PH 724 Data Analysis & Scientific Communication Incorporating Use of Generative AI

Part 11: From Association to Causation

Zhiyang Zhou (zhou67@uwm.edu, zhiyanggeezhou.github.io)

2026/11/16 14:05:40

Association does not imply causation

Our scientific question is:

Does flower treatment affect the probability of detectable infection at the end of the experiment?

The word **affect** suggests a causal question.

A statistically significant treatment coefficient does not, by itself, establish causation.

Does a nonsignificant treatment effect mean no causation?

A nonsignificant result means that we do not have sufficient statistical evidence to reject the null hypothesis.

It does NOT establish that the causal effect is zero.

Possible explanations:

- The true causal effect is zero.
- The effect exists but is small.
- The study has limited statistical power.
- The outcome is highly variable: there is substantial variation in outcomes among experimental units, making it difficult to distinguish a treatment effect from random variation.

What do we mean by a causal effect?

Consider two possible interventions:

- Treatment: a person exercises regularly.
- Control: the same person does not exercise regularly.

A causal question asks:

What would happen to the same person under these two different conditions?

This is fundamentally different from comparing two people who happen to have different exercise habits.

The difficulty is that we cannot observe both outcomes for the same person at the same time.

Randomized controlled trial: the ideal starting point

Suppose we randomly assign participants to the treatment/control group. Because treatment is randomly assigned, the two groups are comparable **in expectation** with respect to pre-treatment characteristics. That

is, if randomization were repeated many times, the two treatment groups would have the same average pre-treatment characteristics, on average across those randomizations.

Under appropriate assumptions, differences in outcomes between the two groups can be attributed to treatment assignment.

This is why randomized controlled trial (RCT) is often regarded as the gold standard for causal inference.

Randomized controlled trial vs. observational study

Suppose that health consciousness affects both:

- Whether a person exercises.
- Their cardiovascular disease risk.

In an observational study:

- Exercise habits may be associated with health consciousness.
- The association between exercise and disease may therefore reflect both exercise and health consciousness.

In an RCT:

- Assignment to the exercise intervention is independent of baseline health consciousness, in expectation.
- Health consciousness is therefore not a systematic confounder of the effect of randomized assignment.

RCT breaks the systematic link between treatment assignment and pre-treatment confounders. However, many important scientific questions cannot be answered using RCTs.

Examples:

- Does long-term air pollution exposure cause lung disease?
- Does smoking cause lung cancer?
- Does education increase lifetime earnings?
- Does neighborhood poverty affect health?

Randomization may be:

- Unethical.
- Impractical.
- Too expensive.
- Impossible for the exposure of interest.

Researchers must often rely on observational studies. The challenge is to determine when observational data can identify a causal effect and support a causal conclusion.

Counterfactual framework

Let A indicate treatment for a randomly selected individual:

- $A = 1$: receives treatment.
- $A = 0$: receives control.

Define two potential outcomes for this individual:

- $Y(1)$: outcome if this individual receives treatment.
- $Y(0)$: outcome if this individual receives control.

The individual causal effect is

$$Y(1) - Y(0).$$

However, we can observe only one of the two potential outcomes for an individual. Under the **consistency assumption**, the observed outcome Y satisfies:

$$Y = A \cdot Y(1) + (1 - A) \cdot Y(0).$$

Instead, we can study the **average treatment effect (ATE)**:

$$\text{ATE} = \text{E}\{Y(1)\} - \text{E}\{Y(0)\}.$$

Importantly, $\text{E}\{Y(a)\}$ is not necessarily the same as $\text{E}(Y \mid A = a)$.

- $\text{E}\{Y(a)\}$: mean outcome if everyone received treatment a.
- $\text{E}(Y \mid A = a)$: mean observed outcome among individuals who actually received treatment a.

The central causal inference question is:

$$\underbrace{\text{E}\{Y(1)\} - \text{E}\{Y(0)\}}_{\text{ATE}} \quad \stackrel{?}{=} \quad \text{E}(Y \mid A = 1) - \text{E}(Y \mid A = 0).$$

When does the above equality hold? We need following assumptions:

- Exchangeability: Treatment assignment is independent of potential outcomes.
- Consistency: The observed outcome corresponds to the treatment actually received.
- Positivity: Each treatment has a positive probability of being received.

When can we identify the ATE?

A central assumption is **exchangeability**. In an RCT, treatment assignment is independent of potential outcomes:

$$\{Y(1), Y(0)\} \perp A.$$

Together with consistency and positivity ($0 < \Pr(A = 1) < 1$) assumptions, exchangeability implies

$$\text{E}\{Y(a)\} = \text{E}(Y \mid A = a).$$

Therefore,

$$\text{ATE} = \text{E}(Y \mid A = 1) - \text{E}(Y \mid A = 0).$$

That is, the difference in population means among the treatment groups identifies the ATE.

What changes in observational studies?

In observational studies, unconditional exchangeability usually does not hold:

$$\{Y(1), Y(0)\} \not\perp A.$$

However, Conditional exchangeability may hold given certain set of pre-treatment confounders X:

$$\{Y(1), Y(0)\} \perp A \mid X.$$

Under consistency, conditional positivity ($0 < \Pr(A = 1 \mid X = x) < 1$), and conditional exchangeability,

$$\text{E}\{Y(a)\} = \text{E}_X\{\text{E}(Y \mid A = a, X)\}.$$

This is the **g-formula**. It provides a way to identify the ATE from observational data under explicit assumptions.

How does regression contribute to causal inference?

Suppose the outcome is binary. We might fit a logistic regression:

$$\text{logit}\{\Pr(Y = 1 \mid A, X)\} = \beta_0 + \beta_1 A + \beta_2 X.$$

For causal inference, we may use the fitted model to predict outcomes under both treatments:

$$\hat{p}_i(a) = \widehat{\Pr}(Y = 1 \mid A = 1, X = x_i), \quad a = 0, 1.$$

Then estimate the ATE by averaging the predicted differences:

$$\widehat{\text{ATE}} = \frac{1}{n} \sum_{i=1}^{n} \{\hat{p}_i(1) - \hat{p}_i(0)\}.$$

This approach is called **regression standardization** (or **parametric g-computation**). It approximates the outer expectation of the g-formula by averaging predictions across all n data points.

Importantly, regression does not guarantee conditional exchangeability, and valid estimation also requires an appropriate model for outcomes.

Causal inference is more than regression

Modern causal inference provides tools for:

- **Representing causal relationships and assumptions:** Directed acyclic graphs (DAGs) and structural equation models.
- **Identifying confounders:** DAGs and the backdoor criterion.
- **Addressing confounding:** Propensity score methods and instrumental variables.
- **Addressing selection bias:** Inverse probability of censoring weighting.
- **Estimating causal effects:** g-computation and propensity score weighting.
- **Assessing robustness:** Sensitivity analysis.

Statistical models describe relationships in observed data.

Causal inference establishes when and how those relationships can be used to identify and estimate causal effects. **This identification depends on the study design and the assumptions required by the chosen causal inference method, not simply on the choice of statistical model.**