FLS 6441 - Methods III: Explanation and Causation

Week 9 - Controlling for Confounding

Jonathan Phillips

May 2019

Classification of Research Designs

		Independence of Treatment Assignment	Researcher Controls Treatment Assignment?
Controlled Experiments	Field Experiments	✓	✓
	Survey and Lab Experiments	✓	✓
Natural Experiments	Natural Experiments	✓	
	Instrumental Variables	✓	
	Discontinuities	✓	
Observational Studies	Difference-in-Differences		
	Controlling for Confounding		
	Matching		
	Comparative Cases and Process Tracing		

Section 1

Controlling for Confounding

Controlling for Confounding

- Adding control variables changes the comparison we are making: **Treatment is associated with higher values of the Outcome...holding constant the values of X**

Controlling for Confounding

- Adding control variables changes the comparison we are making: **Treatment is associated with higher values of the Outcome...holding constant the values of X**
- But when does controlling allow us to say: **Treatment *causes* higher values of the Outcome**?

Controlling for Confounding

- Up until now causal estimates have required treatment assignment to be **independent of potential outcomes**

Controlling for Confounding

- Up until now causal estimates have required treatment assignment to be **independent of potential outcomes**

$$(y_0, y_1) \perp D$$

Controlling for Confounding

- Up until now causal estimates have required treatment assignment to be **independent of potential outcomes**

$$(y_0, y_1) \perp D$$

- But it's also acceptable if treatment assignment is **conditionally independent of potential outcomes**

Controlling for Confounding

- Up until now causal estimates have required treatment assignment to be **independent of potential outcomes**

$$(y_0, y_1) \perp D$$

- But it's also acceptable if treatment assignment is **conditionally independent of potential outcomes**

$$(y_0, y_1) \perp D|X$$

Controlling for Confounding

- Up until now causal estimates have required treatment assignment to be **independent of potential outcomes**

$$(y_0, y_1) \perp D$$

- But it's also acceptable if treatment assignment is **conditionally independent of potential outcomes**

$$(y_0, y_1) \perp D | X$$

- After controlling for X, treatment is independent of potential outcomes: 'No unmeasured confounders'

Controlling for Confounding

- Up until now causal estimates have required treatment assignment to be **independent of potential outcomes**

$$(y_0, y_1) \perp D$$

- But it's also acceptable if treatment assignment is **conditionally independent of potential outcomes**

$$(y_0, y_1) \perp D|X$$

- After controlling for X, treatment is independent of potential outcomes: 'No unmeasured confounders'
- This is an *assumption*

Controlling for Confounding

- Up until now causal estimates have required treatment assignment to be **independent of potential outcomes**

$$(y_0, y_1) \perp D$$

- But it's also acceptable if treatment assignment is **conditionally independent of potential outcomes**

$$(y_0, y_1) \perp D|X$$

- After controlling for X, treatment is independent of potential outcomes: 'No unmeasured confounders'
- This is an *assumption*
  - We cannot directly test it

Controlling for Confounding

- Up until now causal estimates have required treatment assignment to be **independent of potential outcomes**

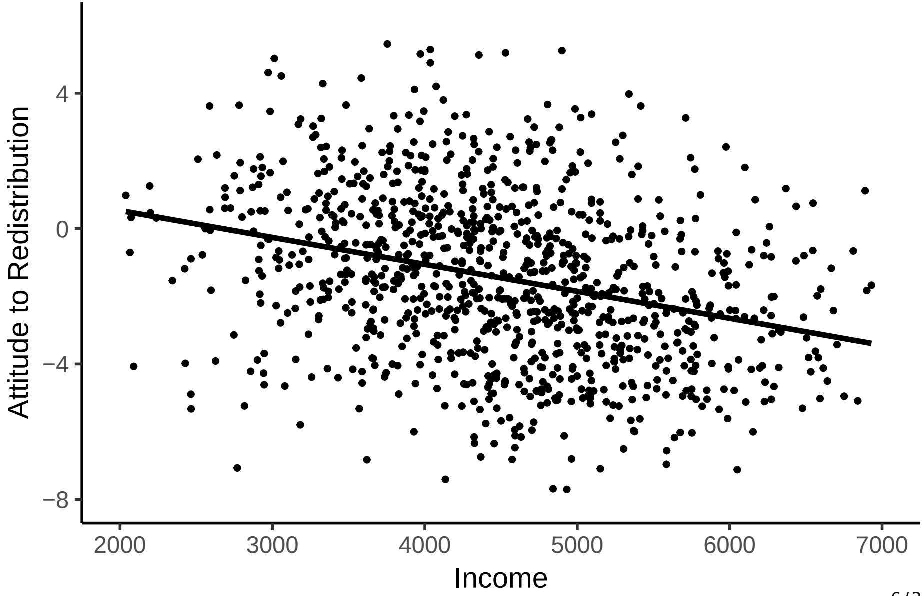
$$(y_0, y_1) \perp D$$

- But it's also acceptable if treatment assignment is **conditionally independent of potential outcomes**

