Causal Diagrams and Identifying Causal Effects

Directed Acyclic Graphs (DAGs)

ECON3500: Econometrics and Applications

Spring 2026

Learning Objectives

By the end of this lecture, you will be able to:

Draw and interpret **directed acyclic graphs (DAGs)**

Identify **causal paths**, **backdoor paths**, and **colliders**

Determine which variables to **control for** to identify a causal effect

Explain why controlling for the wrong variables can **bias** your estimates

Apply DAG logic to real research questions

Reading

Huntington-Klein, *The Effect*: **Chapters 6, 7, and 8** (available free at theeffectbook.net)

What Is Causality?

The Data Generating Process

Every dataset we observe is produced by a **data generating process (DGP)** — the real-world laws, behaviors, and institutions that determine the values we see.

Our challenge as econometricians:

We observe the *data*, not the DGP

Multiple DGPs could produce the same data patterns

We need a framework for reasoning about *which variation* in our data answers our research question

Identification

Identification is the process of figuring out what part of the variation in your data answers your research question.

Association vs. Causation

Association

Two variables move together (correlation). X and Y are related statistically.

Causation

Changing X **causes** Y to change. Formally: if we could *intervene* to change X, the distribution of Y would change as a result.

Association: Useful for prediction, not necessarily for policy

Causation: Required for policy impact, program evaluation, treatment effects

The Framework: Causal Diagrams (DAGs)

A **directed acyclic graph (DAG)** is a visual representation of a data generating process:

Variables (nodes) — circles representing measured or unmeasured quantities

Causal relationships (arrows) — direction of causality

No cycles — causality flows forward (you can't follow arrows back to where you started)

Key insight: Once we draw the DAG, we can **mechanically determine** which variables to control for to identify a causal effect.

> ⚠ **Important assumption**
>
> Every variable and arrow that is *not* on the diagram is an assumption we're making. Drawing a DAG forces you to be explicit about your identifying assumptions.

Drawing and Reading DAGs

DAG Basics: Nodes and Arrows

Node: A variable in the system.

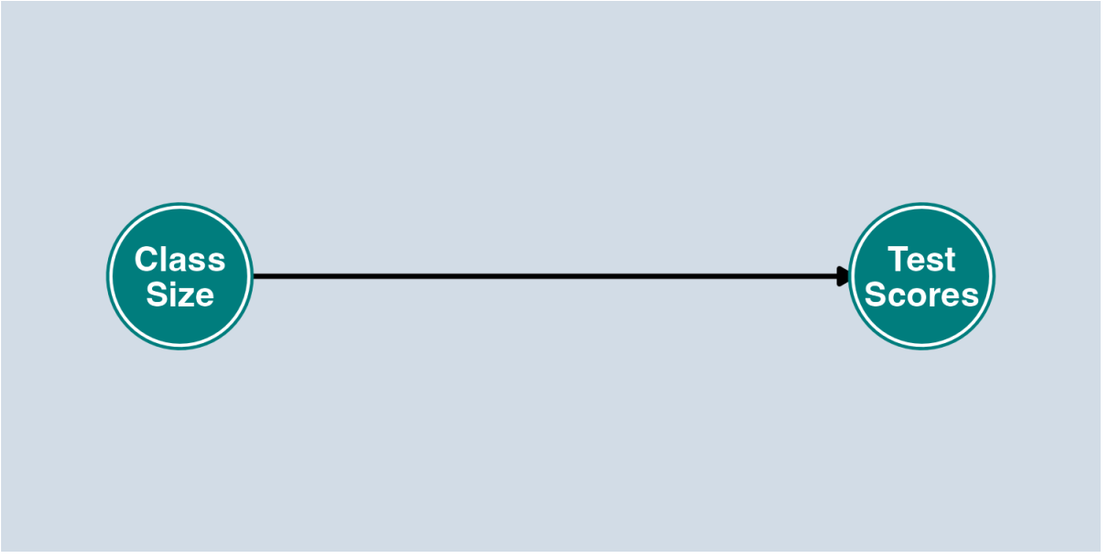
Each circle represents a variable — either observed or unobserved.

