Advanced Topic A03: *Causal Inference vs. Correlation in ML*

Data Science with Python – BSc Advanced Lectures

Joerg Osterrieder

10 Minutes

Motivation: High Accuracy Is Not Enough

Your model predicts default with 95% accuracy. Should you act on it?

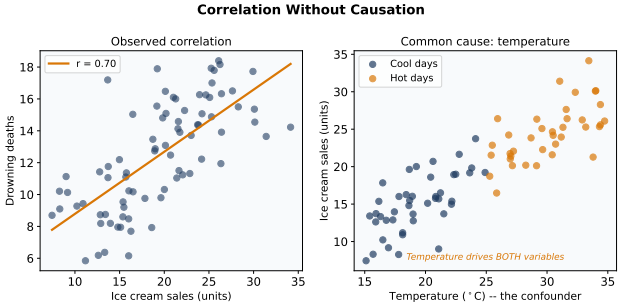
- A model trained on historical data learns correlations, not causes
- Acting on correlations can backfire: changing the input breaks the correlation
- Regulators in credit and hiring increasingly require causal justification
- Correlation tells you what happened; causation tells you what would happen if you intervene

The question is not "does X predict Y?" but "does X cause Y?"

Correlation Without Causation

The classic example: ice cream and drowning

- Ice cream sales and drowning deaths are positively correlated
- Banning ice cream would not reduce drowning
- Common cause: hot weather drives both ice cream consumption and swimming
- ML models happily learn this correlation and use it for prediction



Correlation is sufficient for prediction; it is insufficient for intervention

Three Sources of Correlation

Every observed X-Y correlation has one of three origins

(a) X causes Y:

- Genuine causal effect
- Intervention on X changes Y

(b) Y causes X:

- Reverse causation
- Intervening on X has no effect

(c) Common cause Z:

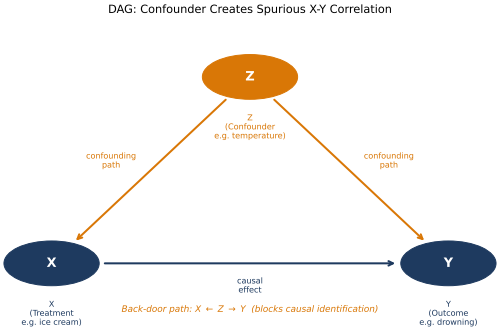
- Confounder Z drives both
- Ice cream / drowning case

Distinguishing these three requires more than observational data alone

DAGs and Confounders

Directed Acyclic Graphs make causal assumptions explicit

- Nodes are variables; directed edges are causal relationships
- A confounder Z causes both X and Y, creating a spurious correlation
- **Back-door criterion (informal):** a set Z blocks all confounding paths between X and Y; conditioning on Z recovers the causal effect
- Key rule: do not condition on a collider (a variable caused by both X and Y)



Back-door adjustment: $P(Y|do(X)) = \sum_z P(Y|X,z)P(z)$

Pearl's Do-Operator

$P(Y|X)$ vs $P(Y|do(X))$: observation vs intervention

- $P(Y|X = x)$: probability of Y given that we *observe* $X = x$ (possibly confounded)
- $P(Y|do(X = x))$: probability of Y if we *set* $X = x$ by intervention, breaking all incoming edges to X in the DAG
- These are equal only when there are no confounders (or they are all measured)
- Example: $P(\text{recovery}|\text{drug})$ vs $P(\text{recovery}|do(\text{drug}))$

Pearl (2000): Causality: Models, Reasoning, and Inference

Why ML Models Are Correlational

ERM minimises prediction error, not causal fidelity

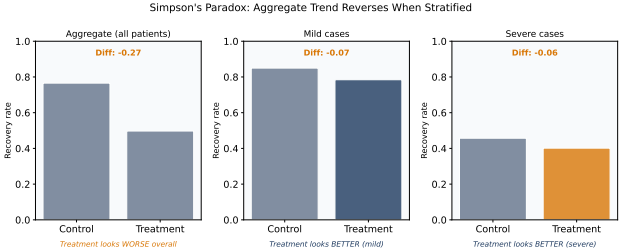
- Empirical risk minimisation exploits any predictive signal, including confounded ones
- A model trained to predict loan default will use neighbourhood as a feature if it predicts default, even if neighbourhood is a proxy for a protected attribute
- Distribution shift breaks correlational models: if the confounder distribution changes (new population, new policy), the model fails
- Causal models are more robust to distribution shift

OOD generalisation is a causal problem, not just a statistical one

Simpson's Paradox

Aggregate trend reverses when stratified by a third variable

- Overall: treatment group has worse outcomes than control
- Stratified by severity: treatment group has better outcomes in *both* mild and severe cases
- Why? Severe cases are more likely to receive treatment (selection bias)
- Resolution: condition on the confounder (severity)



Simpson's paradox: always ask whether a third variable explains the aggregate trend