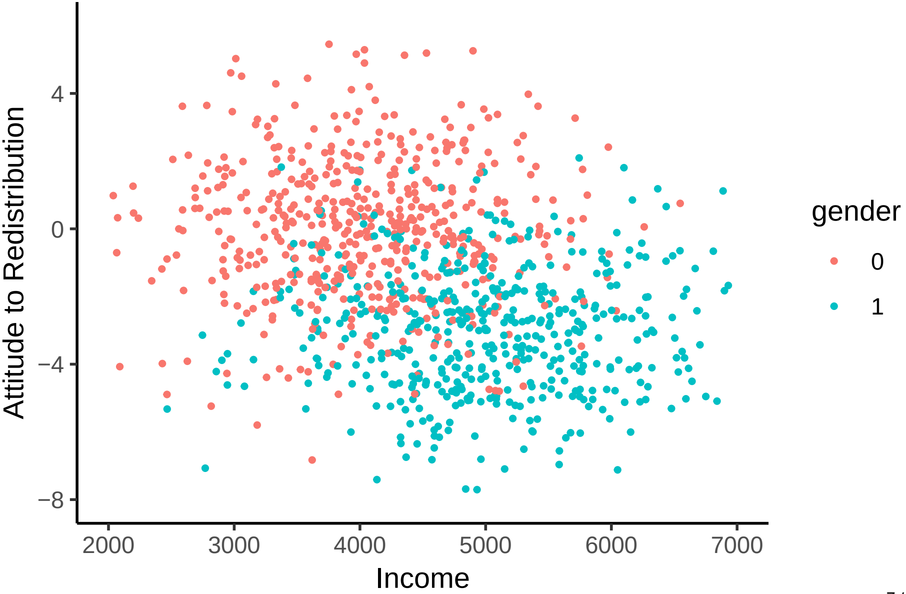
$$(y_0, y_1) \perp D | X$$

- After controlling for X, treatment is independent of potential outcomes: 'No unmeasured confounders'
- This is an *assumption*
  - We cannot directly test it
  - We have to make an argument and provide supporting evidence

