Beyond the Regression

Three Frameworks for Thinking About Causation

Why Causality Matters for Policy

Description

What patterns exist in the data?

Graduates earn more than non-graduates.

Causation

What will happen if we intervene?

If we subsidise degrees, will earnings rise?

The Correlation Trap

Hospitals

A lot of people die in hospitals. Should we close them?

Obviously not. People who are sick go to hospitals. The correlation between hospitals and death runs through illness, not through anything hospitals are doing.

Umbrellas

When more people carry umbrellas, it rains more. Should we ban umbrellas to stop the rain?

Obviously not. Rain causes umbrella use, not the other way around. The correlation runs through weather, not through anything umbrellas are doing.

The Question

> *"What is the causal effect of a university degree on earnings?"*

Our data
Administrative records on university completion and earnings for 50,000 workers.

Our task
A briefing note estimating the return to education for policy advice.

The Counterfactual

A causal effect is the difference between two states of the world:

What actually happened (data)

Person A got a degree and earns $58,000.

What would have happened (counterfactual)

Had Person A not got a degree, what would they earn? We never observe this.

The fundamental problem of causal inference

We only ever observe one outcome per person. The counterfactual is never observed. Comparing degree-holders to non-degree-holders doesn't solve this, because the two groups differ in ways that independently affect earnings.

Every causal framework is a different strategy for filling in the missing counterfactual.

The Naïve Regression

$$\log(\text{Earn}_i) = \alpha + \beta \cdot \text{Uni}_i + \varepsilon_i$$

We estimate $\hat{\beta} = 0.45$. University graduates earn roughly 45% more.

But is this causal?

People who get degrees differ from people who don't: in ability, motivation, family background. These same traits independently affect earnings. The coefficient captures both the causal effect and the selection bias.

Three frameworks.
One diagnosis.

The coefficient from the naïve regression conflates causation with selection.
Each framework responds differently to this issue.

Design-based

Also called Rubin causal model.

Find a source of quasi-random variation in the data. Often very appropriate for ex post policy evaluation.

Basic flavours:
Conditional Independence Assumption
Regression Discontinuity Design
Difference in Differences
Instrumental Variables

DESIGN-BASED

Find a Natural Experiment

Core logic: Causation is about missing data. Each person has a potential outcome with and without a degree. We only observe one. A natural experiment approximates random assignment.

Example: Regression Discontinuity

University admission requires an entrance exam score ≥ 70. Students scoring 69 and 71 are near-identical, but one gets a degree and one doesn't.

Comparing earnings just above and below the threshold gives a credible causal estimate, but only for people near the cutoff.

Result: $\hat{\beta}$ = 0.28, highly credible, but only for students near the cutoff.

DESIGN-BASED

What Does 0.28 Mean?

Students scoring 69 vs 71 differ in earnings by 28%.

But what about students scoring 40? Or 95? Their return to a degree could be very different. High-ability students might gain more from university. Low-ability students might gain less. The RDD can't tell us.

High internal validity

The estimate is credible because the variation is as good as random near the cutoff.

Low external validity

The very thing that makes it credible makes it hard to generalise to the policy question.

Cannot answer: 'What if we expanded access for everyone (people who would score far from 70)?'

DAG-based

Developed by Judea Pearl, also called do-calculus.