Arrow: $X \to Y$ means "X causes Y."

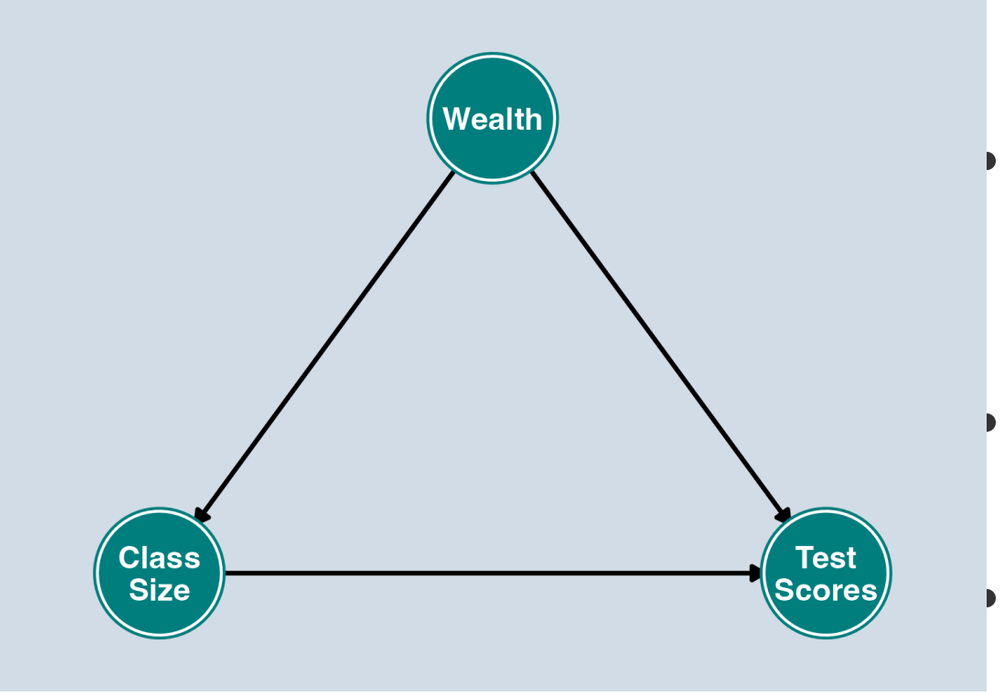
Direction matters (from cause to effect)

X is a **parent** (ancestor) of Y

Y is a **child** (descendant) of X



Example: Class Size and Student Achievement



Variables:

- X = Class Size, Y = Test Scores, Z = Wealth

Causal arrows:

- Wealth → Class Size (wealthy areas have smaller classes)
- Wealth → Test Scores (wealthier students score higher)
- Class Size → Test Scores (smaller classes improve learning)

Causal (Front Door) Paths

$X \to Y$ (or $X \to M \to Y$)

Direct or indirect effect of X on Y, with all arrows pointing *away from* X.

These are **good paths** — they represent what we *want* to estimate.

Example: Class Size → Test Scores

Backdoor Paths

$X \leftarrow Z \rightarrow Y$

X and Y are connected through a common cause Z. These are **bad paths** — they create **confounding**.

Example: Class Size ← Wealth → Test Scores

Connection to OVB

A backdoor path is exactly what creates **omitted variable bias**. If Z is a common cause of X and Y and you leave it out of the regression, you have an open backdoor path — and OVB.

Collider Paths

$X \rightarrow Z \leftarrow Y$

X and Y both cause Z. These paths are **closed by default** — no confounding flows through them *unless* we make a mistake.

Example: Talent → Job Hiring ← Connections

No spurious relationship between Talent and Connections… unless we condition on who got hired.

Open and Closed Paths

Key Concept

A path is **open** if the relationship flows freely between variables along it. A path is **closed** if something blocks that flow.

