estimator for population diff-in-means
  - Now just estimating conditional expectations, etc
  - $\rightsquigarrow$ **after identification, causal inference part done**
  - Purely a statistical question from here on out.

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
    - If identified, we know our causal parameter is some function of $\mathbb{P}$.
    - For example, we worked with the **population** diff-in-means:

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

    - But $\mathbb{P}$ is not directly observable! It's a population distribution!
- Once identified, we need to actually **estimate** functions of $\mathbb{P}$.
    - $\widehat{\tau}_{\text{diff}}$ is an estimator for population diff-in-means
    - Now just estimating conditional expectations, etc
    - $\rightsquigarrow$ **after identification, causal inference part done**
    - Purely a statistical question from here on out.
- Identification comes first, then comes estimation.

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, we worked with the **population** diff-in-means:

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

  - But $\mathbb{P}$ is not directly observable! It's a population distribution!
- Once identified, we need to actually **estimate** functions of $\mathbb{P}$.
  - $\widehat{\tau}_{\text{diff}}$ is an estimator for population diff-in-means
  - Now just estimating conditional expectations, etc
  - $\rightsquigarrow$ **after identification, causal inference part done**
  - Purely a statistical question from here on out.
- Identification comes first, then comes estimation.
  - Without identification, properties of the estimator are unimportant.

Identification versus estimation

- Identification tells us **what** to estimate, not **how**.
  - If identified, we know our causal parameter is some function of $\mathbb{P}$.
  - For example, we worked with the **population** diff-in-means:

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

  - But $\mathbb{P}$ is not directly observable! It's a population distribution!
- Once identified, we need to actually **estimate** functions of $\mathbb{P}$.
  - $\widehat{\tau}_{\text{diff}}$ is an estimator for population diff-in-means
  - Now just estimating conditional expectations, etc
  - $\rightsquigarrow$ **after identification, causal inference part done**
  - Purely a statistical question from here on out.
- Identification comes first, then comes estimation.
  - Without identification, properties of the estimator are unimportant.
  - Keep them separate: estimator shouldn't drive identification.

What is confounding?

- **Confounding**: treatment and potential outcomes are not independent.

What is confounding?

- **Confounding**: treatment and potential outcomes are not independent.
    - Usually because of “common causes” of Y_i and D_i.

What is confounding?

- **Confounding**: treatment and potential outcomes are not independent.
    - Usually because of "common causes" of Y_i and D_i.
    - Main worry in observational studies.

What is confounding?

- **Confounding**: treatment and potential outcomes are not independent.
  - Usually because of "common causes" of Y_i and D_i.
  - Main worry in observational studies.
- Pervasive in the social sciences:

What is confounding?

- **Confounding**: treatment and potential outcomes are not independent.
  - Usually because of “common causes” of Y_i and D_i.
  - Main worry in observational studies.
- Pervasive in the social sciences:
  - effect of income on voting (confounder: age)

What is confounding?

- **Confounding**: treatment and potential outcomes are not independent.
    - Usually because of “common causes” of Y_i and D_i.
    - Main worry in observational studies.
- Pervasive in the social sciences:
    - effect of income on voting (confounder: age)
    - effect of job training program on employment (confounder: motivation)

What is confounding?

- **Confounding**: treatment and potential outcomes are not independent.
    - Usually because of “common causes” of Y_i and D_i.
    - Main worry in observational studies.
- Pervasive in the social sciences:
    - effect of income on voting (confounder: age)
    - effect of job training program on employment (confounder: motivation)
    - effect of political institutions on economic development (confounder: previous economic development)

What is confounding?

- **Confounding**: treatment and potential outcomes are not independent.
    - Usually because of “common causes” of Y_i and D_i.
    - Main worry in observational studies.
- Pervasive in the social sciences:
    - effect of income on voting (confounder: age)
    - effect of job training program on employment (confounder: motivation)
    - effect of political institutions on economic development (confounder: previous economic development)
- Confounding $\rightsquigarrow$ unidentified ATE $\rightsquigarrow$ biased and inconsistent estimators.

What is confounding?

- **Confounding**: treatment and potential outcomes are not independent.
  - Usually because of “common causes” of Y_i and D_i.
  - Main worry in observational studies.
- Pervasive in the social sciences:
  - effect of income on voting (confounder: age)
  - effect of job training program on employment (confounder: motivation)
  - effect of political institutions on economic development (confounder: previous economic development)
- Confounding $\rightsquigarrow$ unidentified ATE $\rightsquigarrow$ biased and inconsistent estimators.
- What to do?

2/ Selection on observables

Observational studies

- Many different sets of identification assumptions that we'll cover.

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability, selection on observables

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables, exogenous

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables, exogenous, conditional exchangeable, etc.

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables, exogenous, conditional exchangeable, etc.
    - Conditional on some covariates, D_i is (effectively) randomly assigned.

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables, exogenous, conditional exchangeable, etc.
    - Conditional on some covariates, D_i is (effectively) randomly assigned.
2. **Positivity** or **overlap**: $0 < \mathbb{P}[D_i = 1 | \mathbf{X}_i] < 1$

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables, exogenous, conditional exchangeable, etc.
    - Conditional on some covariates, D_i is (effectively) randomly assigned.
2. **Positivity** or **overlap**: $0 < \mathbb{P}[D_i = 1|\mathbf{X}_i] < 1$
    - Treatment and control are both possible at every value of $\mathbf{X}_i$.

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables, exogenous, conditional exchangeable, etc.
    - Conditional on some covariates, D_i is (effectively) randomly assigned.
2. **Positivity** or **overlap**: $0 < \mathbb{P}[D_i = 1|\mathbf{X}_i] < 1$
    - Treatment and control are both possible at every value of $\mathbf{X}_i$.

- We'll take $\mathbf{X}$ as given for now and see later how we might choose it.

Observational studies

- Many different sets of identification assumptions that we'll cover.
- Begin with most common observational assumption.

1. **No unmeasured confounding**: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
    - Also called: unconfoundedness, ignorability, selection on observables, no omitted variables, exogenous, conditional exchangeable, etc.
    - Conditional on some covariates, D_i is (effectively) randomly assigned.
2. **Positivity** or **overlap**: $0 < \mathbb{P}[D_i = 1|\mathbf{X}_i] < 1$
    - Treatment and control are both possible at every value of $\mathbf{X}_i$.

- We'll take $\mathbf{X}$ as given for now and see later how we might choose it.
- These are assumptions that **can be wrong**!!

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$\begin{aligned}\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) - Y_i(0) \mid \mathbf{X}_i]\right\}\end{aligned}$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$\begin{aligned}\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) - Y_i(0) \mid \mathbf{X}_i]\right\} \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]\right\}\end{aligned}$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) - Y_i(0) \mid \mathbf{X}_i]\right\} \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]\right\} \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid D_i = 0, \mathbf{X}_i]\right\}
\end{aligned}
$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$
\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) - Y_i(0) \mid \mathbf{X}_i]\right\} \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]\right\} \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid D_i = 0, \mathbf{X}_i]\right\} \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i]\right\}
\end{aligned}
$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$\begin{aligned}
\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) - Y_i(0) \mid \mathbf{X}_i]\right\} \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]\right\} \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid D_i = 0, \mathbf{X}_i]\right\} \\
&= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i]\right\}
\end{aligned}$$

- Useful to write the treated and control CEFs:

$$\mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}], \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}]$$

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$\begin{aligned}\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) - Y_i(0) \mid \mathbf{X}_i]\right\} \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]\right\} \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid D_i = 0, \mathbf{X}_i]\right\} \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i]\right\}\end{aligned}$$

- Useful to write the treated and control CEFs:

$$\mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}], \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}]$$

  - How the mean of the potential outcomes vary with the covariates.

Identification of the ATE

- Positivity and no unmeasured confounders will identify the PATE:

$$\begin{aligned}\tau &= \mathbb{E}[Y_i(1) - Y_i(0)] \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) - Y_i(0) \mid \mathbf{X}_i]\right\} \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]\right\} \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i(1) \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid D_i = 0, \mathbf{X}_i]\right\} \\ &= \mathbb{E}_{\mathbf{X}}\left\{E[Y_i \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i]\right\}\end{aligned}$$

- Useful to write the treated and control CEFs:

$$\mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}], \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}]$$

  - How the mean of the potential outcomes vary with the covariates.

- Key part of the above proof:

$$\underbrace{\mu_1(\mathbf{x})}_{\text{counterfactual}} = \underbrace{\mathbb{E}[Y_i \mid D_i = 1, \mathbf{X}_i = \mathbf{x}]}_{\text{observational}}, \qquad \mu_0(\mathbf{x}) = \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i = \mathbf{x}]$$

Regression estimation of the ATE

- Identification done, estimation has just begun!

Regression estimation of the ATE

- Identification done, estimation has just begun!
- Regression estimators $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$.

Regression estimation of the ATE

- Identification done, estimation has just begun!
- Regression estimators $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$.
    - Might be linear or nonlinear models

Regression estimation of the ATE

- Identification done, estimation has just begun!
- Regression estimators $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$.
    - Might be linear or nonlinear models
    - Safest practice: estimate separate regressions in each treatment group.

Regression estimation of the ATE

- Identification done, estimation has just begun!
- Regression estimators $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$.
  - Might be linear or nonlinear models
  - Safest practice: estimate separate regressions in each treatment group.
- Regression estimator of the ATE:

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

Regression estimation of the ATE

- Identification done, estimation has just begun!
- Regression estimators $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$.
    - Might be linear or nonlinear models
    - Safest practice: estimate separate regressions in each treatment group.
- Regression estimator of the ATE:

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Procedure:

Regression estimation of the ATE

- Identification done, estimation has just begun!
- Regression estimators $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$.
    - Might be linear or nonlinear models
    - Safest practice: estimate separate regressions in each treatment group.
- Regression estimator of the ATE:

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Procedure:
    - Obtain predicted values for all units when $D_i = 1$.

Regression estimation of the ATE

- Identification done, estimation has just begun!
- Regression estimators $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$.
  - Might be linear or nonlinear models
  - Safest practice: estimate separate regressions in each treatment group.
- Regression estimator of the ATE:

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Procedure:
  - Obtain predicted values for all units when $D_i = 1$.
  - Obtain predicted values for all units when $D_i = 0$.