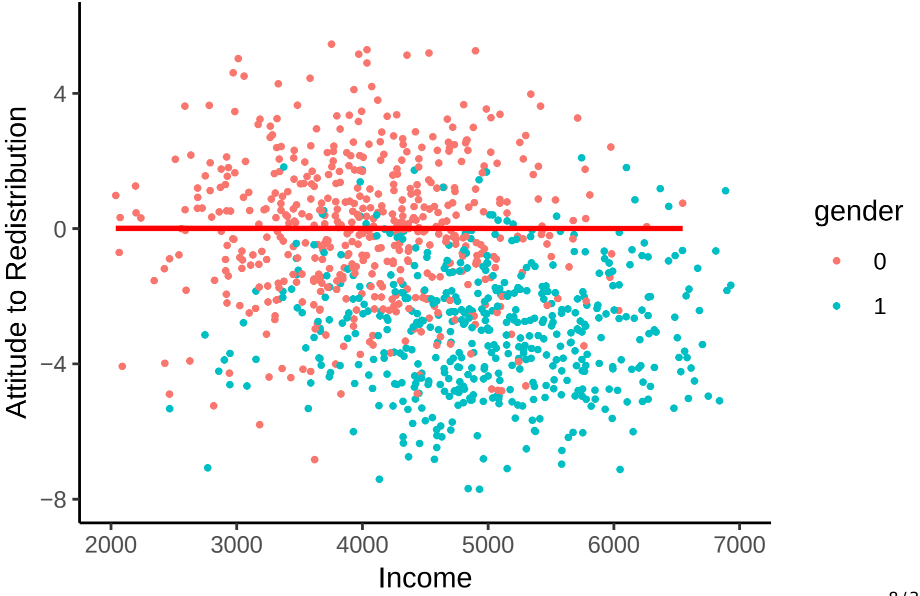
Controlling for Confounding



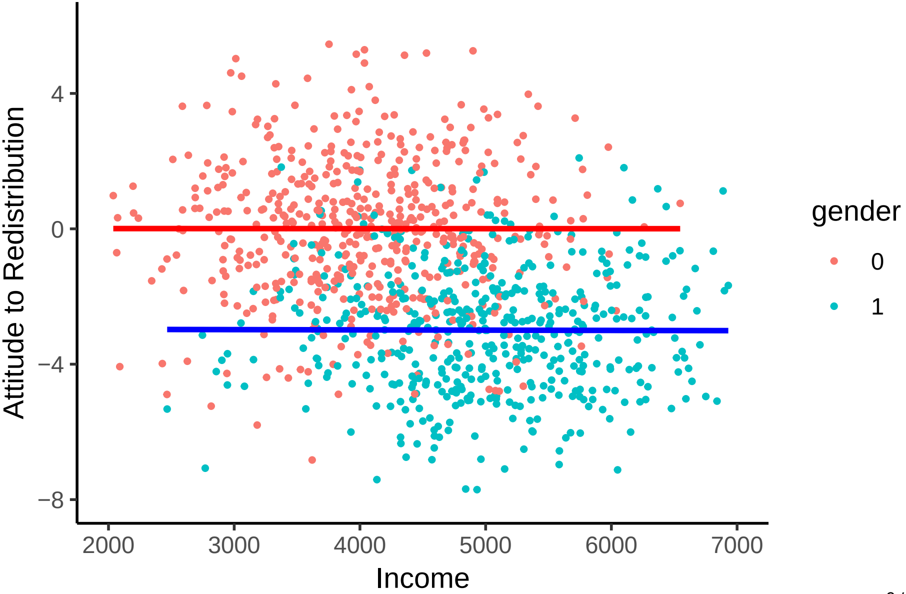
Controlling for Confounding



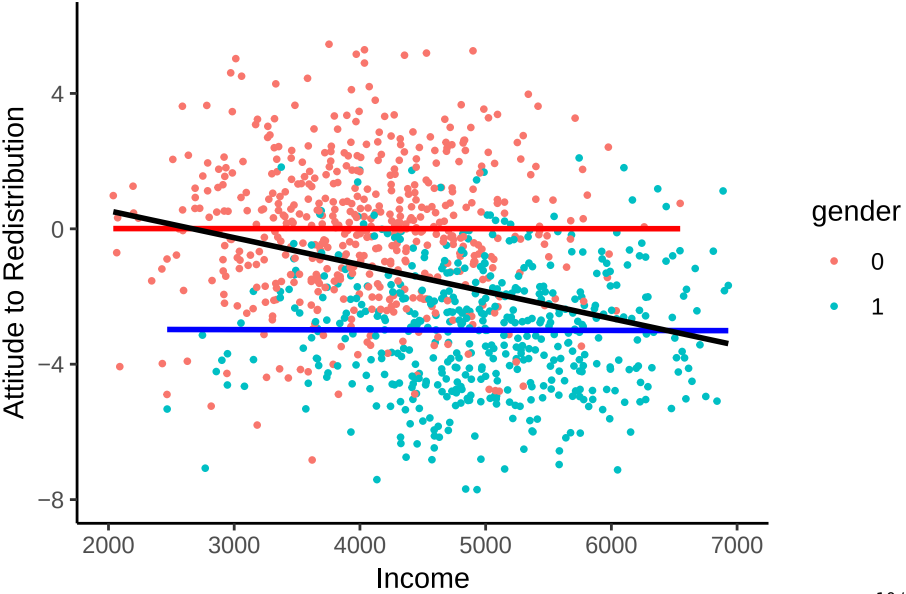
Controlling for Confounding



Controlling for Confounding



Controlling for Confounding



Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:

Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:
- ’True’ regression with all confounders:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:
- ’True’ regression with all confounders:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

- The ’wrong’ regression with a missing confounder:

Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:
- 'True' regression with all confounders:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

- The 'wrong' regression with a missing confounder:

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:
- 'True' regression with all confounders:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

- The 'wrong' regression with a missing confounder:

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

- What happens to our coefficient estimate?

Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:
- 'True' regression with all confounders:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

- The 'wrong' regression with a missing confounder:

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

- What happens to our coefficient estimate?

$$X_i = \psi + \delta D_i + \epsilon_i$$

Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:
- 'True' regression with all confounders:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

- The 'wrong' regression with a missing confounder:

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

- What happens to our coefficient estimate?

$$X_i = \psi + \delta D_i + \epsilon_i$$

$$Y_i = \alpha + \beta D_i + \gamma(\psi + \delta D_i + \epsilon_i) + \epsilon_i$$

Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:
- 'True' regression with all confounders:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

- The 'wrong' regression with a missing confounder:

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

- What happens to our coefficient estimate?

$$X_i = \psi + \delta D_i + \epsilon_i$$

$$Y_i = \alpha + \beta D_i + \gamma(\psi + \delta D_i + \epsilon_i) + \epsilon_i$$

$$Y_i = \alpha + D_i(\beta + \gamma\delta) + \gamma(\psi + \epsilon_i) + \epsilon_i$$

Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:
- 'True' regression with all confounders:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

- The 'wrong' regression with a missing confounder:

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

- What happens to our coefficient estimate?

$$X_i = \psi + \delta D_i + \epsilon_i$$

$$Y_i = \alpha + \beta D_i + \gamma(\psi + \delta D_i + \epsilon_i) + \epsilon_i$$

$$Y_i = \alpha + D_i(\beta + \gamma\delta) + \gamma(\psi + \epsilon_i) + \epsilon_i$$

- So the coefficient we estimate is wrong by this amount:

Controlling for Confounding

- If we omit a confounding variable, we bias our regression estimate:
- 'True' regression with all confounders:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$$

- The 'wrong' regression with a missing confounder:

$$Y_i = \alpha + \beta D_i + \epsilon_i$$

- What happens to our coefficient estimate?

$$X_i = \psi + \delta D_i + \epsilon_i$$

$$Y_i = \alpha + \beta D_i + \gamma(\psi + \delta D_i + \epsilon_i) + \epsilon_i$$

$$Y_i = \alpha + D_i(\beta + \gamma\delta) + \gamma(\psi + \epsilon_i) + \epsilon_i$$

- So the coefficient we estimate is wrong by this amount:

$$\beta_{wrong} = \beta_{true} + \gamma\delta$$

Controlling for Confounding

- What does controlling **do**?

Controlling for Confounding

- What does controlling **do**?
  - It means **removing the variation** in the data due to the confounder

Controlling for Confounding

- What does controlling **do**?
  - It means **removing the variation** in the data due to the confounder
  - Equivalently, it means separating our data for each value of the confounder: **Subclassification**

Controlling for Confounding

- What does controlling **do**?
  - It means **removing the variation** in the data due to the confounder
  - Equivalently, it means separating our data for each value of the confounder: **Subclassification**
  - Then, within each group, the confounder is **constant** and can't affect the relationship between D and Y.

Controlling for Confounding

- What does controlling **do**?
    - It means **removing the variation** in the data due to the confounder
    - Equivalently, it means separating our data for each value of the confounder: **Subclassification**
    - Then, within each group, the confounder is **constant** and can't affect the relationship between D and Y.
    - We have **created balance** between the treated and control groups on all other characteristics

Controlling for Confounding

- We receive a dataset with twenty variables: D, Y and 18 more.
  - Which variables should we include as controls?

Controlling for Confounding

- We receive a dataset with twenty variables: D, Y and 18 more.
  - Which variables should we include as controls?
- We usually should **NOT** include all of them

Controlling for Confounding

- We receive a dataset with twenty variables: D, Y and 18 more.
  - Which variables should we include as controls?
- We usually should **NOT** include all of them
  - Only the variables **necessary** to stop confounding

Controlling for Confounding

- We receive a dataset with twenty variables: D, Y and 18 more.
  - Which variables should we include as controls?
- We usually should **NOT** include all of them
  - Only the variables **necessary** to stop confounding
  - Including unnecessary variables can produce bias

Controlling for Confounding

- We receive a dataset with twenty variables: D, Y and 18 more.
  - Which variables should we include as controls?
- We usually should **NOT** include all of them
  - Only the variables **necessary** to stop confounding
  - Including unnecessary variables can produce bias
  - We lose power (degrees of freedom) for every control we add

Controlling for Confounding

- We receive a dataset with twenty variables: D, Y and 18 more.
  - Which variables should we include as controls?
- We usually should **NOT** include all of them
  - Only the variables **necessary** to stop confounding
  - Including unnecessary variables can produce bias
  - We lose power (degrees of freedom) for every control we add
  - And additional variables reduce overlap (increase model-dependence)

Controlling for Confounding

- We receive a dataset with twenty variables: D, Y and 18 more.
  - Which variables should we include as controls?
- We usually should **NOT** include all of them
  - Only the variables **necessary** to stop confounding
  - Including unnecessary variables can produce bias
  - We lose power (degrees of freedom) for every control we add
  - And additional variables reduce overlap (increase model-dependence)