Path Type	Default Status	What controls do
Backdoor ($X \leftarrow Z \rightarrow Y$)	**Open**	Controlling for Z **closes** it
Collider ($X \rightarrow Z \leftarrow Y$)	**Closed**	Controlling for Z **opens** it

Our goal: Close all bad (backdoor) paths while keeping good (causal) paths open.

Confounding and the OVB Connection

The Confounding Problem

Confounding

When X and Y share a common cause Z, they are **confounded**. A regression of Y on X will pick up both the causal effect AND the confounding association.

In the class size example:

$$\text{Association} = \underbrace{\text{Causal Effect}}_{\text{Class Size} \rightarrow \text{Scores}} + \underbrace{\text{Confounding Bias}}_{\text{via Wealth}}$$

OVB Is an Open Backdoor Path

Remember **omitted variable bias** from Chapter 6?

OVB occurs when a variable that belongs in the regression is left out. In DAG language:

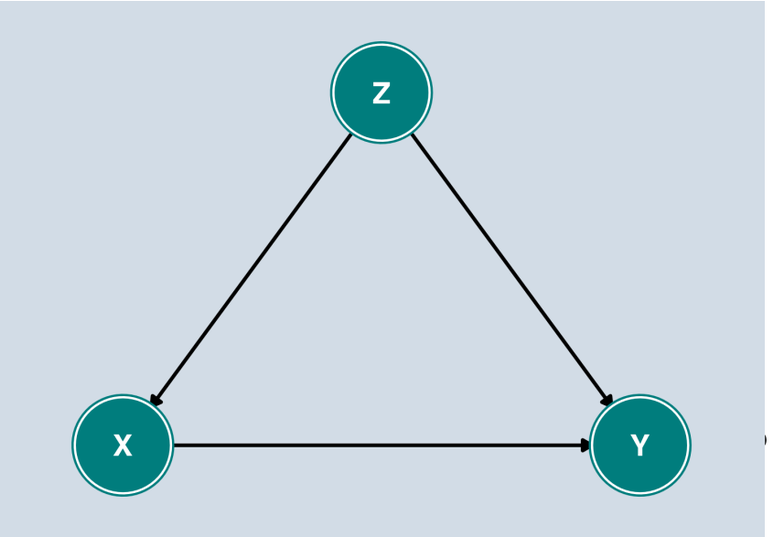
The omitted variable creates a **backdoor path** between X and Y

Leaving it out of the regression means the path stays **open**

The bias flows through that open path into your estimate of $\hat{\beta_1}$

> ⓘ **OVB ↔ Backdoor Paths**
>
> **Every case of OVB is an open backdoor path.** The DAG just gives you a visual way to see it — and to figure out what to control for.

Blocking Backdoor Paths



Blocking a Backdoor Path

Control for (include in regression) the confounder Z to "close" the backdoor path $X \leftarrow Z \rightarrow Y$.

In the class size example:

- **Without controlling for Wealth:** Regression picks up confounding bias (path is **open**)
- **With controlling for Wealth:** Regression isolates the causal effect (path is **closed**)

Knowledge Check: Direction of Bias

Think About It

If we regress Test Scores on Class Size *without* controlling for Wealth, will the class size coefficient be biased up or down?

Answer: The coefficient on class size is biased *toward zero* (understates the negative effect).

Wealthy areas → smaller classes AND higher scores

This creates a *positive* confounding association between class size and scores

The positive confounding partially cancels the true negative causal effect

Which Variables to Control For?

The Backdoor Criterion

To identify the causal effect of X on Y:

Backdoor Criterion

Control for a set of variables $\mathbf{Z}$ such that:

1. No variable in $\mathbf{Z}$ is a **descendant** of X
2. Controlling for $\mathbf{Z}$ **closes all backdoor paths** from X to Y