Randomized Controlled Trials

The gold standard: randomisation breaks confounding

- Randomly assign units to treatment and control
- Randomisation ensures $X \perp\!\!\!\perp Z$ for all confounders Z
- Result: observed correlation $P(Y|X)$ equals causal effect $P(Y|do(X))$
- Limitation: expensive, slow, sometimes unethical, often impossible in finance

A/B testing is an RCT; the only difference is the industry vocabulary

Observational Methods

Causal inference from data you did not randomise

- **Instrumental variables:** find a variable Z that causes X but affects Y only through X; use Z to isolate the causal $X \to Y$ channel
- **Regression discontinuity:** units just above and below a threshold are nearly random; use the discontinuity as a natural experiment
- **Propensity score matching:** match treated and control units with similar probability of receiving treatment; controls for observed confounders

Each method requires different untestable assumptions; document them explicitly

Counterfactual Reasoning

"Would this applicant have defaulted if we had approved them?"

- The fundamental problem of causal inference: we cannot observe both potential outcomes for the same unit
- Individual treatment effect: $\tau_i = Y_i(1) - Y_i(0)$; only one is observed
- Causal forests (Wager & Athey 2018): estimate heterogeneous treatment effects via random forests
- Double ML (Chernozhukov et al. 2018): orthogonalise to remove nuisance bias
- Tools: EconML, DoWhy, CausalML

The counterfactual you cannot observe is the core challenge of causal inference

The Ladder of Causation

Pearl's three rungs of increasingly powerful reasoning

Rung 1: Association

- See and predict
- $P(Y|X)$
- Standard ML lives here

Rung 2: Intervention

- Do and act
- $P(Y|do(X))$
- RCTs, A/B tests

Rung 3: Counterfactuals

- Imagine and explain
- $P(Y_x|X = x')$
- What-if reasoning

Most regulators require Rung 2 reasoning; standard ML only provides Rung 1

Limitations

Causal inference is not a free lunch

- Cannot assume away unobserved confounders: they must be measured or the identifying assumptions must be stated and justified
- Identifiability: causal effects can only be estimated if the causal structure is partially known (the DAG must be specified or estimated)
- Causal discovery from purely observational data is hard and identifiable only up to Markov equivalence class
- Sample sizes required for heterogeneous treatment effects are often larger than available datasets

Every causal claim rests on assumptions; state them or stop making the claim

Code Pointer

DoWhy and EconML for causal estimation

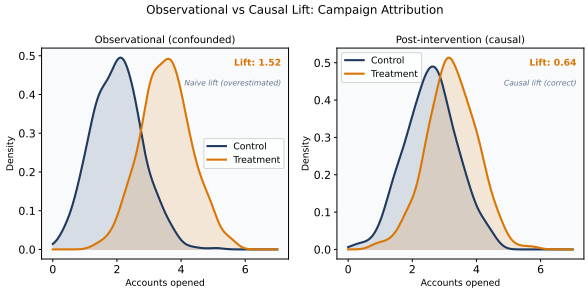
- `pip install dowhy econml`
- DoWhy: define a causal graph, identify, estimate, refute
- EconML: `CausalForestDML` for heterogeneous effects
- `from econml.dml import CausalForestDML`
- Fit on observational data; estimate CATE (conditional average treatment effect)

DoWhy's refute step runs sensitivity checks automatically: a major advantage

Finance Application: Campaign Attribution

Does a marketing campaign cause new account opens, or is it seasonal?

- Observed correlation: customers who receive the campaign open more accounts
- Confounder: campaign targets high-engagement customers who would open accounts anyway
- Intervention: $P(\text{opens}|do(\text{campaign}))$ requires controlling for engagement
- Method: propensity score matching on engagement features before computing lift



Naive lift = correlation; causal lift = $P(Y|do(X=1)) - P(Y|do(X=0))$

Why It Matters for Deployment

Acting on correlations in production can cause feedback loops

- Recommender systems trained on clicks amplify existing preferences; adding causality breaks the filter bubble
- Credit models using neighbourhood as proxy may violate fair lending law if neighbourhood is a proxy for a protected characteristic
- Policy interventions based on correlational models often fail in the field (the streetlight effect: measuring what is easy, not what causes the outcome)

Deployment = intervention; correlation-based models were not built for this

Practical Takeaway

High accuracy on observational data does not justify causal claims

- Before deploying: ask "will the input distribution change when we act?"
- If yes: you need causal reasoning, not just predictive accuracy
- Start with DAGs: drawing the assumed causal graph forces hidden assumptions into the open
- Use A/B tests when possible; use instrumental variables or RD when they are not

Rule: every policy recommendation needs a causal story, not just a prediction

Summary and Open Questions

Three things to remember

- Correlation $\neq$ causation; ML models are correlational and can fail when deployed as interventions
- DAGs make assumptions explicit; the back-door criterion recovers causal effects from observational data when all confounders are measured
- Finance: always distinguish predictive lift from causal lift; propensity matching or RCT required for the latter

Open questions:

1. When can ML safely make causal claims?
2. How do we test causal assumptions in practice?
3. Can foundation models perform counterfactual reasoning?

Further reading: Pearl (2018) The Book of Why; Imbens & Rubin (2015) Causal Inference