Recap: Causal Graphs and Estimating Effects

DS701 Session 18 — Wed Nov 4, 2026

Today’s plan

Knowledge-Check Review

KC 1: Three building blocks

Name the three elementary structures every path in a causal DAG is built from, and for each say (i) what the middle node is called and (ii) whether the path through it is open or blocked ***by default****, and what happens when you condition on that middle node.*

KC 2: “Controlling for more can only help”

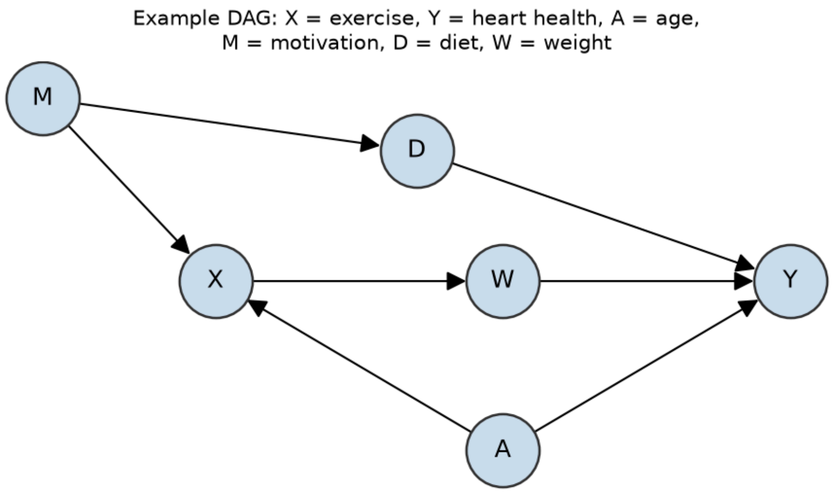
In the tutoring example, X *= tutoring and* Y *= final grade, with* $\mathrm{GPA}_{\text{prior}} \to X$*,* $\mathrm{GPA}_{\text{prior}} \to Y$*,* $X \to \mathrm{Hours} \to Y$ *and* $X \to \mathrm{HonorRoll} \leftarrow Y$*. A colleague argues that adjusting for* $\{\mathrm{GPA}_{\text{prior}}, \mathrm{Hours}, \mathrm{HonorRoll}\}$ *must be* at least as good* as adjusting for $\{\mathrm{GPA}_{\text{prior}}\}$ alone, “because controlling for more can only help.” Using the backdoor procedure, explain why the colleague is wrong about the **total** effect of tutoring — say what each of Hours and HonorRoll does to the estimate.*

KC 3: When the confounder is not in the data

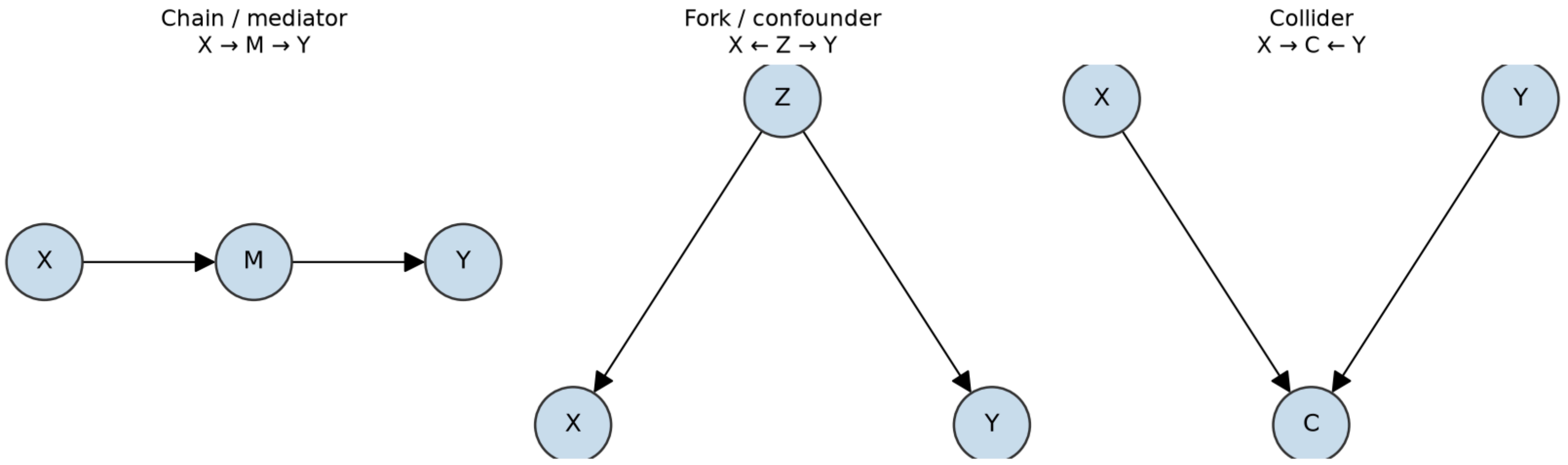
Lecture I showed that randomization removes selection bias; this lecture showed that adjusting for a valid backdoor set does the same job with observational data. Suppose you believe there is a confounder you ***cannot measure*** *(e.g. "motivation" driving both who takes a training program and later earnings). Explain in a sentence why regression adjustment cannot fix this, then name* ***one*** *of the three natural-experiment designs from the lecture, say in a sentence what "as good as random" variation it exploits, and state the key assumption it relies on instead.*