Regression estimation of the ATE

- Identification done, estimation has just begun!
- Regression estimators $\widehat{\mu}_1(\mathbf{x})$ and $\widehat{\mu}_0(\mathbf{x})$.
  - Might be linear or nonlinear models
  - Safest practice: estimate separate regressions in each treatment group.
- Regression estimator of the ATE:

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

- Procedure:
  - Obtain predicted values for all units when $D_i = 1$.
  - Obtain predicted values for all units when $D_i = 0$.
  - Take the average difference between these predicted values.

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\mu}_1(\mathbf{X}_i) - \widehat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:
  - $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are from the same OLS model `Y ~ D + X`.

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:
  - $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are from the same OLS model `Y ~ D + X`.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:
  - $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are from the same OLS model `Y ~ D + X`.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i
- Fully interacted OLS:

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:
  - $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are from the same OLS model `Y ~ D + X`.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i
- Fully interacted OLS:
  - $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are from fully interacted OLS with centered covariates.

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:
  - $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are from the same OLS model `Y ~ D + X`.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i
- Fully interacted OLS:
  - $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are from fully interacted OLS with centered covariates.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i

Coefficients?

$$\widehat{\tau}_{\text{reg}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_i)$$

- Under linear models, $\widehat{\tau}_{\text{reg}}$ is sometimes equivalent to a coefficient.
- Uninteracted OLS:
  - $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are from the same OLS model `Y ~ D + X`.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i
- Fully interacted OLS:
  - $\hat{\mu}_1(\mathbf{x})$ and $\hat{\mu}_0(\mathbf{x})$ are from fully interacted OLS with centered covariates.
  - $\widehat{\tau}_{\text{reg}} \equiv$ estimated coefficient on D_i
- These make two very different assumptions about the CEFs!

Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?

Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- If an OLS coefficient $\rightsquigarrow$ use EHW variance estimator.

Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- If an OLS coefficient $\rightsquigarrow$ use EHW variance estimator.
- Analytic expressions can be derived, but complicated!

Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- If an OLS coefficient $\rightsquigarrow$ use EHW variance estimator.
- Analytic expressions can be derived, but complicated!
- Computational alternative: **nonparametric bootstrap**

Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- If an OLS coefficient $\rightsquigarrow$ use EHW variance estimator.
- Analytic expressions can be derived, but complicated!
- Computational alternative: **nonparametric bootstrap**
    - Randomly resample n rows of the data with replacement

Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- If an OLS coefficient $\rightsquigarrow$ use EHW variance estimator.
- Analytic expressions can be derived, but complicated!
- Computational alternative: **nonparametric bootstrap**
    - Randomly resample n rows of the data with replacement
    - Refit the regressions on the bootstrapped data.

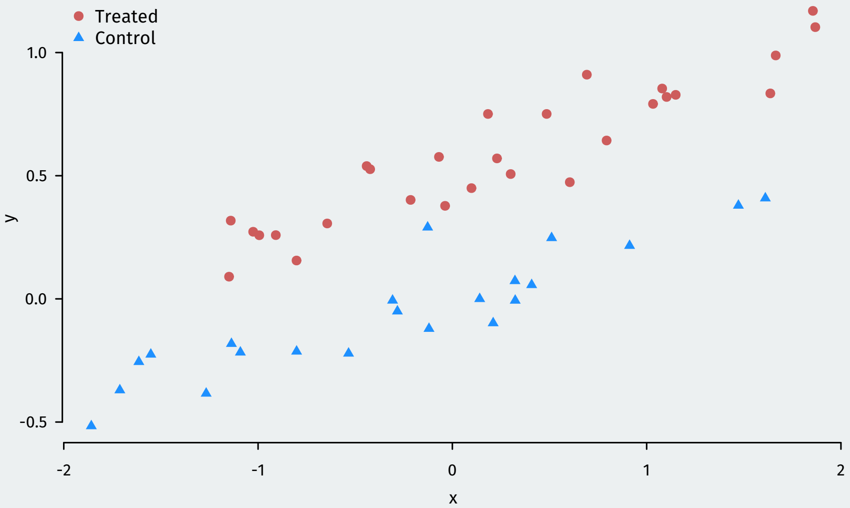
Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- If an OLS coefficient $\rightsquigarrow$ use EHW variance estimator.
- Analytic expressions can be derived, but complicated!
- Computational alternative: **nonparametric bootstrap**
    - Randomly resample n rows of the data with replacement
    - Refit the regressions on the bootstrapped data.
    - Calculate $\widehat{\tau}_{\text{reg}}$ in each bootstrap

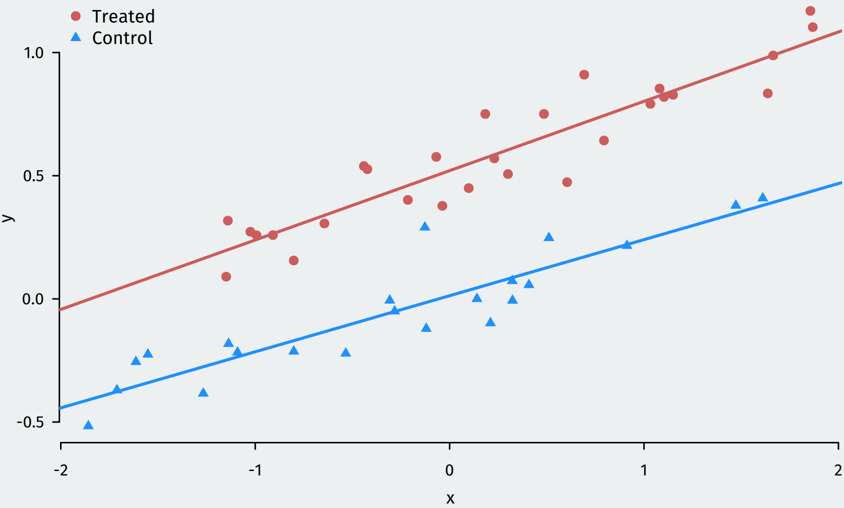
Variance estimation

- How do we get estimates of the variance of $\widehat{\tau}_{\text{reg}}$?
- If an OLS coefficient $\rightsquigarrow$ use EHW variance estimator.
- Analytic expressions can be derived, but complicated!
- Computational alternative: **nonparametric bootstrap**
    - Randomly resample n rows of the data with replacement
    - Refit the regressions on the bootstrapped data.
    - Calculate $\widehat{\tau}_{\text{reg}}$ in each bootstrap
    - Repeat several times and use empirical variance of the bootstraps

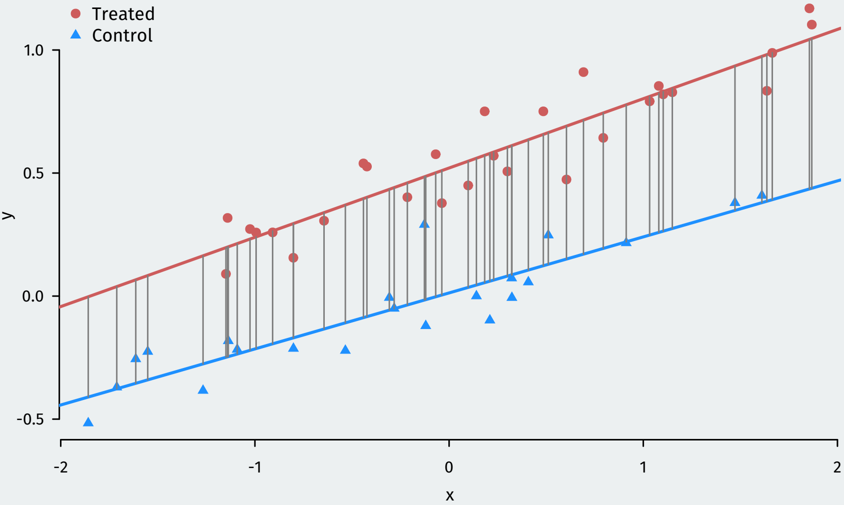
Imputation estimator visualization



Imputation estimator visualization

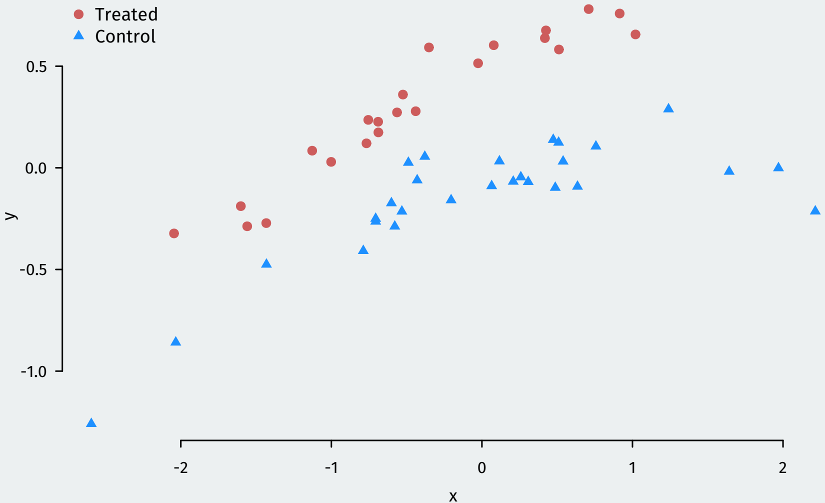


Imputation estimator visualization



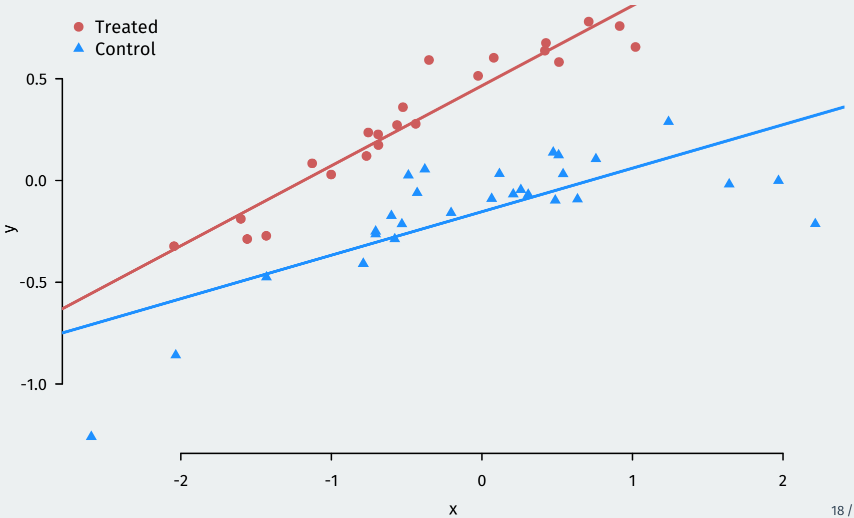
Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