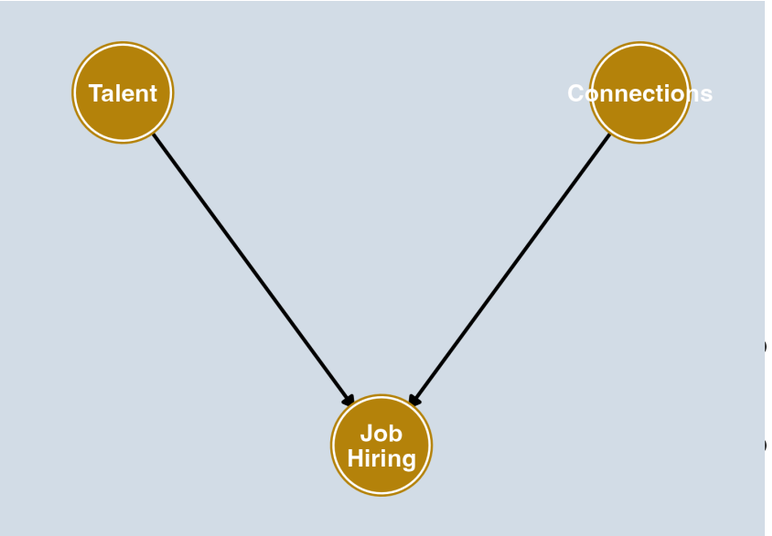
In plain language:

Do control for common causes (confounders) — they create open backdoor paths

Don't control for variables that X affects — that would block part of the causal effect

Don't control for colliders — that would *open* a path that was safely closed

What NOT to Control For: Colliders



> ⚠ **Collider Bias**
>
> If you control for a **collider** — a variable caused by *both* X and Y — you **open** a previously closed path and create bias.

- Talent and Connections both cause Job Hiring
- The path Talent → Job Hiring ← Connections is **closed by default**
- If we condition on people who were hired (control for the collider), the path **opens**
- Among the hired: low Talent becomes correlated with high Connections

What NOT to Control For: Mediators

⚠ **Bad Control: Mediator**

If $X \to M \to Y$, then M is a **mediator**. Controlling for M blocks part of the causal effect of X on Y.

Example: Does education affect earnings?

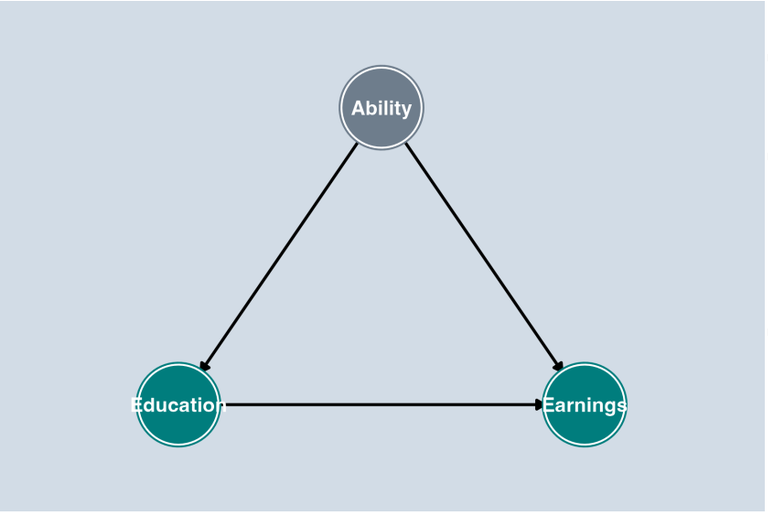
Education → Job Type → Earnings

If we control for Job Type, we block the indirect effect of education that works *through* job placement

We'd only estimate the effect of education holding job type fixed — not the total causal effect

Applications: Identifying Causal Effects

Does Education Cause Earnings?