Highlights

The graph is your assumptions, not the data's



Every path is chains, forks and colliders



The golden rule: adjust for forks, not colliders or mediators

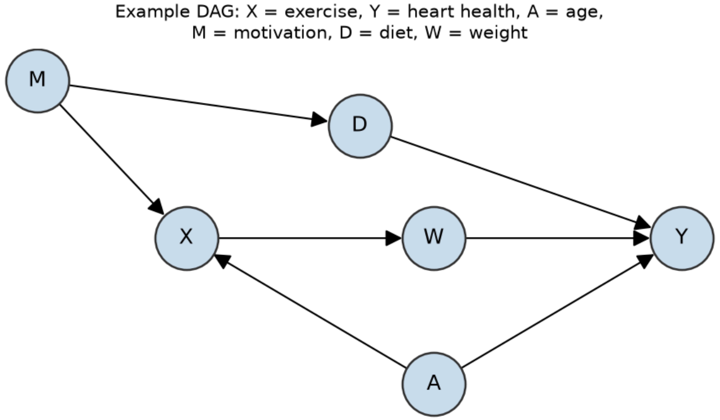
structure	middle node	adjust for it?	why
fork $X \leftarrow Z \rightarrow Y$	confounder	**yes**	blocks the spurious backdoor path
chain $X \rightarrow M \rightarrow Y$	mediator	**no***	it *is* the effect you want (overcontrol)
collider $X \rightarrow C \leftarrow Y$	collider	**no!**	conditioning *creates* association

* unless you specifically want the *direct*, not the total, effect.

The backdoor criterion is a procedure

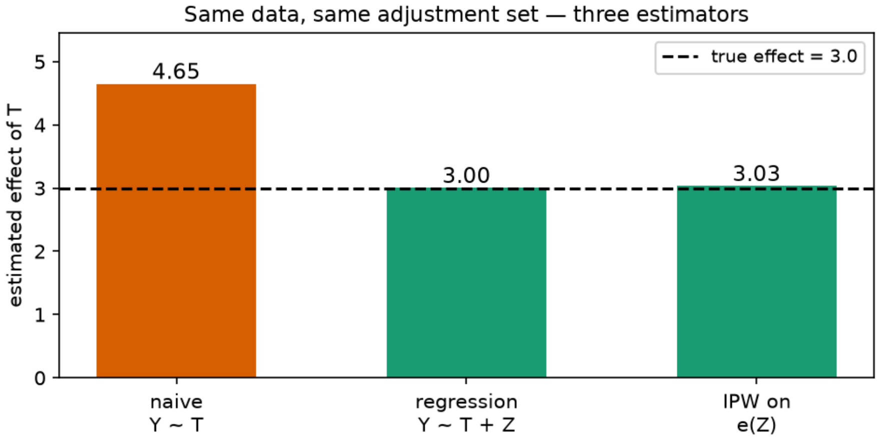
A **backdoor path** starts with an arrow *into* X — confounding in graph language. S is a valid adjustment set if it **blocks every backdoor path** and contains **no descendant of** X.

Worked example, five steps



$$A \to X,\ A \to Y,\ M \to X,\ M \to D,\ D \to Y,\ X \to W,\ W \to Y$$

After identification: the estimator is the easy part



Given a valid S (**ignorability**), any of these recovers the effect:

Lecture simulation: $Y = 3T + 2Z + \varepsilon$, sicker ($Z$) more likely treated.

Unmeasured confounders: nature's randomization

Adjustment only blocks paths through variables you **measured**. Otherwise look for treatment variation that is *as good as random*:

In-Class Activity

Activity: backdoor identification by hand, then watch a wrong adjustment set lie

Goal: on five DAGs (drawn with `networkx`), list every backdoor path, mark it open or blocked, code the two rules into a checker, and propose one valid and one *invalid* adjustment set per DAG. Then simulate data from one of those DAGs with a **known** effect and estimate it four ways — naive, correct set, wrong set (mediator / collider), inverse-propensity weighting — and see which ones recover the truth.

- Work in groups of 2–3. Open **your section's** notebook — the DAG subset and the simulated effect differ by section.
- Parts marked **(autograded)** are submitted to Gradescope; the rest is participation.
- Staff will circulate — be ready to explain any part of your work.
- Dependencies: `numpy`, `pandas`, `matplotlib`, `networkx`, `statsmodels`, `scikit-learn`.

Going deeper

Full lecture notes: Causal Inference II: Causal Graphs and Estimating Effects

- Reading a causal graph — node, arrow, directed path, descendant, DAG
- Open and blocked paths — the two rules
- The golden rule and Confounders: the backdoor — criterion + adjustment formula
- The backdoor procedure, by hand and the Worked example
- Colliders: adjusting can create bias — the talent/looks simulation
- Regression vs. stratification and Propensity scores — the code the stretch section builds on
- Natural experiments: DiD, IV & RD, and Assumptions you can't skip
- Where confounding first appeared: Causal Inference I