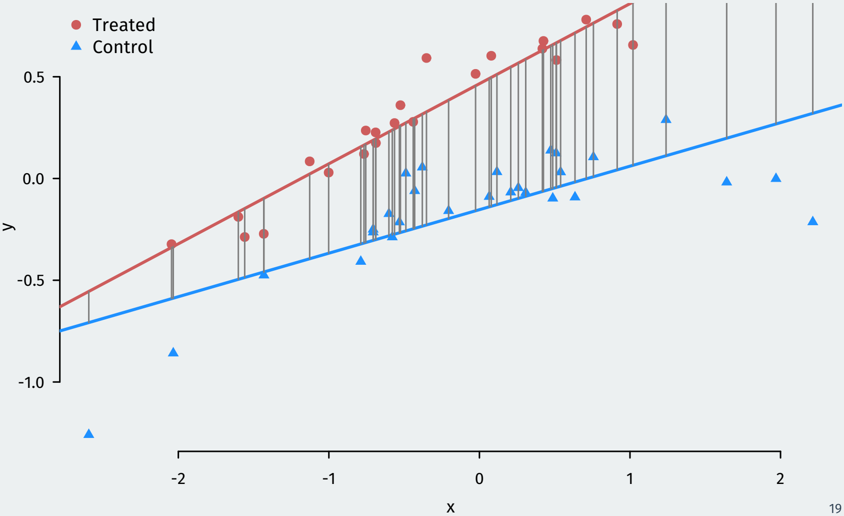
Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



Nonlinear relationships

- Same idea but with nonlinear relationship between Y_i and X_i:



Using semiparametric regression

- Here, CEFs are nonlinear, but we don't know their form.

Using semiparametric regression

- Here, CEFs are nonlinear, but we don't know their form.
- We can use GAMs from the `mgcv` package to for flexible estimate:

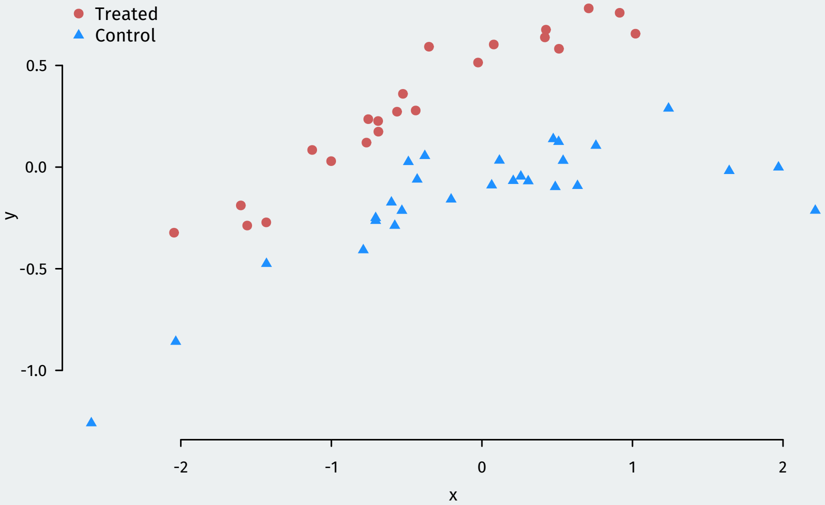
Using semiparametric regression

- Here, CEFs are nonlinear, but we don't know their form.
- We can use GAMs from the `mgcv` package to for flexible estimate:

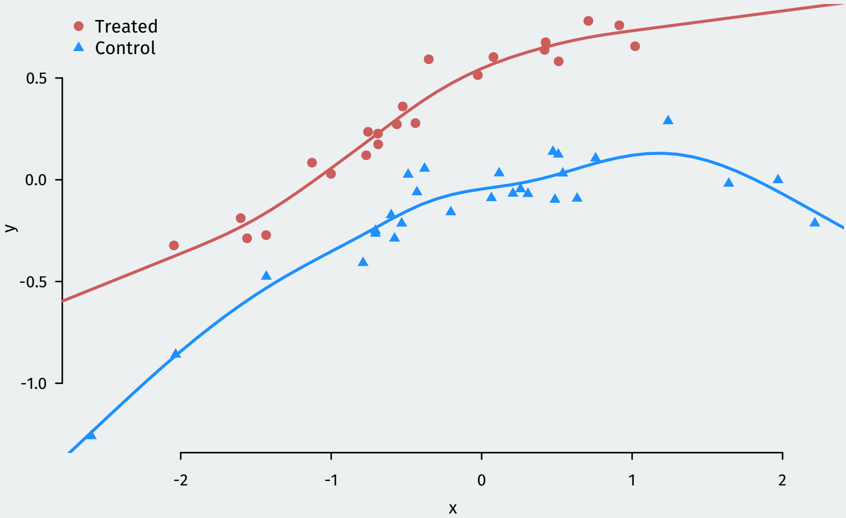
```
library(mgcv)
mod0 <- gam(y~s(x), subset = d==0)
summary(mod0)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ s(x)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.154      0.019    -8.1  5.1e-08 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##       edf Ref.df    F p-value
## s(x) 5.17   6.29 36.9  <2e-16 ***
## ---
```

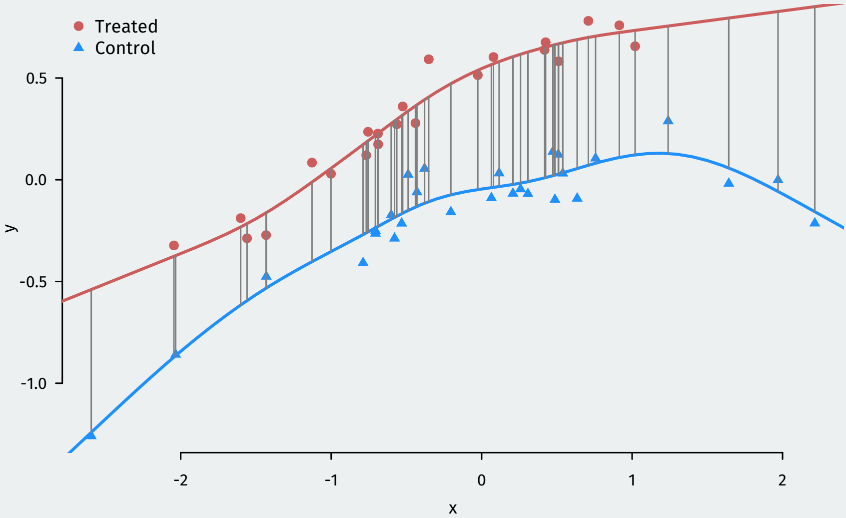
Using GAMs



Using GAMs



Using GAMs



3/ DAGs

Choosing the conditioning set

- How do we know if no unmeasured confounders holds?

Choosing the conditioning set

- How do we know if no unmeasured confounders holds?
- Put differently:

Choosing the conditioning set

- How do we know if no unmeasured confounders holds?
- Put differently:
    - What covariates do we need to condition on?

Choosing the conditioning set

- How do we know if no unmeasured confounders holds?
- Put differently:
    - What covariates do we need to condition on?
    - What covariates do we need to include in our regressions?

Choosing the conditioning set

- How do we know if no unmeasured confounders holds?
- Put differently:
  - What covariates do we need to condition on?
  - What covariates do we need to include in our regressions?
- One way, from the assumption itself: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$

Choosing the conditioning set

- How do we know if no unmeasured confounders holds?
- Put differently:
  - What covariates do we need to condition on?
  - What covariates do we need to include in our regressions?
- One way, from the assumption itself: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
  - Include covariates such that, conditional on them, the treatment assignment does not depend on the potential outcomes.

Choosing the conditioning set

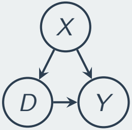
- How do we know if no unmeasured confounders holds?
- Put differently:
  - What covariates do we need to condition on?
  - What covariates do we need to include in our regressions?
- One way, from the assumption itself: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
  - Include covariates such that, conditional on them, the treatment assignment does not depend on the potential outcomes.
  - Somewhat circular

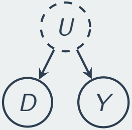
Choosing the conditioning set

- How do we know if no unmeasured confounders holds?
- Put differently:
  - What covariates do we need to condition on?
  - What covariates do we need to include in our regressions?
- One way, from the assumption itself: $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$
  - Include covariates such that, conditional on them, the treatment assignment does not depend on the potential outcomes.
  - Somewhat circular
- Another way: use DAGs and look at back-door paths.

Directed Acyclic Graphs

- **Directed acyclic graphs** (DAGs) describe the causal structure of variables





Directed Acyclic Graphs

- **Directed acyclic graphs** (DAGs) describe the causal structure of variables

- **Nodes/vertices**: observed (solid) or unobserved (dashed) variables.

Directed Acyclic Graphs

- **Directed acyclic graphs** (DAGs) describe the causal structure of variables

- **Nodes/vertices**: observed (solid) or unobserved (dashed) variables.
- **Edges**: arrows that encodes the presence or absence of a causal effect.

Directed Acyclic Graphs

- **Directed acyclic graphs** (DAGs) describe the causal structure of variables

- **Nodes/vertices**: observed (solid) or unobserved (dashed) variables.
- **Edges**: arrows that encodes the presence or absence of a causal effect.
  - Arrow present = a direct causal effect: $Y_i(d) \neq Y_i(d')$ for some i and d.

Directed Acyclic Graphs

- **Directed acyclic graphs** (DAGs) describe the causal structure of variables

- **Nodes/vertices**: observed (solid) or unobserved (dashed) variables.
- **Edges**: arrows that encodes the presence or absence of a causal effect.
  - Arrow present = a direct causal effect: $Y_i(d) \neq Y_i(d')$ for some i and d.
  - Lack of an arrow = no causal effect: $Y_i(d) = Y_i(d')$ for all i and d.

Directed Acyclic Graphs

- **Directed acyclic graphs** (DAGs) describe the causal structure of variables

- **Nodes/vertices**: observed (solid) or unobserved (dashed) variables.
- **Edges**: arrows that encodes the presence or absence of a causal effect.
  - Arrow present = a direct causal effect: $Y_i(d) \neq Y_i(d')$ for some i and d.
  - Lack of an arrow = no causal effect: $Y_i(d) = Y_i(d')$ for all i and d.
  - Missing variables = no other common causes of any variables.

