Bias

POLS 602
Fall 2025

Dr. Mike Burnham
Texas A&M Political Science

Test notes

- Be precise in your language.
  - U_i = B_ip+C_i is not about whether or not someone will vote. It is strictly a statement about expected utility
  - “relationship” is not a precise word. Correlation. causal effect. predicted value of Y with a change in X.
  - estimate vs. estimator vs. estimand
  - “benefit someone gets depends on the probability that they cast the deciding vote”
    - Not accurate. It is actually the difference between the benefit they would get if they did vote, vs the benefit they would get if they didn’t vote.
  - Some reliance on previous knowledge without a deep understanding of the relationships between everything.

Gauss Markov Theorem

Gauss-Markov Theorem

Under the following assumptions, OLS is the best linear unbiased estimator (BLUE):

Best = lowest sampling variance among all linear unbiased estimators

1. Linearity of parameters
2. Independence (no autocorrelation, or uncorrelated errors)
3. Homoscedasticity
4. No Perfect multicolinearity
5. Zero conditional mean (exogeneity)

Linearity of parameters

- The population models parameters are linear. The change in Y associated with a change in X is constant

- This does not imply that independent variables must be linear or that linear regression can only model linear phenomena

    - linear: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

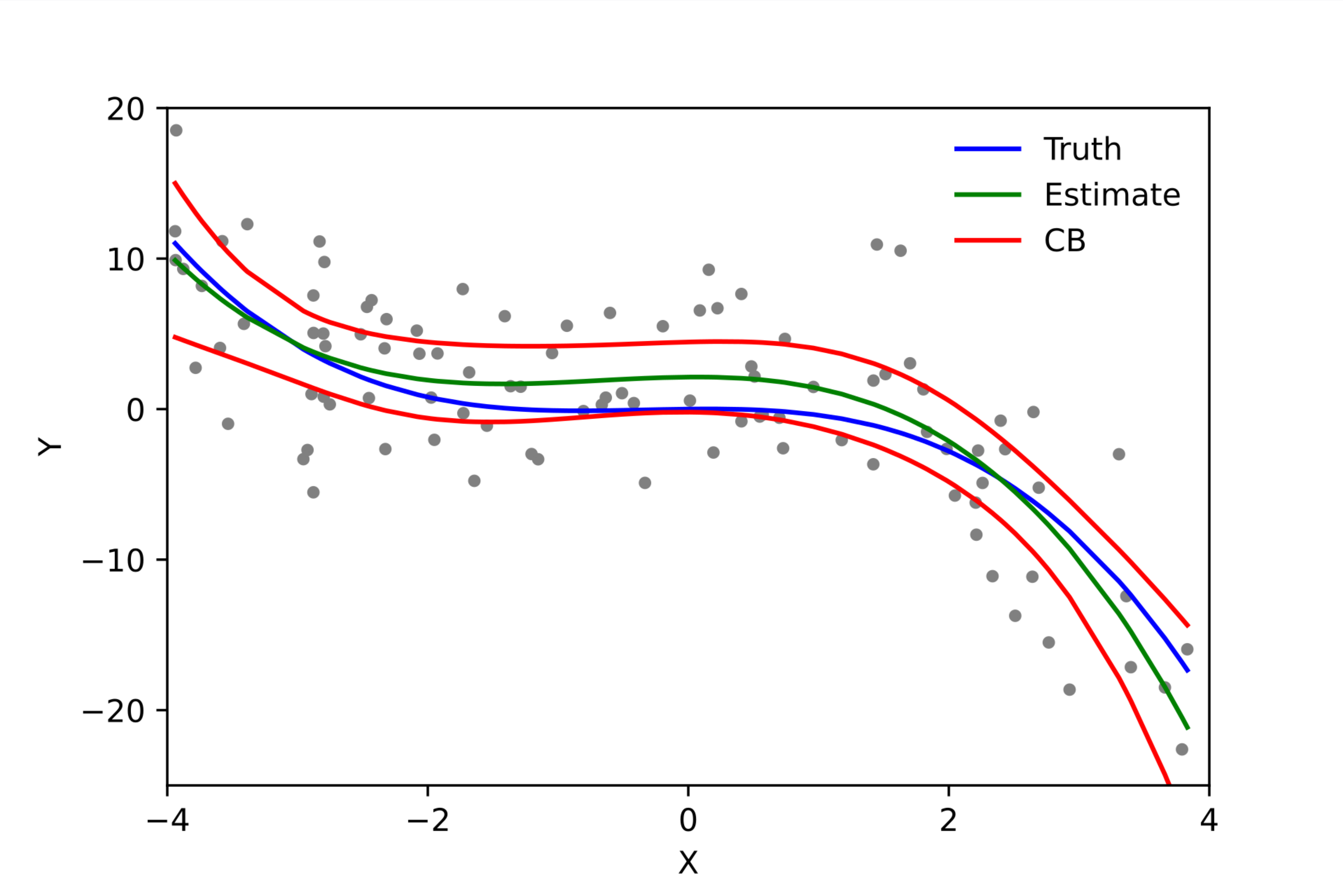
    - linear: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$