Section 2

Which Variables to Control For

Causal Diagrams (DAGs)

- To know which variables to control for, it helps to draw a causal diagram

Causal Diagrams (DAGs)

- To know which variables to control for, it helps to draw a causal diagram
  - A **Directed Acyclical Graph** (DAG)

Causal Diagrams (DAGs)

- To know which variables to control for, it helps to draw a causal diagram
  - A **Directed Acyclical Graph** (DAG)
  - Arrows only in one direction

Causal Diagrams (DAGs)

- To know which variables to control for, it helps to draw a causal diagram
  - A **Directed Acyclical Graph** (DAG)
  - Arrows only in one direction
  - No circular loops!

Causal Diagrams (DAGs)

$$\text{Treatment} \longrightarrow \text{Outcome}$$

Causal Diagrams (DAGs)

- Causation is like **Water**, flowing along the graph
  - We want to focus on one 'flow' of causation from treatment to outcomes

Causal Diagrams (DAGs)

- Causation is like **Water**, flowing along the graph
  - We want to focus on one ’flow’ of causation from treatment to outcomes
  - Avoiding mixing with the other flows of causation in the network

Causal Diagrams (DAGs)

Causal Diagrams (DAGs)

- Three Rules to achieve Conditional Independence:
  1. Include as controls enough variables to **block all back-door paths** from treatment to the outcome

Causal Diagrams (DAGs)

- Three Rules to achieve Conditional Independence:
  1. Include as controls enough variables to **block all back-door paths** from treatment to the outcome
  2. Exclude any variables that are **post-treatment**

Causal Diagrams (DAGs)

- Three Rules to achieve Conditional Independence:
  1. Include as controls enough variables to **block all back-door paths** from treatment to the outcome
  2. Exclude any variables that are **post-treatment**
  3. Exclude any variables that are **colliders**

1. Back-door Paths

- To identify back-door paths:

1. Back-door Paths

- To identify back-door paths:
  - Start with an arrow pointing at treatment

1. Back-door Paths

- To identify back-door paths:
  - Start with an arrow pointing at treatment
  - Trace the path ’backwards’ (the direction of the arrows doesn’t matter) until you reach the outcome

1. Back-door Paths

- To identify back-door paths:
  - Start with an arrow pointing at treatment
  - Trace the path 'backwards' (the direction of the arrows doesn't matter) until you reach the outcome
  - Repeat for every possible path from treatment to outcome

1. Back-door Paths

- To identify back-door paths:
  - Start with an arrow pointing at treatment
  - Trace the path ’backwards’ (the direction of the arrows doesn’t matter) until you reach the outcome
  - Repeat for every possible path from treatment to outcome
- **Block back-door paths** by controlling for any variable along the path

1. Back-door Paths

- To identify back-door paths:
  - Start with an arrow pointing at treatment
  - Trace the path 'backwards' (the direction of the arrows doesn't matter) until you reach the outcome
  - Repeat for every possible path from treatment to outcome
- **Block back-door paths** by controlling for any variable along the path
- Identify the **Minimum set of controls** to achieve conditional independence

1. Back-door Paths

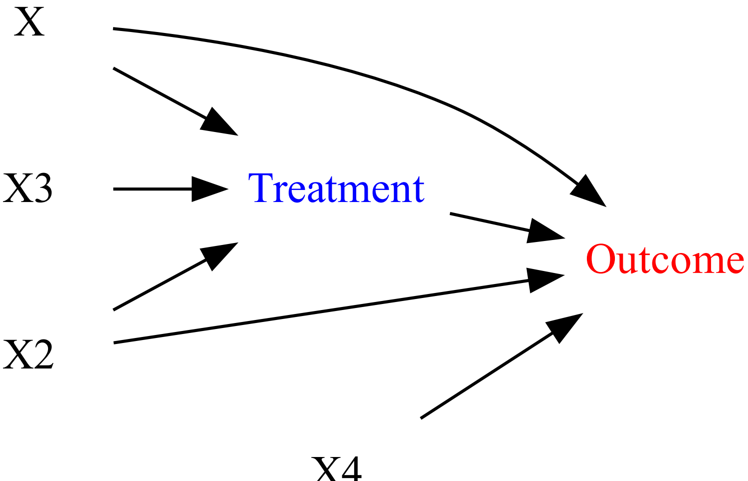
- To identify back-door paths:
  - Start with an arrow pointing at treatment
  - Trace the path 'backwards' (the direction of the arrows doesn't matter) until you reach the outcome
  - Repeat for every possible path from treatment to outcome
- **Block back-door paths** by controlling for any variable along the path
- Identify the **Minimum set of controls** to achieve conditional independence
  - *Any* set of variables which blocks *All* back-door paths