Directed Acyclic Graphs

- **Directed acyclic graphs** (DAGs) describe the causal structure of variables

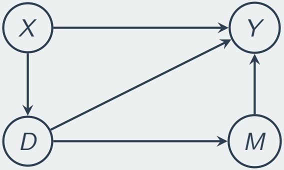
- **Nodes/vertices**: observed (solid) or unobserved (dashed) variables.
- **Edges**: arrows that encodes the presence or absence of a causal effect.
  - Arrow present = a direct causal effect: $Y_i(d) \neq Y_i(d')$ for some i and d.
  - Lack of an arrow = no causal effect: $Y_i(d) = Y_i(d')$ for all i and d.
  - Missing variables = no other common causes of any variables.
- **Directed**: each arrow implies a direction

Directed Acyclic Graphs

- **Directed acyclic graphs** (DAGs) describe the causal structure of variables

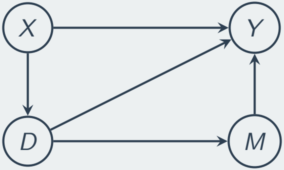
- **Nodes/vertices**: observed (solid) or unobserved (dashed) variables.
- **Edges**: arrows that encodes the presence or absence of a causal effect.
  - Arrow present = a direct causal effect: $Y_i(d) \neq Y_i(d')$ for some i and d.
  - Lack of an arrow = no causal effect: $Y_i(d) = Y_i(d')$ for all i and d.
  - Missing variables = no other common causes of any variables.
- **Directed**: each arrow implies a direction
- **Acyclic**: no cycles: a variable cannot cause itself

DAG terminology



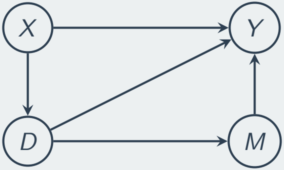
- **Path**: a sequence of edges that connect two nodes.

DAG terminology



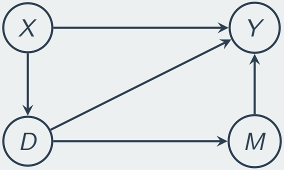
- **Path**: a sequence of edges that connect two nodes.
    - A **directed** or **causal** path is all in the same causal direction.

DAG terminology



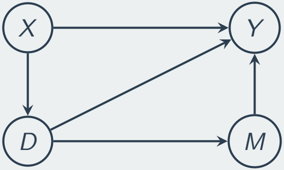
- **Path**: a sequence of edges that connect two nodes.
    - A **directed** or **causal** path is all in the same causal direction.
    - Non-causal path example: $D \leftarrow X \rightarrow Y$

DAG terminology



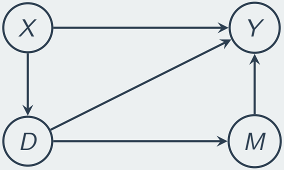
- **Path**: a sequence of edges that connect two nodes.
  - A **directed** or **causal** path is all in the same causal direction.
  - Non-causal path example: $D \leftarrow X \rightarrow Y$
- **Descendants**: nodes on a directed path away from some other node.

DAG terminology



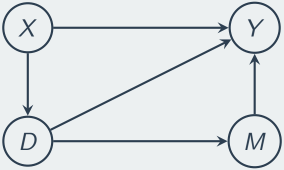
- **Path**: a sequence of edges that connect two nodes.
  - A **directed** or **causal** path is all in the same causal direction.
  - Non-causal path example: $D \leftarrow X \rightarrow Y$
- **Descendants**: nodes on a directed path away from some other node.
  - M is a descendant of D and X.

DAG terminology



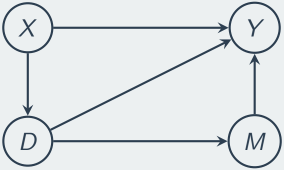
- **Path**: a sequence of edges that connect two nodes.
    - A **directed** or **causal** path is all in the same causal direction.
    - Non-causal path example: $D \leftarrow X \rightarrow Y$
- **Descendants**: nodes on a directed path away from some other node.
    - M is a descendant of D and X.
    - Ancestors is the reverse: X is an ancestor of M

DAG terminology



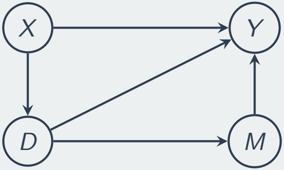
- **Path**: a sequence of edges that connect two nodes.
  - A **directed** or **causal** path is all in the same causal direction.
  - Non-causal path example: $D \leftarrow X \rightarrow Y$
- **Descendants**: nodes on a directed path away from some other node.
  - M is a descendant of D and X.
  - Ancestors is the reverse: X is an ancestor of M
- **Parents** immediate causes of a node

DAG terminology



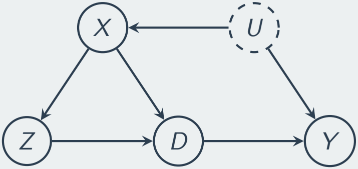
- **Path**: a sequence of edges that connect two nodes.
    - A **directed** or **causal** path is all in the same causal direction.
    - Non-causal path example: $D \leftarrow X \rightarrow Y$
- **Descendants**: nodes on a directed path away from some other node.
    - M is a descendant of D and X.
    - Ancestors is the reverse: X is an ancestor of M
- **Parents** immediate causes of a node
    - D is the parent of Y and M.

DAG terminology



- **Path**: a sequence of edges that connect two nodes.
    - A **directed** or **causal** path is all in the same causal direction.
    - Non-causal path example: $D \leftarrow X \rightarrow Y$
- **Descendants**: nodes on a directed path away from some other node.
    - M is a descendant of D and X.
    - Ancestors is the reverse: X is an ancestor of M
- **Parents** immediate causes of a node
    - D is the parent of Y and M.
    - **Children** are the reverse: M is a child of D

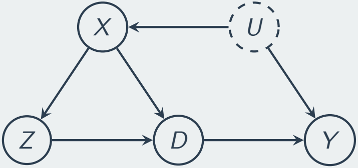
DAGs to distributions



$$Y = f_y(D, U, \varepsilon_y)$$
$$D = f_d(Z, X, \varepsilon_d)$$
$$X = f_x(U, \varepsilon_x)$$
$$Z = f_z(X, \varepsilon_z)$$

- Causal DAGs equivalent to **nonparametric structural equation models**

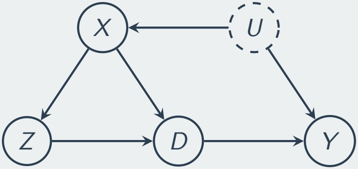
DAGs to distributions



$$Y = f_y(D, U, \varepsilon_y)$$
$$D = f_d(Z, X, \varepsilon_d)$$
$$X = f_x(U, \varepsilon_x)$$
$$Z = f_z(X, \varepsilon_z)$$

- Causal DAGs equivalent to **nonparametric structural equation models**
- NPSEM have a **causal interpreation**, but completely flexible.

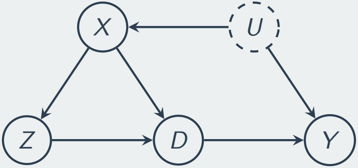
DAGs to distributions



$$Y = f_y(D, U, \varepsilon_y)$$
$$D = f_d(Z, X, \varepsilon_d)$$
$$X = f_x(U, \varepsilon_x)$$
$$Z = f_z(X, \varepsilon_z)$$

- Causal DAGs equivalent to **nonparametric structural equation models**
- NPSEM have a **causal interpreation**, but completely flexible.
    - No specification of a functional form or interactions, etc.

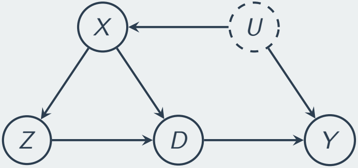
DAGs to distributions



$$Y = f_y(D, U, \varepsilon_y)$$
$$D = f_d(Z, X, \varepsilon_d)$$
$$X = f_x(U, \varepsilon_x)$$
$$Z = f_z(X, \varepsilon_z)$$

- Causal DAGs equivalent to **nonparametric structural equation models**
- NPSEM have a **causal interpreation**, but completely flexible.
    - No specification of a functional form or interactions, etc.
    - More standard linear SEM is a special case.

DAGs to distributions



$$Y = f_y(D, U, \varepsilon_y)$$
$$D = f_d(Z, X, \varepsilon_d)$$
$$X = f_x(U, \varepsilon_x)$$
$$Z = f_z(X, \varepsilon_z)$$

- Causal DAGs equivalent to **nonparametric structural equation models**
- NPSEM have a **causal interpreation**, but completely flexible.
    - No specification of a functional form or interactions, etc.
    - More standard linear SEM is a special case.
- Causal DAGs imply the following factorization (some conditions apply):

$$\mathbb{P}(X_1, X_2, \dots, X_J) = \prod_{j=1}^{J} \mathbb{P}(X_j \mid \text{pa}(X_j)) \quad \text{where} \quad \text{pa}(X_j) = \text{parents of } X_j$$

d-separation

- Can we determine conditional independence from our causal DAG?

d-separation

- Can we determine conditional independence from our causal DAG?
- Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:

d-separation

- Can we determine conditional independence from our causal DAG?
- Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
    1. Find all paths from any vertex in A to any vertex in B.

d-separation

- Can we determine conditional independence from our causal DAG?
- Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
    1. Find all paths from any vertex in A to any vertex in B.
    2. Check is each path is **blocked**.

d-separation

- Can we determine conditional independence from our causal DAG?
- Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
  1. Find all paths from any vertex in A to any vertex in B.
  2. Check is each path is **blocked**.
  3. If all paths are blocked, then A is **d-separated** from B by C

d-separation

- Can we determine conditional independence from our causal DAG?
- Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
  1. Find all paths from any vertex in A to any vertex in B.
  2. Check is each path is **blocked**.
  3. If all paths are blocked, then A is **d-separated** from B by C
- A path is **blocked** conditional on C if:

d-separation

- Can we determine conditional independence from our causal DAG?
- Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
  1. Find all paths from any vertex in A to any vertex in B.
  2. Check is each path is **blocked**.
  3. If all paths are blocked, then A is **d-separated** from B by C
- A path is **blocked** conditional on C if:
  1. C includes a non-collider on that path **OR**

d-separation

- Can we determine conditional independence from our causal DAG?
- Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
  1. Find all paths from any vertex in A to any vertex in B.
  2. Check is each path is **blocked**.
  3. If all paths are blocked, then A is **d-separated** from B by C
- A path is **blocked** conditional on C if:
  1. C includes a non-collider on that path **OR**
  2. Path includes a collider not in C and no descendant of any collider is in C.