    - non-linear: $y_i = \beta_0 + \beta_1^{\beta_2 x_i} + \epsilon_i$

Linearity of parameters

- The population models parameters are linear. The change in Y associated with a change in X is constant
- This does not imply that independent variables must be linear or that linear regression can only model linear phenomena
  - linear: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
  - linear: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$
  - non-linear: $y_i = \beta_0 + \beta_1^{\beta_2 x_i} + \epsilon_i$

Linear

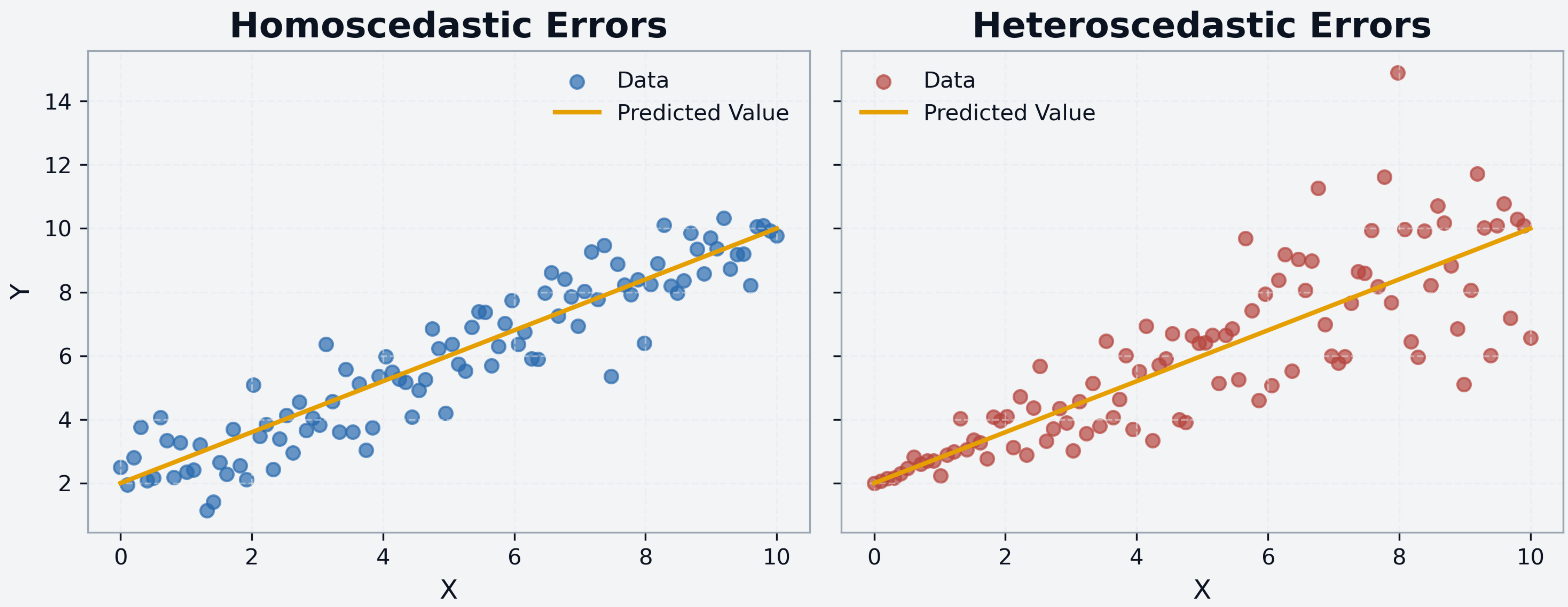


Independence

- The errors of any two observations are not correlated with each other
- The error for one data point provides no information about the error for another data point
- $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$

Homoscedasticity

- The variance of the error term is constant for all observations



No perfect multicollinearity

- No independent variable is a perfect linear function of any other independent variable
- If two variables are perfectly collinear, it is impossible to isolate the effect of the individual variables

Zero conditional mean (exogeneity)

- The expected value of the residuals is zero
- $\mathbb{E}(\epsilon_i | X) = 0$
- $Cov(X, \epsilon) = 0$

Bias

Validity

Internal Validity: The degree to which your study supports the claims being made about the studied population.