1. Back-door Paths

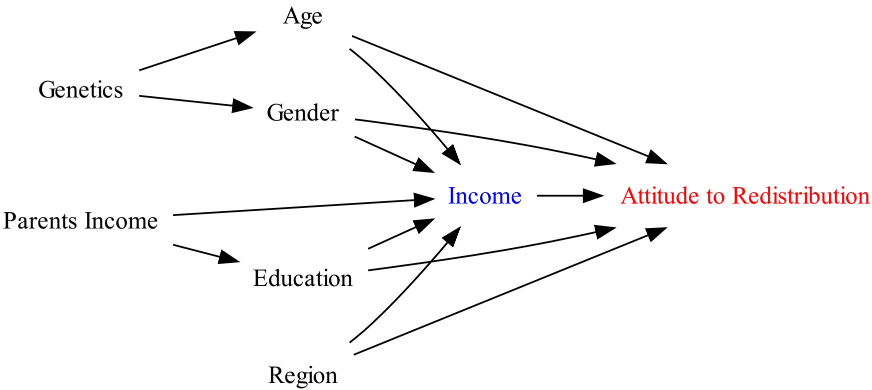
- To identify back-door paths:
  - Start with an arrow pointing at treatment
  - Trace the path 'backwards' (the direction of the arrows doesn't matter) until you reach the outcome
  - Repeat for every possible path from treatment to outcome
- **Block back-door paths** by controlling for any variable along the path
- Identify the **Minimum set of controls** to achieve conditional independence
  - *Any* set of variables which blocks *All* back-door paths
  - Include these as control variables in our regression

1. Back-door Paths

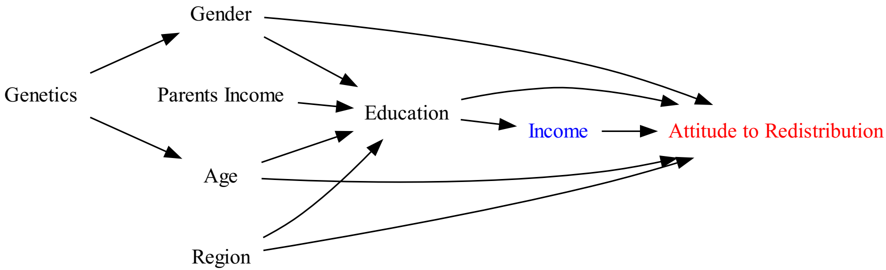
1. Back-door Paths



1. Back-door Paths



1. Back-door Paths



2. Post-treatment Variables

- Including **post-treatment** variables will introduce bias

2. Post-treatment Variables

- Including **post-treatment** variables will introduce bias
  - Because variables measured 'after' treatment can also be affected by treatment

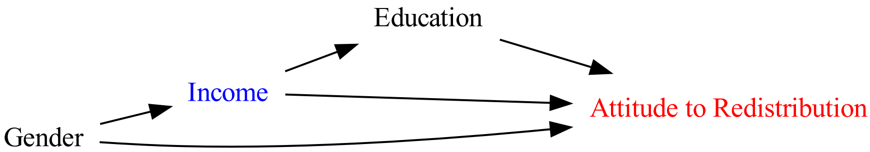
2. Post-treatment Variables

- Including **post-treatment** variables will introduce bias
  - Because variables measured ’after’ treatment can also be affected by treatment
  - They’re not confounders, but **mediating variables**

2. Post-treatment Variables

- Including **post-treatment** variables will introduce bias
    - Because variables measured ’after’ treatment can also be affected by treatment
    - They’re not confounders, but **mediating variables**
    - Controlling for them changes the definition of the causal effect we are estimating

2. Post-treatment Variables



3. Colliders

- Colliders are variables on back-door paths which have arrows pointing both into them and out of them

3. Colliders

- Colliders are variables on back-door paths which have arrows pointing both into them and out of them
- The water 'collides' in both directions but the source variables are not correlated, so produce no bias

3. Colliders

- Colliders are variables on back-door paths which have arrows pointing both into them and out of them
- The water ’collides’ in both directions but the source variables are not correlated, so produce no bias
- But if we do ’control’ for a collider we *introduce* a bias in the relationship between D and Y