1. **Causal path?** Yes: Education → Earnings
2. **Backdoor paths?** Yes: Education ← Ability → Earnings
3. **What to control for?** Ability

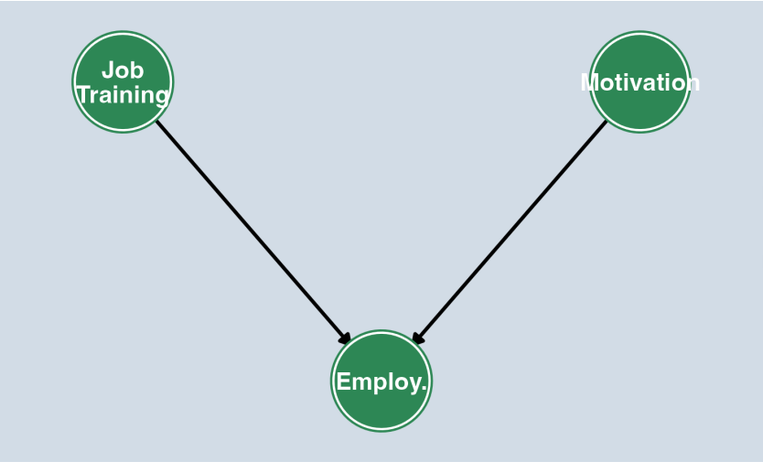
Gray node = unobserved variable

Does Education Cause Earnings? (cont.)

Challenge: Ability is **unobserved**! We can't directly control for it.

This is why economists turn to **identification strategies** — instruments, experiments, panel data — to close backdoor paths when confounders are unobservable.

Does Job Training Improve Employment?



The problem without randomization:

- Motivated workers are more likely to sign up for training
- Motivated workers are also more likely to find jobs
- Backdoor path: Training ← Motivation → Employment

. . .

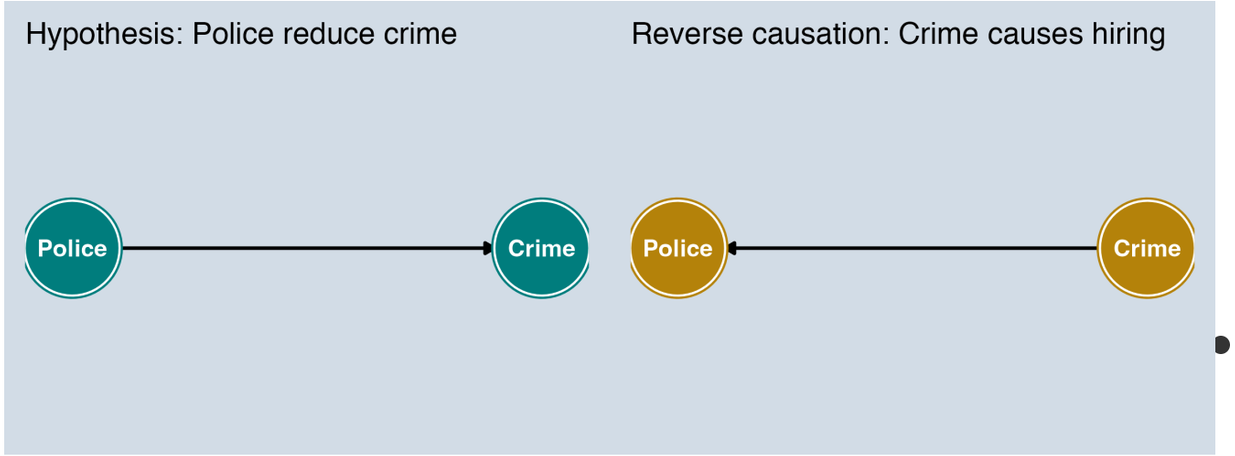
With an RCT:

- Randomly assign workers to training
- Now Motivation does **not** cause Training — that arrow is severed
- All backdoor paths are closed *by design*

Important

This is why **randomized controlled trials (RCTs)** are the gold standard for causal inference — they eliminate confounding mechanically.