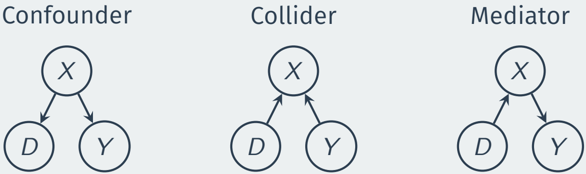
d-separation

- Can we determine conditional independence from our causal DAG?
- Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
  1. Find all paths from any vertex in A to any vertex in B.
  2. Check is each path is **blocked**.
  3. If all paths are blocked, then A is **d-separated** from B by C
- A path is **blocked** conditional on C if:
  1. C includes a non-collider on that path **OR**
  2. Path includes a collider not in C and no descendant of any collider is in C.
- If A and B are d-separated, then we have $A \perp\!\!\!\perp B \mid C$.

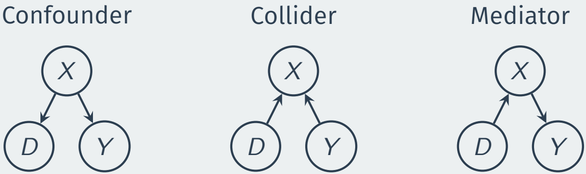
d-separation

- Can we determine conditional independence from our causal DAG?
- Yes! To verify that $A \perp\!\!\!\perp B \mid C$ where each is a set of nodes:
    1. Find all paths from any vertex in A to any vertex in B.
    2. Check is each path is **blocked**.
    3. If all paths are blocked, then A is **d-separated** from B by C
- A path is **blocked** conditional on C if:
    1. C includes a non-collider on that path **OR**
    2. Path includes a collider not in C and no descendant of any collider is in C.
- If A and B are d-separated, then we have $A \perp\!\!\!\perp B \mid C$.
    - If not, then d-connected and A and B dependence conditional on C is compatible with the DAG.

Common structures

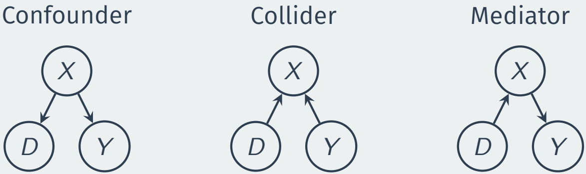


Common structures



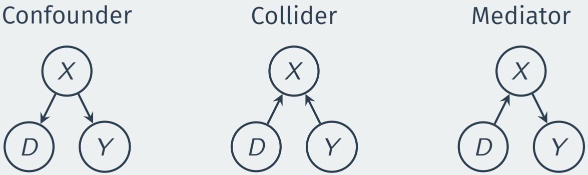
- **Confounder**: common cause of two variables.

Common structures



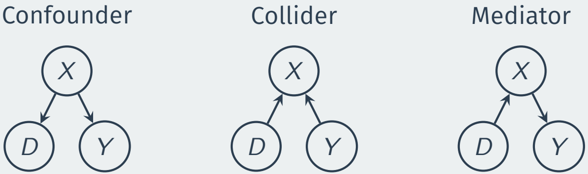
- **Confounder**: common cause of two variables.
    - D and Y unconditionally dependent, conditionally independent.

Common structures



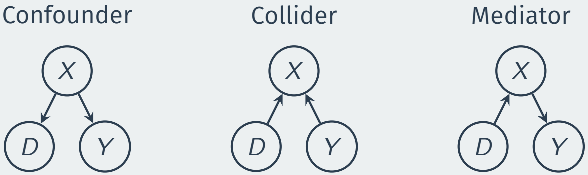
- **Confounder**: common cause of two variables.
    - D and Y unconditionally dependent, conditionally independent.
- **Collider**: common descendant of two variables.

Common structures



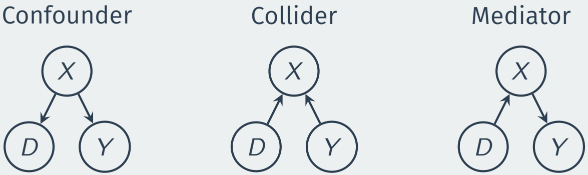
- **Confounder**: common cause of two variables.
  - D and Y unconditionally dependent, conditionally independent.
- **Collider**: common descendant of two variables.
  - D and Y unconditionally independent, conditionally dependent.

Common structures



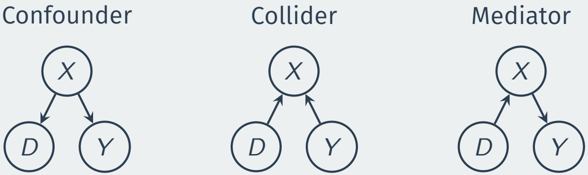
- **Confounder**: common cause of two variables.
  - D and Y unconditionally dependent, conditionally independent.
- **Collider**: common descendant of two variables.
  - D and Y unconditionally independent, conditionally dependent.
  - X “blocks” the relationship between them when not conditioned on.

Common structures



- **Confounder**: common cause of two variables.
    - D and Y unconditionally dependent, conditionally independent.
- **Collider**: common descendant of two variables.
    - D and Y unconditionally independent, conditionally dependent.
    - X “blocks” the relationship between them when not conditioned on.
- **Mediator**: variable on the path from one variable to another.

Common structures



- **Confounder**: common cause of two variables.
    - D and Y unconditionally dependent, conditionally independent.
- **Collider**: common descendant of two variables.
    - D and Y unconditionally independent, conditionally dependent.
    - X "blocks" the relationship between them when not conditioned on.
- **Mediator**: variable on the path from one variable to another.
    - D and Y unconditionally dependent.

Backdoor paths and blocking paths

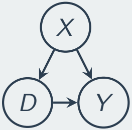
- **Backdoor path**: is a non-causal path from D to Y.

Backdoor paths and blocking paths

- **Backdoor path**: is a non-causal path from D to Y.
  - Would remain if we removed any arrows pointing out of D.

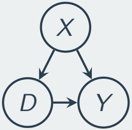
Backdoor paths and blocking paths

- **Backdoor path**: is a non-causal path from D to Y.
  - Would remain if we removed any arrows pointing out of D.
- Backdoor paths between D and Y $\rightsquigarrow$ common causes of D and Y:



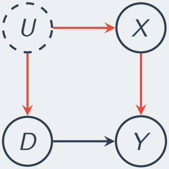
Backdoor paths and blocking paths

- **Backdoor path**: is a non-causal path from D to Y.
    - Would remain if we removed any arrows pointing out of D.
- Backdoor paths between D and $Y \rightsquigarrow$ common causes of D and Y:



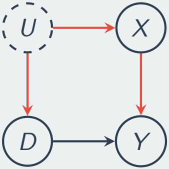
- Here: backdoor path $D \leftarrow X \rightarrow Y$

Other types of confounding



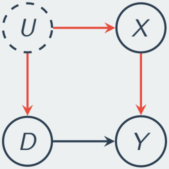
- D is enrolling in a job training program.

Other types of confounding



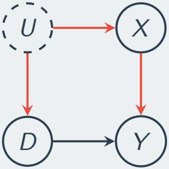
- D is enrolling in a job training program.
- Y is getting a job.

Other types of confounding



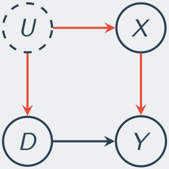
- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated

Other types of confounding



- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated
- X is number of job applications sent out.

Other types of confounding



- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated
- X is number of job applications sent out.
- Big assumption here: no arrow from U to Y.

Backdoor criterion

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

- Can we use a DAG to evaluate no unmeasured confounders?

Backdoor criterion

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

- Can we use a DAG to evaluate no unmeasured confounders?
- Holds if the **backdoor criterion** on a causal DAG is met:

Backdoor criterion

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

- Can we use a DAG to evaluate no unmeasured confounders?
- Holds if the **backdoor criterion** on a causal DAG is met:
    1. No vertex in **X** is a descend of *D* (**no post-treatment bias**), and

Backdoor criterion

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

- Can we use a DAG to evaluate no unmeasured confounders?
- Holds if the **backdoor criterion** on a causal DAG is met:
    1. No vertex in $\mathbf{X}$ is a descend of D (**no post-treatment bias**), and
    2. $\mathbf{X}$ blocks all backdoor paths from D to Y.

Backdoor criterion

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

- Can we use a DAG to evaluate no unmeasured confounders?
- Holds if the **backdoor criterion** on a causal DAG is met:
  1. No vertex in $\mathbf{X}$ is a descend of D (**no post-treatment bias**), and
  2. $\mathbf{X}$ blocks all backdoor paths from D to Y.
- The backdoor criterion is fairly powerful. Tells us:

Backdoor criterion

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

- Can we use a DAG to evaluate no unmeasured confounders?
- Holds if the **backdoor criterion** on a causal DAG is met:
  1. No vertex in $\mathbf{X}$ is a descend of D (**no post-treatment bias**), and
  2. $\mathbf{X}$ blocks all backdoor paths from D to Y.
- The backdoor criterion is fairly powerful. Tells us:
  - if there confounding given this DAG,

Backdoor criterion

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

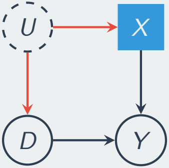
- Can we use a DAG to evaluate no unmeasured confounders?
- Holds if the **backdoor criterion** on a causal DAG is met:
  1. No vertex in $\mathbf{X}$ is a descend of D (**no post-treatment bias**), and
  2. $\mathbf{X}$ blocks all backdoor paths from D to Y.
- The backdoor criterion is fairly powerful. Tells us:
  - if there confounding given this DAG,
  - if it is possible to removing the confounding, and

Backdoor criterion

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

- Can we use a DAG to evaluate no unmeasured confounders?
- Holds if the **backdoor criterion** on a causal DAG is met:
  1. No vertex in $\mathbf{X}$ is a descend of D (**no post-treatment bias**), and
  2. $\mathbf{X}$ blocks all backdoor paths from D to Y.
- The backdoor criterion is fairly powerful. Tells us:
  - if there confounding given this DAG,
  - if it is possible to removing the confounding, and
  - what variables to condition on to eliminate the confounding.

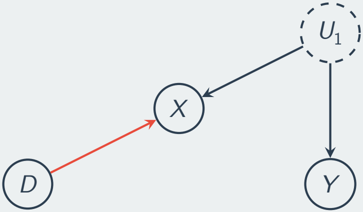
Other types of confounding



- D is enrolling in a job training program.
- Y is getting a job.
- U is being motivated
- X is number of job applications sent out.
- Big assumption here: no arrow from U to Y.