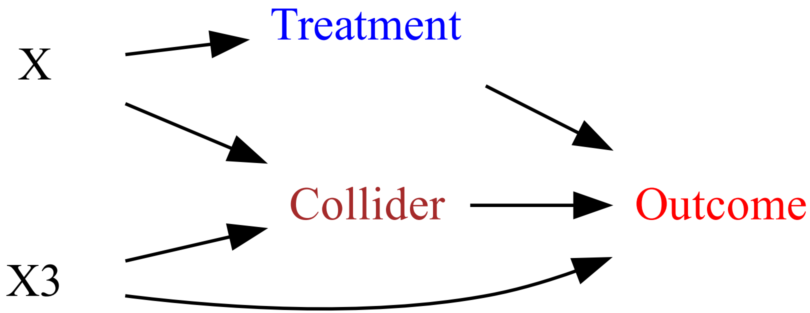
3. Colliders

- Colliders are variables on back-door paths which have arrows pointing both into them and out of them
- The water 'collides' in both directions but the source variables are not correlated, so produce no bias
- But if we do 'control' for a collider we *introduce* a bias in the relationship between D and Y
- So we must avoid controlling for them

3. Colliders

- Colliders are variables on back-door paths which have arrows pointing both into them and out of them
- The water ’collides’ in both directions but the source variables are not correlated, so produce no bias
- But if we do ’control’ for a collider we *introduce* a bias in the relationship between D and Y
- So we must avoid controlling for them
- Hard!

3. Colliders



3. Colliders

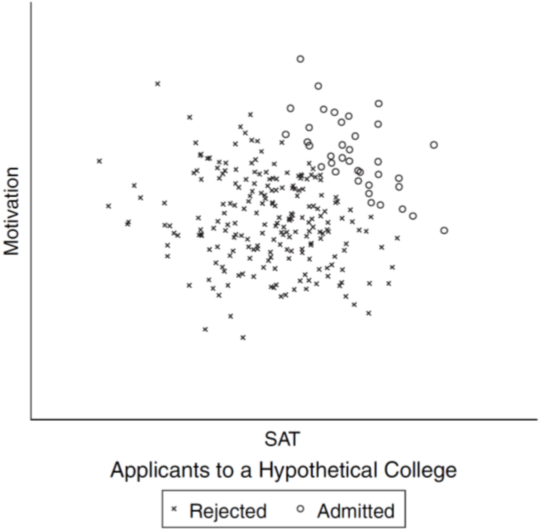
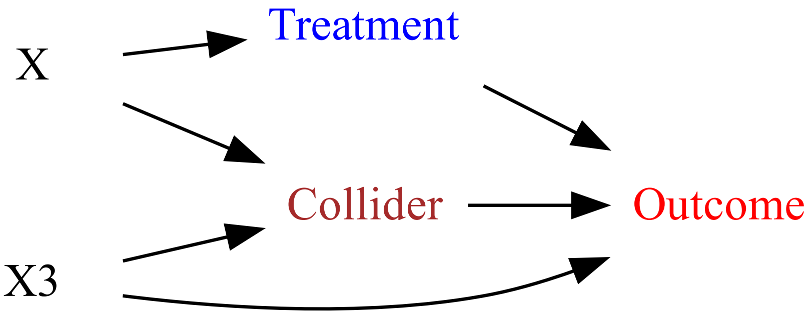
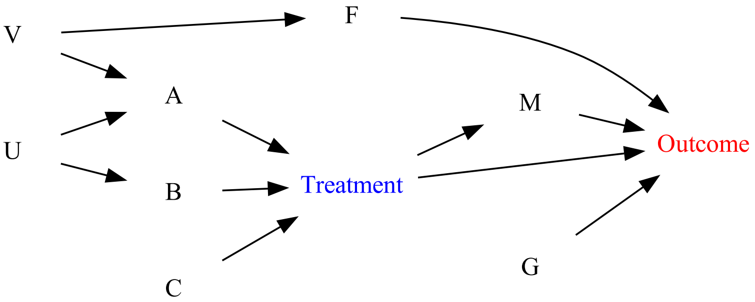


Figure 3.4: Simulation of conditional dependence within values of a collider variable.