Reverse Causation

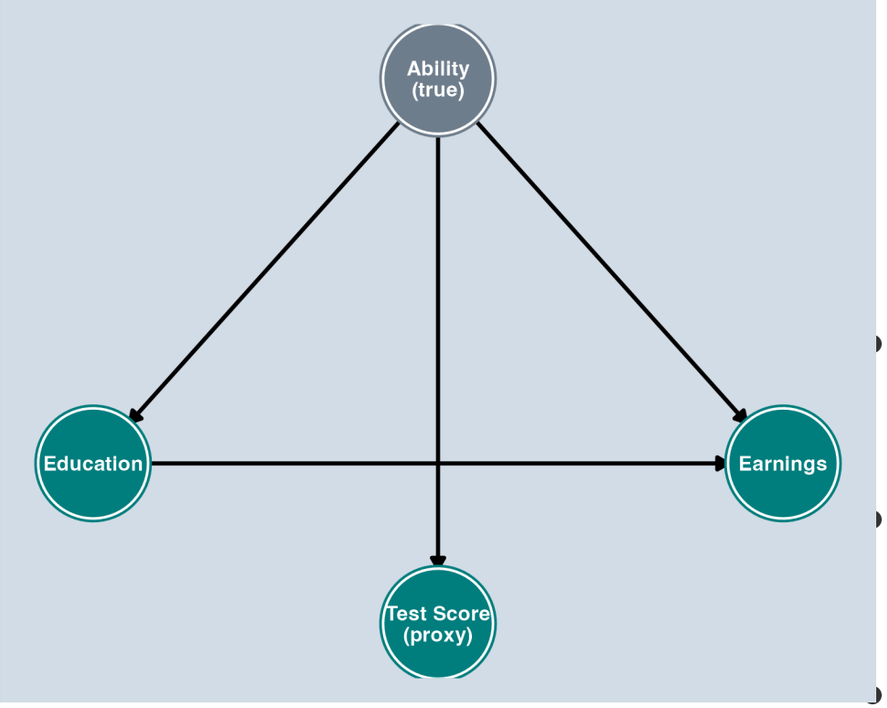


> ⚠ **Reverse Causation**
>
> Sometimes the arrow between X and Y runs in the **opposite direction** from what we assume.

- We observe a positive correlation between police and crime
- **Our hypothesis:** More police → less crime
- **The problem:** Cities with more crime hire more police

In DAG terms: we've drawn the wrong DAG. The true DGP has the arrow reversed (or both arrows exist — **simultaneity**).

Measurement Error



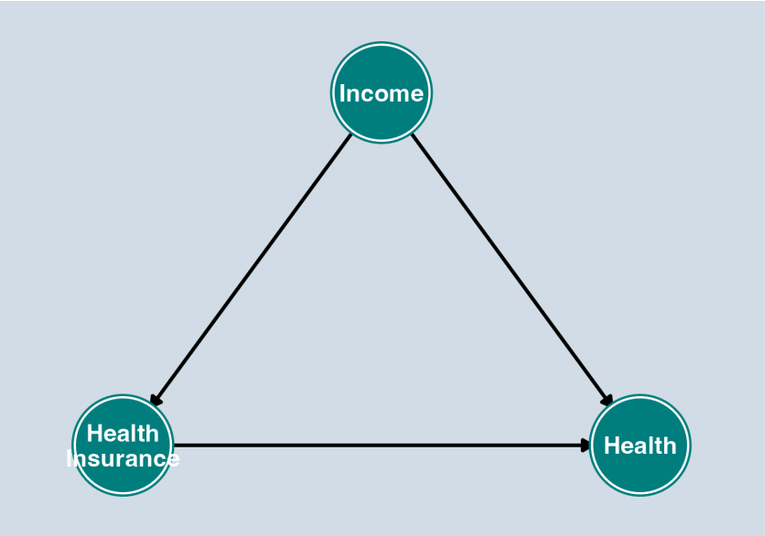
Gray = unobserved

> ⚠ **Measurement Error**
>
> When we can't perfectly measure a variable in our DAG, we introduce **measurement error**.

- We want to control for Ability to close a backdoor path
- We can't measure Ability directly, so we use a proxy (Test Score)
- But Test Score ≠ Ability (noisy measurement)
- Controlling for the proxy only *partially* closes the backdoor path
- Result: our estimate is still biased — **attenuation bias**

Knowledge Checks and Practice

Knowledge Check 1: Health Insurance



Does Income confound the effect of Health Insurance on Health?

To estimate the effect of Health Insurance on Health, should we control for Income?