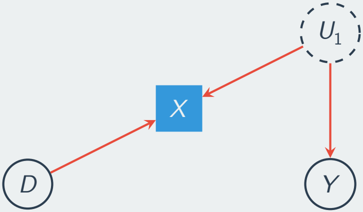
- Conditioning on X blocks all backdoor paths.

Why not condition on descendants?



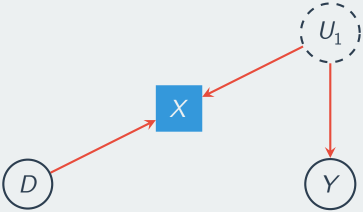
- No causal or statistical relationship between D and Y

Why not condition on descendants?



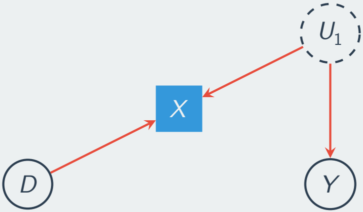
- No causal or statistical relationship between D and Y
- Conditioning on the posttreatment variables opens non-causal paths

Why not condition on descendants?



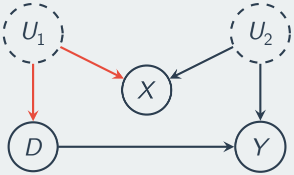
- No causal or statistical relationship between D and Y
- Conditioning on the posttreatment variables opens non-causal paths
  - $\rightsquigarrow$ statistical relationship between D and Y conditional on X

Why not condition on descendants?



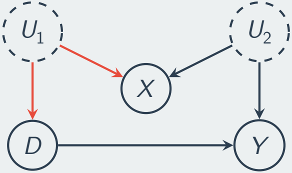
- No causal or statistical relationship between D and Y
- Conditioning on the posttreatment variables opens non-causal paths
  - $\rightsquigarrow$ statistical relationship between D and Y conditional on X
  - But still no causal relationship $\rightsquigarrow$ selection bias.

M-bias



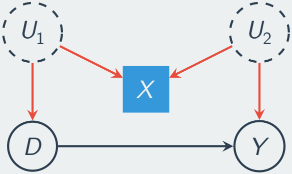
- Not all backdoor paths induce confounding.

M-bias



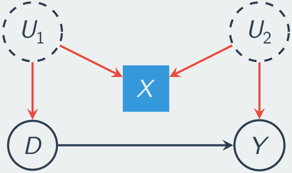
- Not all backdoor paths induce confounding.
- No conditioning: backdoor path blocked by the collider X_i.

M-bias



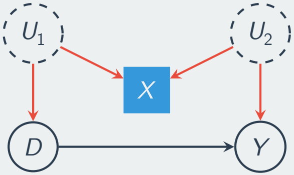
- Not all backdoor paths induce confounding.
- No conditioning: backdoor path blocked by the collider X_i.
- If we control for $X_i \rightsquigarrow$ opens the path and induces confounding.

M-bias



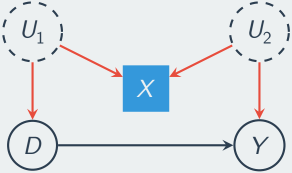
- Not all backdoor paths induce confounding.
- No conditioning: backdoor path blocked by the collider X_i.
- If we control for X_i $\rightsquigarrow$ opens the path and induces confounding.
    - Sometimes called **M-bias** or **collider bias**.

M-bias



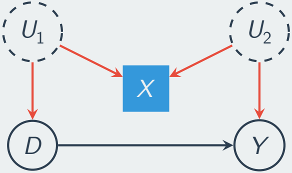
- Not all backdoor paths induce confounding.
- No conditioning: backdoor path blocked by the collider X_i.
- If we control for $X_i \rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called **M-bias** or **collider bias**.
- Controversial because of differing views on what to control for:

M-bias



- Not all backdoor paths induce confounding.
- No conditioning: backdoor path blocked by the collider X_i.
- If we control for $X_i \rightsquigarrow$ opens the path and induces confounding.
  - Sometimes called **M-bias** or **collider bias**.
- Controversial because of differing views on what to control for:
  - Rubin thinks that M-bias is a “mathematical curiosity” and we should control for all pretreatment variables

M-bias



- Not all backdoor paths induce confounding.
- No conditioning: backdoor path blocked by the collider X_i.
- If we control for $X_i \rightsquigarrow$ opens the path and induces confounding.
    - Sometimes called **M-bias** or **collider bias**.
- Controversial because of differing views on what to control for:
    - Rubin thinks that M-bias is a “mathematical curiosity” and we should control for all pretreatment variables
    - Pearl and others think M-bias is a real threat.

4/ Sensitivity analysis

Where are we? Where are we going?

- Saw how to estimate the ATE with regression under selection on observables.

Where are we? Where are we going?

- Saw how to estimate the ATE with regression under selection on observables.
- What if this assumption doesn't hold? Two potential solutions:

Where are we? Where are we going?

- Saw how to estimate the ATE with regression under selection on observables.
- What if this assumption doesn't hold? Two potential solutions:
    1. **Sensitivity analysis**: try to vary the amount of unmeasured confounding to see if it changes the effect.

Where are we? Where are we going?

- Saw how to estimate the ATE with regression under selection on observables.
- What if this assumption doesn't hold? Two potential solutions:
    1. **Sensitivity analysis**: try to vary the amount of unmeasured confounding to see if it changes the effect.
    2. **Partial identification**: abandon point identification and try to find bounds for the ATE under different assumptions.

Sensitivity analysis for regression

- Standard regression estimator of the ATE:

$$Y_i = \hat{\alpha} + \hat{\tau} D_i + \mathbf{X}_i' \hat{\beta} + \hat{\varepsilon}_i$$

Sensitivity analysis for regression

- Standard regression estimator of the ATE:

$$Y_i = \hat{\alpha} + \widehat{\tau} D_i + \mathbf{X}_i' \widehat{\beta} + \hat{\varepsilon}_i$$

- What if the true regression model contained U_i which we omitted?

$$Y_i = \alpha + \tau D_i + \mathbf{X}_i' \beta + \gamma U_i + \varepsilon_i \qquad \widehat{\tau} \overset{p}{\to} \tau + \gamma \times \underbrace{\frac{\text{cov}(D_i^{\perp \mathbf{X}}, U_i^{\perp \mathbf{X}})}{\mathbb{V}(D_i^{\perp \mathbf{X}})}}_{\text{regression of } U_i^{\perp \mathbf{X}} \text{ on } D_i^{\perp \mathbf{X}}}$$

Sensitivity analysis for regression

- Standard regression estimator of the ATE:

$$Y_i = \hat{\alpha} + \widehat{\tau} D_i + \mathbf{X}_i' \widehat{\beta} + \hat{\varepsilon}_i$$

- What if the true regression model contained U_i which we omitted?

$$Y_i = \alpha + \tau D_i + \mathbf{X}_i' \beta + \gamma U_i + \varepsilon_i \qquad \widehat{\tau} \xrightarrow{p} \tau + \gamma \times \underbrace{\frac{\text{cov}(D_i^{\perp \mathbf{X}}, U_i^{\perp \mathbf{X}})}{\mathbb{V}(D_i^{\perp \mathbf{X}})}}_{\text{regression of } U_i^{\perp \mathbf{X}} \text{ on } D_i^{\perp \mathbf{X}}}$$

- Standard **omitted variable bias formula**:

$$\widehat{\tau} \xrightarrow{p} \tau + \gamma \times \underbrace{\frac{\text{cov}(D_i^{\perp \mathbf{X}}, U_i^{\perp \mathbf{X}})}{\mathbb{V}(D_i^{\perp \mathbf{X}})}}_{\text{regression of } U_i^{\perp \mathbf{X}} \text{ on } D_i^{\perp \mathbf{X}}}$$

Partial R-squared interpretations

- Regression coefficients with unknowns are difficult to reason about.

Partial R-squared interpretations

- Regression coefficients with unknowns are difficult to reason about.
- Easier to reason with partial R^2 version of OVB (Cinelli and Hazlett, JRSSB, 2019):

$$|\text{bias}| = \sqrt{\frac{R^2_{Y\sim U|D,\mathbf{X}} R^2_{D\sim U|\mathbf{X}}}{1 - R^2_{D\sim U|\mathbf{X}}} \frac{\mathbb{V}(Y^{\perp \mathbf{X},D})}{\mathbb{V}(D_i^{\perp \mathbf{X}})}}$$

Partial R-squared interpretations

- Regression coefficients with unknowns are difficult to reason about.
- Easier to reason with partial R^2 version of OVB (Cinelli and Hazlett, JRSSB, 2019):

$$|\text{bias}| = \sqrt{\frac{R^2_{Y\sim U|D,\mathbf{X}} R^2_{D\sim U|\mathbf{X}}}{1 - R^2_{D\sim U|\mathbf{X}}} \frac{\mathbb{V}(Y^{\perp \mathbf{X},D})}{\mathbb{V}(D_i^{\perp \mathbf{X}})}}$$

- Partial R^2 is the incremental predictive value of one variable:

$$R^2_{Y\sim U|D,\mathbf{X}} = \frac{R^2_{Y\sim D+\mathbf{X}+U} - R^2_{Y\sim D+\mathbf{X}}}{1 - R^2_{Y\sim D+\mathbf{X}}} = \frac{\text{additional variance explained by } U}{\text{variance unexplained by } D, \mathbf{X}}$$

Partial R-squared interpretations

- Regression coefficients with unknowns are difficult to reason about.
- Easier to reason with partial R^2 version of OVB (Cinelli and Hazlett, JRSSB, 2019):

$$|\text{bias}| = \sqrt{\frac{R^2_{Y\sim U|D,\mathbf{X}} R^2_{D\sim U|\mathbf{X}}}{1 - R^2_{D\sim U|\mathbf{X}}} \frac{\mathbb{V}(Y^{\perp \mathbf{X},D})}{\mathbb{V}(D_i^{\perp \mathbf{X}})}}$$

- Partial R^2 is the incremental predictive value of one variable:

$$R^2_{Y\sim U|D,\mathbf{X}} = \frac{R^2_{Y\sim D+\mathbf{X}+U} - R^2_{Y\sim D+\mathbf{X}}}{1 - R^2_{Y\sim D+\mathbf{X}}} = \frac{\text{additional variance explained by } U}{\text{variance unexplained by } D, \mathbf{X}}$$