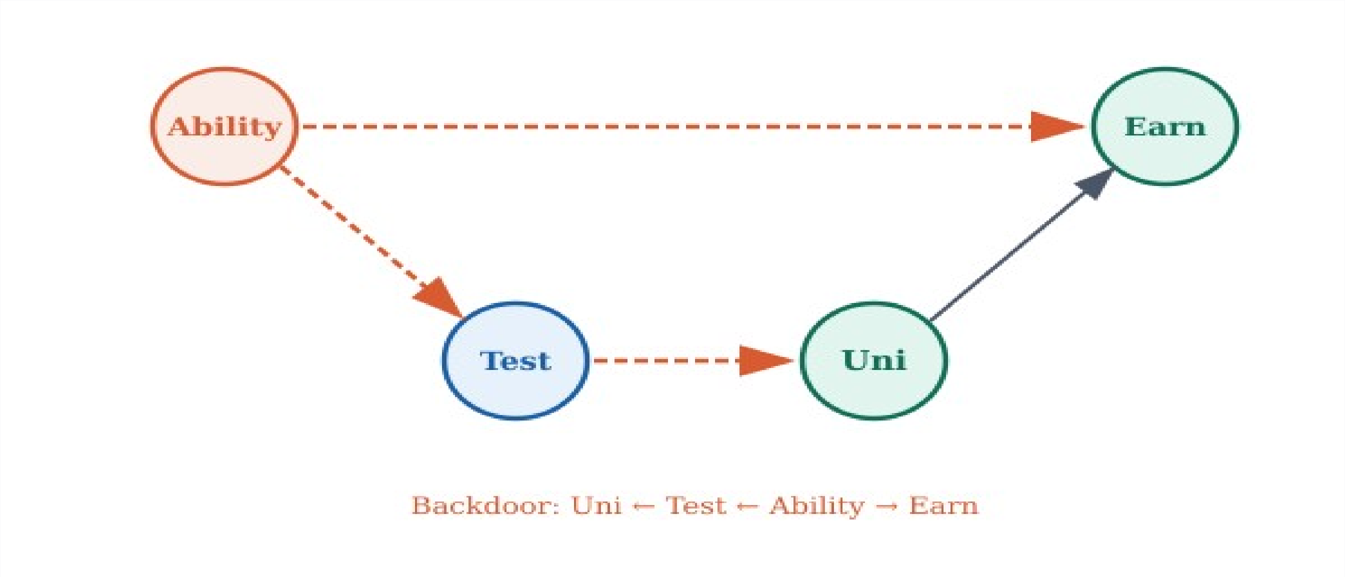
Excels in determining causality with assumptions on structure of problem. Very easy to use.

Write down what we believe about the causal structure. A directed acyclic graph makes our assumptions explicit and tells us whether identification is possible with the data we have.