Knowledge Check 1: Answer

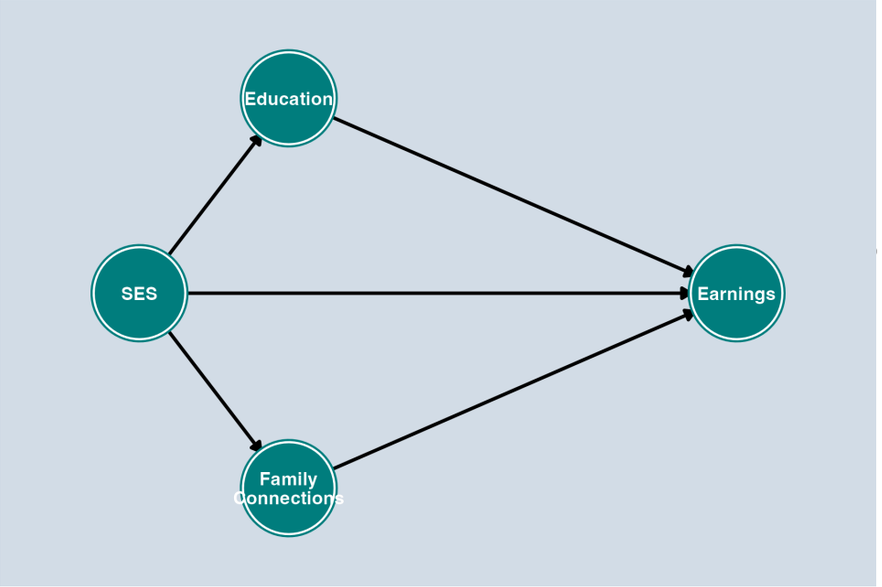
Answer

There is a backdoor path: Health Insurance ← Income → Health

Income is a common cause of both Health Insurance and Health

So yes, we **should** control for Income — it's a confounder

Knowledge Check 2: SES and Earnings



Questions:

. What are the backdoor paths from SES to Earnings?

. What minimal set of variables should we control for?

Knowledge Check 2: Answer

Answer

1. **Backdoor path:** SES ← (shared causes) → Earnings via Family Connections
2. **Controls:** Family Connections (to close the backdoor)
   - Do **NOT** control for Education — it is a *descendant* of SES (mediator)

Practice: Draw Your Own DAG

Think of a causal question from your own interest or research paper:

What is the treatment (X) and outcome (Y)?

What variables might confound this relationship?

Draw a simple DAG (3–5 variables)

Identify the backdoor paths

What would you need to control for?

> **Tip from *The Effect***
>
> Start with your treatment and outcome. Then ask: “What are all the things that cause my treatment? What are all the things that cause my outcome? Do any of those overlap?” Those overlapping causes are your confounders.

Summary

Key Takeaways

The DGP is what we're trying to understand. A DAG is our best model of that process.

Association ≠ Causation. DAGs help us think systematically about *why*.

OVB = an open backdoor path. If a confounder is omitted, the path stays open and your estimate is biased.

The Backdoor Criterion is mechanical. Once you draw the DAG, you can determine what to control for.

Collider bias is real. Controlling for the wrong variable can *introduce* bias by opening a closed path.

Randomization is powerful. Random assignment closes backdoor paths by design.

What's Next?

These ideas connect directly to **Chapter 9: Assessing Regression Validity**:

Omitted variable bias = an open backdoor path you haven't closed

Measurement error = your DAG nodes don't match what you actually measured

Simultaneity / reverse causation = the arrows in your DAG might be wrong

The DAG framework gives you a visual language for diagnosing every threat to internal validity.

For the Curious: d-Separation

The backdoor criterion is actually a special case of a more general concept called **d-separation**.

Two variables are **d-separated** by $\mathbf{Z}$ if every path between them is closed after conditioning on $\mathbf{Z}$

If d-separated, the variables are conditionally independent given $\mathbf{Z}$

You don't need to know d-separation for this course — the backdoor criterion is the tool you'll use. But if you want to go deeper, see *The Effect* Chapter 8 or Pearl (2009).