- **Sensitivity analysis** can then vary two unknown parameters:

Partial R-squared interpretations

- Regression coefficients with unknowns are difficult to reason about.
- Easier to reason with partial R^2 version of OVB (Cinelli and Hazlett, JRSSB, 2019):

$$|\text{bias}| = \sqrt{\frac{R^2_{Y\sim U|D,\mathbf{X}} R^2_{D\sim U|\mathbf{X}}}{1 - R^2_{D\sim U|\mathbf{X}}} \frac{\mathbb{V}(Y^{\perp \mathbf{X},D})}{\mathbb{V}(D_i^{\perp \mathbf{X}})}}$$

- Partial R^2 is the incremental predictive value of one variable:

$$R^2_{Y\sim U|D,\mathbf{X}} = \frac{R^2_{Y\sim D+\mathbf{X}+U} - R^2_{Y\sim D+\mathbf{X}}}{1 - R^2_{Y\sim D+\mathbf{X}}} = \frac{\text{additional variance explained by } U}{\text{variance unexplained by } D, \mathbf{X}}$$

- **Sensitivity analysis** can then vary two unknown parameters:
    - $R^2_{Y\sim U|D,\mathbf{X}} \in [0, 1]$ incremental predictive value of U for the outcome

Partial R-squared interpretations

- Regression coefficients with unknowns are difficult to reason about.
- Easier to reason with partial R^2 version of OVB (Cinelli and Hazlett, JRSSB, 2019):

$$|\text{bias}| = \sqrt{\frac{R^2_{Y\sim U|D,\mathbf{X}} R^2_{D\sim U|\mathbf{X}}}{1 - R^2_{D\sim U|\mathbf{X}}} \frac{\mathbb{V}(Y^{\perp \mathbf{X},D})}{\mathbb{V}(D_i^{\perp \mathbf{X}})}}$$

- Partial R^2 is the incremental predictive value of one variable:

$$R^2_{Y\sim U|D,\mathbf{X}} = \frac{R^2_{Y\sim D+\mathbf{X}+U} - R^2_{Y\sim D+\mathbf{X}}}{1 - R^2_{Y\sim D+\mathbf{X}}} = \frac{\text{additional variance explained by } U}{\text{variance unexplained by } D, \mathbf{X}}$$

- **Sensitivity analysis** can then vary two unknown parameters:
    - $R^2_{Y\sim U|D,\mathbf{X}} \in [0, 1]$ incremental predictive value of U for the outcome
    - $R^2_{D\sim U|\mathbf{X}} \in [0, 1]$ incremental predictive value of U for treatment

Partial R-squared interpretations

- Regression coefficients with unknowns are difficult to reason about.
- Easier to reason with partial R^2 version of OVB (Cinelli and Hazlett, JRSSB, 2019):

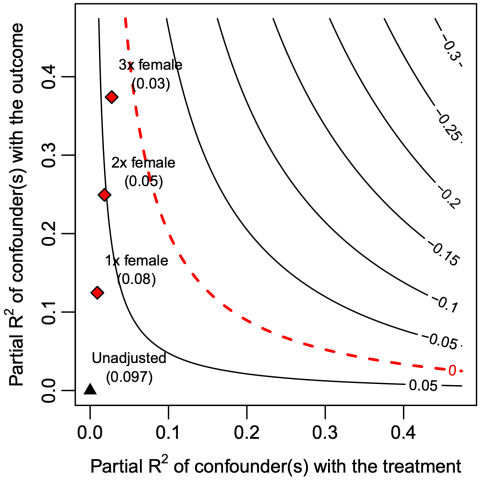
$$|\text{bias}| = \sqrt{\frac{R^2_{Y\sim U|D,\mathbf{X}} R^2_{D\sim U|\mathbf{X}}}{1 - R^2_{D\sim U|\mathbf{X}}} \frac{\mathbb{V}(Y^{\perp \mathbf{X},D})}{\mathbb{V}(D_i^{\perp \mathbf{X}})}}$$

- Partial R^2 is the incremental predictive value of one variable:

$$R^2_{Y\sim U|D,\mathbf{X}} = \frac{R^2_{Y\sim D+\mathbf{X}+U} - R^2_{Y\sim D+\mathbf{X}}}{1 - R^2_{Y\sim D+\mathbf{X}}} = \frac{\text{additional variance explained by } U}{\text{variance unexplained by } D, \mathbf{X}}$$

- **Sensitivity analysis** can then vary two unknown parameters:
    - $R^2_{Y\sim U|D,\mathbf{X}} \in [0,1]$ incremental predictive value of U for the outcome
    - $R^2_{D\sim U|\mathbf{X}} \in [0,1]$ incremental predictive value of U for treatment
    - From these we can determine the bias and thus the true value of τ

Sensitivity analysis example



5/ Partial identification and bounds

No assumption bounds

- **Law of decreasing credibility** (Manski): credibility of inferences decreases with strength of assumptions

No assumption bounds

- **Law of decreasing credibility** (Manski): credibility of inferences decreases with strength of assumptions
  - Idea: pick assumptions and then figure out what you can learn.

No assumption bounds

- **Law of decreasing credibility** (Manski): credibility of inferences decreases with strength of assumptions
    - Idea: pick assumptions and then figure out what you can learn.
    - May not be point identified, but maybe we can bound the effect.

No assumption bounds

- **Law of decreasing credibility** (Manski): credibility of inferences decreases with strength of assumptions
    - Idea: pick assumptions and then figure out what you can learn.
    - May not be point identified, but maybe we can bound the effect.
- If Y is bounded $[y_L, y_U]$, τ logically must be in $[y_L - y_U, y_U - y_L]$.

No assumption bounds

- **Law of decreasing credibility** (Manski): credibility of inferences decreases with strength of assumptions
    - Idea: pick assumptions and then figure out what you can learn.
    - May not be point identified, but maybe we can bound the effect.
- If Y is bounded $[y_L, y_U]$, τ logically must be in $[y_L - y_U, y_U - y_L]$.
- Can we improve using data? Rewrite the ATE with $p = \mathbb{P}(D_i = 1)$:

$$\tau = \mathbb{E}[Y_i \mid D_i = 1]p + \mathbb{E}[Y_i(1) \mid D_i = 0](1 - p)$$
$$- \mathbb{E}[Y_i(0) \mid D_i = 1]p - \mathbb{E}[Y_i \mid D_i = 0](1 - p)$$

No assumption bounds

- **Law of decreasing credibility** (Manski): credibility of inferences decreases with strength of assumptions
  - Idea: pick assumptions and then figure out what you can learn.
  - May not be point identified, but maybe we can bound the effect.
- If Y is bounded $[y_L, y_U]$, τ logically must be in $[y_L - y_U, y_U - y_L]$.
- Can we improve using data? Rewrite the ATE with $p = \mathbb{P}(D_i = 1)$:

$$\begin{aligned}\tau = \; & \mathbb{E}[Y_i \mid D_i = 1]p + \mathbb{E}[Y_i(1) \mid D_i = 0](1-p) \\ & - \mathbb{E}[Y_i(0) \mid D_i = 1]p - \mathbb{E}[Y_i \mid D_i = 0](1-p)\end{aligned}$$

- Plug in y_L and y_U for the counterfactual means to get bounds for τ:

$$\tau \geq \mathbb{E}[Y_i \mid D_i = 1]p + y_L(1-p) - y_U p - \mathbb{E}[Y_i \mid D_i = 0](1-p)$$
$$\tau \leq \mathbb{E}[Y_i \mid D_i = 1]p + y_U(1-p) - y_L p - \mathbb{E}[Y_i \mid D_i = 0](1-p)$$

No assumption bounds

- **Law of decreasing credibility** (Manski): credibility of inferences decreases with strength of assumptions
  - Idea: pick assumptions and then figure out what you can learn.
  - May not be point identified, but maybe we can bound the effect.
- If Y is bounded $[y_L, y_U]$, τ logically must be in $[y_L - y_U, y_U - y_L]$.
- Can we improve using data? Rewrite the ATE with $p = \mathbb{P}(D_i = 1)$:

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

- Plug in y_L and y_U for the counterfactual means to get bounds for τ:

$$\tau \geq \mathbb{E}[Y_i \mid D_i = 1]p + y_L(1-p) - y_U p - \mathbb{E}[Y_i \mid D_i = 0](1-p)$$
$$\tau \leq \mathbb{E}[Y_i \mid D_i = 1]p + y_U(1-p) - y_L p - \mathbb{E}[Y_i \mid D_i = 0](1-p)$$

  - These bounds have width of $|y_U - y_L|$ which is half of the logical bounds.

No assumption bounds

- **Law of decreasing credibility** (Manski): credibility of inferences decreases with strength of assumptions
  - Idea: pick assumptions and then figure out what you can learn.
  - May not be point identified, but maybe we can bound the effect.
- If Y is bounded $[y_L, y_U]$, τ logically must be in $[y_L - y_U, y_U - y_L]$.
- Can we improve using data? Rewrite the ATE with $p = \mathbb{P}(D_i = 1)$:

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

- Plug in y_L and y_U for the counterfactual means to get bounds for τ:

$$\tau \geq \mathbb{E}[Y_i \mid D_i = 1]p + y_L(1-p) - y_U p - \mathbb{E}[Y_i \mid D_i = 0](1-p)$$
$$\tau \leq \mathbb{E}[Y_i \mid D_i = 1]p + y_U(1-p) - y_L p - \mathbb{E}[Y_i \mid D_i = 0](1-p)$$

  - These bounds have width of $|y_U - y_L|$ which is half of the logical bounds.
  - But always will contain 0. Weak assumptions $\rightsquigarrow$ weak inferences

Optimized treatment choice

- **Assumptions** can narrow the bounds even further.

Optimized treatment choice

- **Assumptions** can narrow the bounds even further.
- Assumption: people choose the treatment with the highest outcome.