DAG-BASED

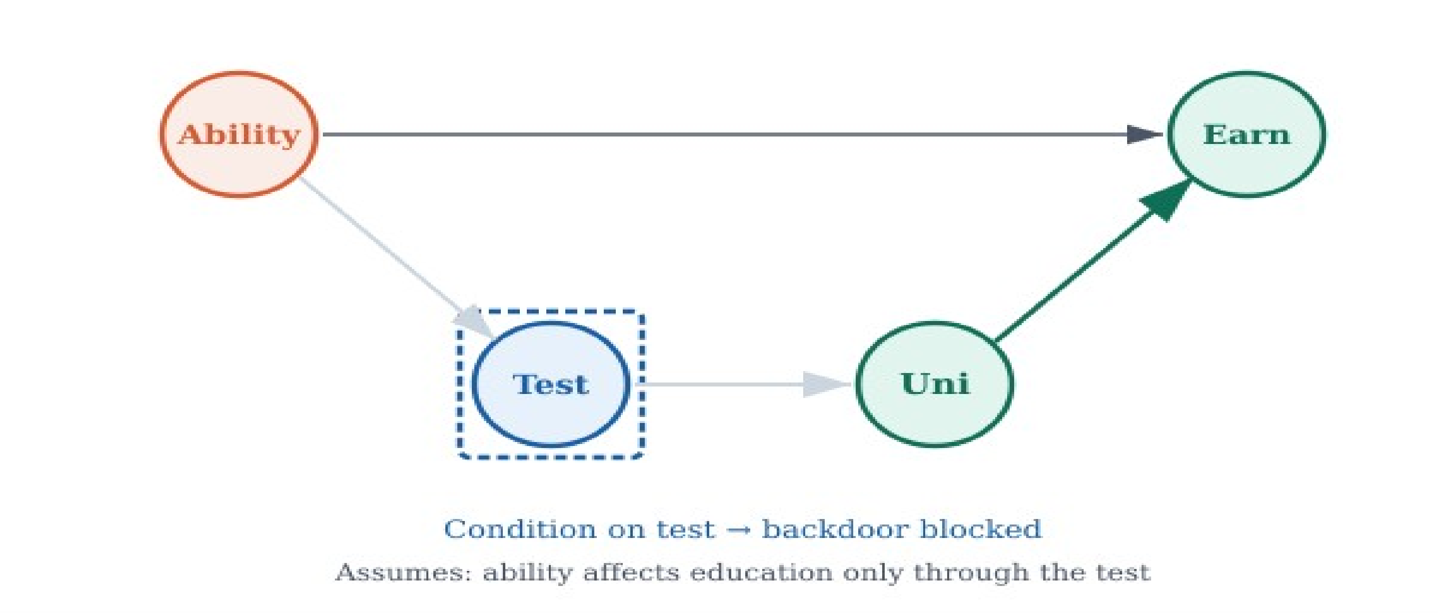
Draw the Graph

Core logic: Make our causal assumptions explicit as a directed acyclic graph. The graph tells us which variables to condition on, and which ones to leave alone.



DAG-BASED

Blocking the Backdoor



Key insight: The graph makes our assumptions visible. It tells us *what we need to condition on*, and whether our data is sufficient to do it.

Model-based

Long tradition in economics, but recently formalised by Heckman and Pinto.

Write down a model of decisions made by agents. The upside is that allows for ex ante policy evaluation, including distributional effects and welfare analysis.

Very relevant for government policy design.

MODEL-BASED

Model the Decision

Core logic: Education is a choice. Model why people choose it, model how it affects earnings, and use the selection process to correct the causal estimate.

1. Selection equation

$$Uni_i = 1 \text{ if } \gamma_1 ParEd_i - \gamma_2 Tuition_i + u_i > 0$$

Who gets a degree? Parental education and tuition costs are observed. Unobserved ability lives in u_i.

2. Outcome equation

$$\log(Earn_i) = \alpha + \beta \cdot Uni_i + \rho\lambda(\cdot) + \varepsilon_i$$

$\lambda(\cdot)$ is the correction term from the selection equation. Without it, β absorbs selection bias.

3. The connection

The correction term $\lambda(\cdot)$ comes from the selection equation. It captures how the unobservables driving the education decision correlate with the unobservables driving earnings. If $\rho > 0$, people who select into education would earn more anyway. Naive OLS overstates the return. The selection equation estimates this, and the correction term in the outcome equation removes it.

MODEL-BASED

Simulating a Policy Change

With the structural model estimated, we can ask: what happens if we subsidise tuition by $5,000?

Before subsidy	After $5k subsidy
42%	**51%**
get a degree	get a degree
Population avg: $48,200*	Population avg: $50,800*
Degree-holder avg: $62,100*	Degree-holder avg: $58,900*

Population earnings rise (more people have degrees), but degree-holder earnings fall (new entrants have lower baseline earnings).

Side by Side

	Design-based	DAG-based	Model-based
Response to bias	Find external variation	Map the causal structure	Model the decision process
Key assumption	As-if random assignment near the cutoff	The DAG is correct	The structural model is correct
Estimand	ATE at the cutoff	Causal effect conditional on graph	ATE corrected for selection
What it can't do	Generalise beyond the current situation	Simulate policy changes	Wrong model, wrong answer
Requires	A natural experiment	A credible causal graph	A structural model of the decision

Frameworks for Government Work

The right lens depends on the question we're facing.

Design-based

Ideal when a natural experiment exists. Gives us the most credible single estimate, but for a narrow group. The very thing that makes it credible (local variation) limits its relevance to broader policy questions. Usually sufficient for ex post policy evaluation.

DAG-based

Immediately useful even without running a formal causal analysis. The DAG is a thinking tool that disciplines which controls help and which ones hurt. Everyone here can use it tomorrow.

Model-based

Best suited for the question government usually asks: 'What would happen if we changed the policy?' Lets us simulate counterfactuals for people not yet affected. Demands more structure and stronger assumptions. Useful for ex ante policy evaluation, and for analysis of mechanisms.

The fundamental trade-off

Design

DAG

Model

Internal validity

Credibility of the estimate

External validity

Relevance to the policy question

More credible estimates ⟵⟶ More policy-relevant answers

The framework we reach for depends on the question we are trying to answer.