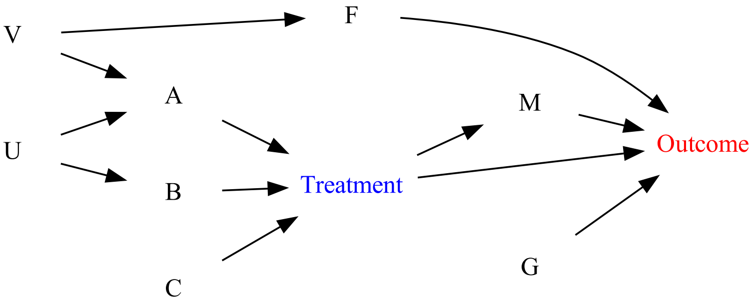
3. Colliders



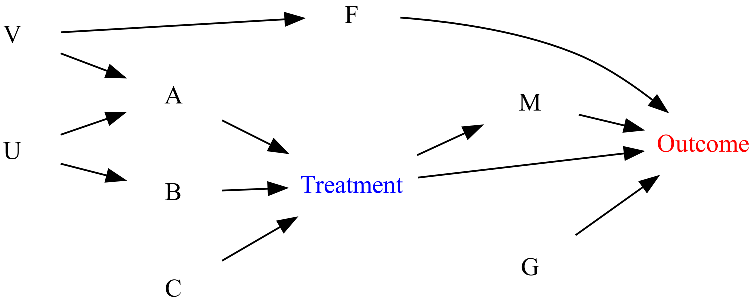
Example adapted from Morgan and Winship, p.72



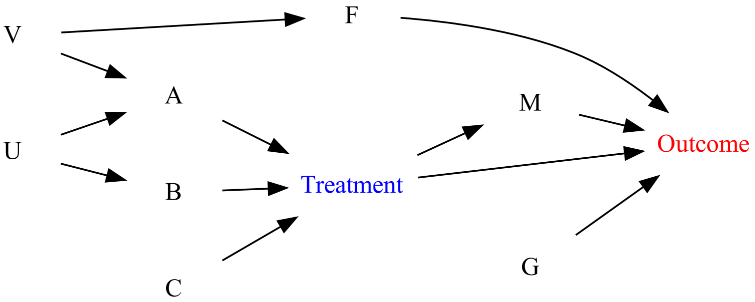
Example adapted from Morgan and Winship, p.72



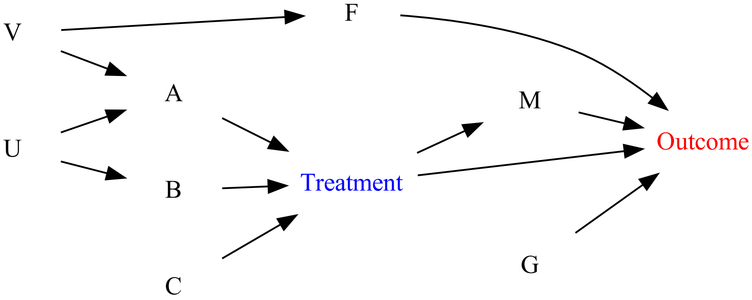
Example adapted from Morgan and Winship, p.72



Example adapted from Morgan and Winship, p.72



Example adapted from Morgan and Winship, p.72



Causal Diagrams (DAGs)

- Three Rules to achieve Conditional Independence:
  1. Include as controls enough variables to **block all back-door paths** from treatment to the outcome

Causal Diagrams (DAGs)

- Three Rules to achieve Conditional Independence:
  1. Include as controls enough variables to **block all back-door paths** from treatment to the outcome
     - In practice, variables which theory and past studies identify as potential confounderss

Causal Diagrams (DAGs)

- Three Rules to achieve Conditional Independence:
  1. Include as controls enough variables to **block all back-door paths** from treatment to the outcome
     - In practice, variables which theory and past studies identify as potential confounderss
  2. Exclude any variables that are **post-treatment**

Causal Diagrams (DAGs)

- Three Rules to achieve Conditional Independence:
    1. Include as controls enough variables to **block all back-door paths** from treatment to the outcome
        - In practice, variables which theory and past studies identify as potential confounderss
    2. Exclude any variables that are **post-treatment**
        - In practice, know when your variables were measured

Causal Diagrams (DAGs)

- Three Rules to achieve Conditional Independence:
  1. Include as controls enough variables to **block all back-door paths** from treatment to the outcome
     - In practice, variables which theory and past studies identify as potential confounderss
  2. Exclude any variables that are **post-treatment**
     - In practice, know when your variables were measured
  3. Exclude any variables that are **colliders**

Causal Diagrams (DAGs)

- Three Rules to achieve Conditional Independence:
  1. Include as controls enough variables to **block all back-door paths** from treatment to the outcome
     - In practice, variables which theory and past studies identify as potential confounderss
  2. Exclude any variables that are **post-treatment**
     - In practice, know when your variables were measured
  3. Exclude any variables that are **colliders**
     - In practice, don't include unnecessary controls