Optimized treatment choice

- **Assumptions** can narrow the bounds even further.
- Assumption: people choose the treatment with the highest outcome.
  - $\mathbb{E}[Y_i(0) \mid D_i = 1] \leq \mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}[Y_i \mid D_i = 1] = \mu(1)$

Optimized treatment choice

- **Assumptions** can narrow the bounds even further.
- Assumption: people choose the treatment with the highest outcome.
  - $\mathbb{E}[Y_i(0) \mid D_i = 1] \leq \mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}[Y_i \mid D_i = 1] = \mu(1)$
  - $\mathbb{E}[Y_i(1) \mid D_i = 0] \leq \mathbb{E}[Y_i(0) \mid D_i = 0] = \mathbb{E}[Y_i \mid D_i = 0] = \mu(0)$

Optimized treatment choice

- **Assumptions** can narrow the bounds even further.
- Assumption: people choose the treatment with the highest outcome.
  - $\mathbb{E}[Y_i(0) \mid D_i = 1] \leq \mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}[Y_i \mid D_i = 1] = \mu(1)$
  - $\mathbb{E}[Y_i(1) \mid D_i = 0] \leq \mathbb{E}[Y_i(0) \mid D_i = 0] = \mathbb{E}[Y_i \mid D_i = 0] = \mu(0)$
- Implies the following bounds for τ:

$$\tau \in \left[(1-p)\left(y_L - \mathbb{E}[Y_i \mid D_i = 0]\right),\ \ p\left(\mathbb{E}[Y_i \mid D_i = 1] - y_L\right)\right]$$

Optimized treatment choice

- **Assumptions** can narrow the bounds even further.
- Assumption: people choose the treatment with the highest outcome.
  - $\mathbb{E}[Y_i(0) \mid D_i = 1] \leq \mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}[Y_i \mid D_i = 1] = \mu(1)$
  - $\mathbb{E}[Y_i(1) \mid D_i = 0] \leq \mathbb{E}[Y_i(0) \mid D_i = 0] = \mathbb{E}[Y_i \mid D_i = 0] = \mu(0)$
- Implies the following bounds for τ:

$$\tau \in \left[(1-p)\left(y_L - \mathbb{E}[Y_i \mid D_i = 0]\right),\;\; p\left(\mathbb{E}[Y_i \mid D_i = 1] - y_L\right)\right]$$

- Width of these bounds: $\mathbb{E}[Y_i] - y_L$

Optimized treatment choice

- **Assumptions** can narrow the bounds even further.
- Assumption: people choose the treatment with the highest outcome.
  - $\mathbb{E}[Y_i(0) \mid D_i = 1] \leq \mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}[Y_i \mid D_i = 1] = \mu(1)$
  - $\mathbb{E}[Y_i(1) \mid D_i = 0] \leq \mathbb{E}[Y_i(0) \mid D_i = 0] = \mathbb{E}[Y_i \mid D_i = 0] = \mu(0)$
- Implies the following bounds for τ:

$$\tau \in \left[(1-p)\left(y_L - \mathbb{E}[Y_i \mid D_i = 0]\right),\ \ p\left(\mathbb{E}[Y_i \mid D_i = 1] - y_L\right)\right]$$

- Width of these bounds: $\mathbb{E}[Y_i] - y_L$
  - Width now depends on the observed data!

Optimized treatment choice

- **Assumptions** can narrow the bounds even further.
- Assumption: people choose the treatment with the highest outcome.
  - $\mathbb{E}[Y_i(0) \mid D_i = 1] \leq \mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}[Y_i \mid D_i = 1] = \mu(1)$
  - $\mathbb{E}[Y_i(1) \mid D_i = 0] \leq \mathbb{E}[Y_i(0) \mid D_i = 0] = \mathbb{E}[Y_i \mid D_i = 0] = \mu(0)$
- Implies the following bounds for τ:

$$\tau \in \left[(1-p)\left(y_L - \mathbb{E}[Y_i \mid D_i = 0]\right),\ \ p\left(\mathbb{E}[Y_i \mid D_i = 1] - y_L\right)\right]$$

- Width of these bounds: $\mathbb{E}[Y_i] - y_L$
  - Width now depends on the observed data!
  - Interval will still always include zero.

Confidence regions for bounds

- More general setup:

Confidence regions for bounds

- More general setup:
    - True bounds $[\delta_L, \delta_U]$ also called the **identification region**

Confidence regions for bounds

- More general setup:
  - True bounds $[\delta_L, \delta_U]$ also called the **identification region**
  - Estimated bounds $[\hat{\delta}_L, \hat{\delta}_U]$.

Confidence regions for bounds

- More general setup:
  - True bounds $[\delta_L, \delta_U]$ also called the **identification region**
  - Estimated bounds $[\hat{\delta}_L, \hat{\delta}_U]$.
  - $\widehat{\text{se}}(\hat{\delta}_L), \widehat{\text{se}}(\hat{\delta}_U)$ are the standard errors of the estimated bounds

Confidence regions for bounds

- More general setup:
    - True bounds $[\delta_L, \delta_U]$ also called the **identification region**
    - Estimated bounds $[\hat{\delta}_L, \hat{\delta}_U]$.
    - $\widehat{\text{se}}(\hat{\delta}_L), \widehat{\text{se}}(\hat{\delta}_U)$ are the standard errors of the estimated bounds
- Two possible CI approaches that find intervals that...

Confidence regions for bounds

- More general setup:
  - True bounds $[\delta_L, \delta_U]$ also called the **identification region**
  - Estimated bounds $[\hat{\delta}_L, \hat{\delta}_U]$.
  - $\widehat{\text{se}}(\hat{\delta}_L), \widehat{\text{se}}(\hat{\delta}_U)$ are the standard errors of the estimated bounds
- Two possible CI approaches that find intervals that...

  1. Covers the identified region with probability $1 - \alpha$

$$\mathbb{P}(\hat{\delta}_L \leq \delta_L, \hat{\delta}_U \geq \delta_U) \geq 1 - \alpha$$

Confidence regions for bounds

- More general setup:
  - True bounds $[\delta_L, \delta_U]$ also called the **identification region**
  - Estimated bounds $[\hat{\delta}_L, \hat{\delta}_U]$.
  - $\widehat{\text{se}}(\hat{\delta}_L), \widehat{\text{se}}(\hat{\delta}_U)$ are the standard errors of the estimated bounds
- Two possible CI approaches that find intervals that...

1. Covers the identified region with probability $1 - \alpha$

$$\mathbb{P}(\hat{\delta}_L \leq \delta_L, \hat{\delta}_U \geq \delta_U) \geq 1 - \alpha$$

2. Covers the true value of the parameter with probability $1 - \alpha$

$$\mathbb{P}(\tau \in [\hat{\delta}_L, \hat{\delta}_U]) \geq 1 - \alpha$$

Calculating confidence intervals

- Case 1: covering the identified region $\mathbb{P}(\hat{\delta}_L \leq \delta_L, \hat{\delta}_U \geq \delta_U) \geq 1 - \alpha$

$$[\hat{\delta}_L - z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_L), \quad \hat{\delta}_U + z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_U)]$$

Calculating confidence intervals

- Case 1: covering the identified region $\mathbb{P}(\hat{\delta}_L \leq \delta_L, \hat{\delta}_U \geq \delta_U) \geq 1 - \alpha$

$$[\hat{\delta}_L - z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_L), \quad \hat{\delta}_U + z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_U)]$$

- Works because of **Bonferroni inequality**:

$$\mathbb{P}(\hat{\delta}_L \leq \delta_L \text{ and } \hat{\delta}_U \geq \delta_U) \geq \mathbb{P}(\hat{\delta}_L \leq \delta_L) + \mathbb{P}(\hat{\delta}_U \leq \delta_U) - 1 = 1 - \alpha$$

Calculating confidence intervals

- Case 1: covering the identified region $\mathbb{P}(\hat{\delta}_L \leq \delta_L, \hat{\delta}_U \geq \delta_U) \geq 1 - \alpha$

$$[\hat{\delta}_L - z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_L), \quad \hat{\delta}_U + z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_U)]$$

- Works because of **Bonferroni inequality**:

$$\mathbb{P}(\hat{\delta}_L \leq \delta_L \text{ and } \hat{\delta}_U \geq \delta_U) \geq \mathbb{P}(\hat{\delta}_L \leq \delta_L) + \mathbb{P}(\hat{\delta}_U \leq \delta_U) - 1 = 1 - \alpha$$

- Case 2: cover the true parameter, τ.

$$[\hat{\delta}_L - z_{1-\alpha}\widehat{\text{se}}(\hat{\delta}_L), \quad \hat{\delta}_U + z_{1-\alpha}\widehat{\text{se}}(\hat{\delta}_U)]$$

Calculating confidence intervals

- Case 1: covering the identified region $\mathbb{P}(\hat{\delta}_L \leq \delta_L, \hat{\delta}_U \geq \delta_U) \geq 1 - \alpha$

$$[\hat{\delta}_L - z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_L), \quad \hat{\delta}_U + z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_U)]$$

- Works because of **Bonferroni inequality**:

$$\mathbb{P}(\hat{\delta}_L \leq \delta_L \text{ and } \hat{\delta}_U \geq \delta_U) \geq \mathbb{P}(\hat{\delta}_L \leq \delta_L) + \mathbb{P}(\hat{\delta}_U \leq \delta_U) - 1 = 1 - \alpha$$

- Case 2: cover the true parameter, τ.

$$[\hat{\delta}_L - z_{1-\alpha}\widehat{\text{se}}(\hat{\delta}_L), \quad \hat{\delta}_U + z_{1-\alpha}\widehat{\text{se}}(\hat{\delta}_U)]$$

  - If $\tau = \delta_L$ or $\tau = \delta_U$, then coverage converges to $1 - \alpha$

Calculating confidence intervals

- Case 1: covering the identified region $\mathbb{P}(\hat{\delta}_L \leq \delta_L, \hat{\delta}_U \geq \delta_U) \geq 1 - \alpha$

$$[\hat{\delta}_L - z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_L), \quad \hat{\delta}_U + z_{1-\alpha/2}\widehat{\text{se}}(\hat{\delta}_U)]$$

- Works because of **Bonferroni inequality**:

$$\mathbb{P}(\hat{\delta}_L \leq \delta_L \text{ and } \hat{\delta}_U \geq \delta_U) \geq \mathbb{P}(\hat{\delta}_L \leq \delta_L) + \mathbb{P}(\hat{\delta}_U \leq \delta_U) - 1 = 1 - \alpha$$

- Case 2: cover the true parameter, τ.

$$[\hat{\delta}_L - z_{1-\alpha}\widehat{\text{se}}(\hat{\delta}_L), \quad \hat{\delta}_U + z_{1-\alpha}\widehat{\text{se}}(\hat{\delta}_U)]$$

  - If $\tau = \delta_L$ or $\tau = \delta_U$, then coverage converges to $1 - \alpha$
  - If $\delta_L < \tau < \delta_U$, then coverage converges to 1.