External Validity: The extend to which results can be extrapolated to other populations or contexts.

Bias

The average estimation error

$\mathbb{E}(\text{estimation error})$

$\mathbb{E}(\text{estimated value} - \text{true value})$

Omitted variable bias

- Bias induced by excluding one or more relevant variables

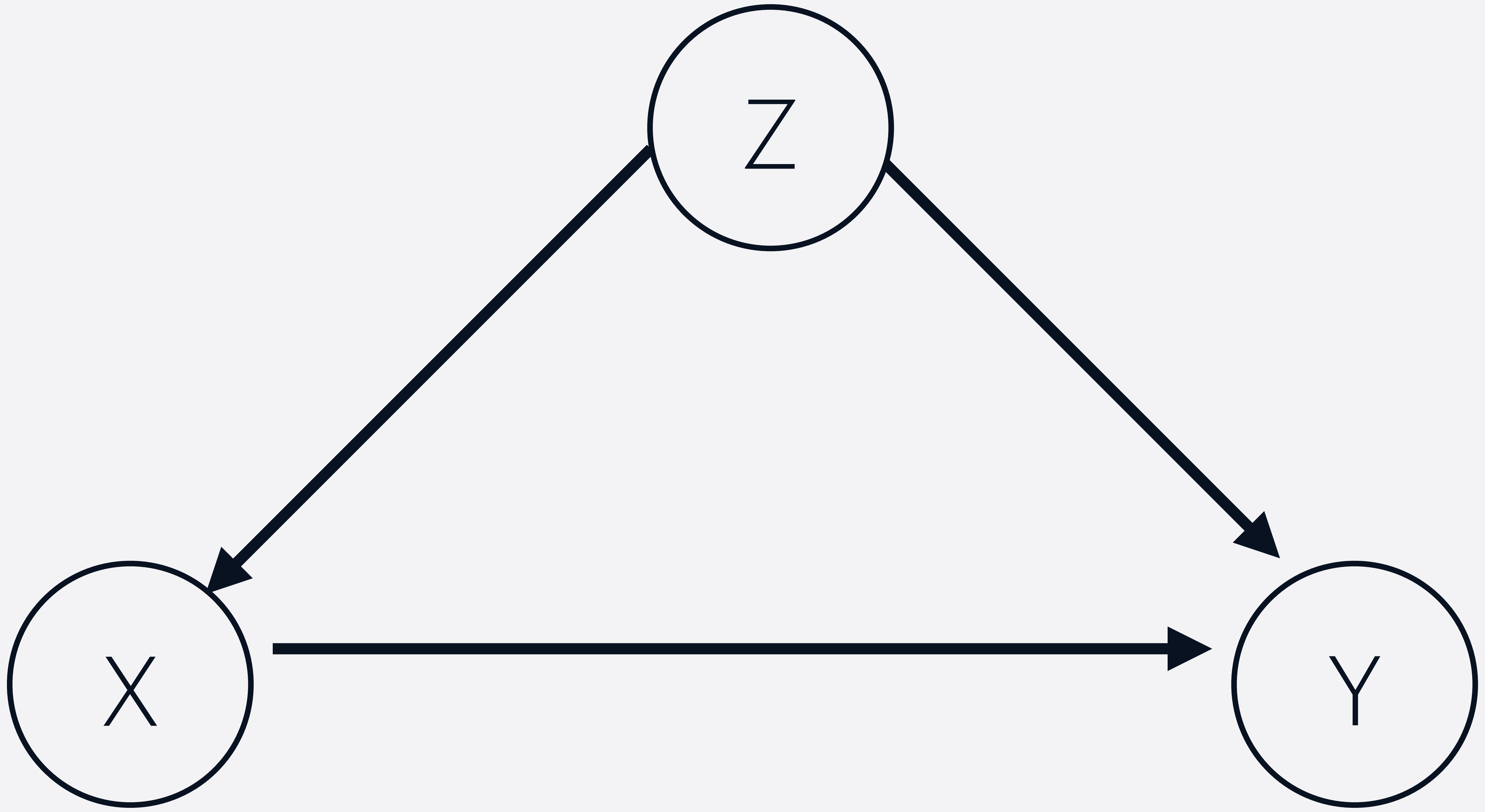
True DGP: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \epsilon_i$

Theorized model: $y_i = \beta_0 + \beta_1 x_i + u_i$

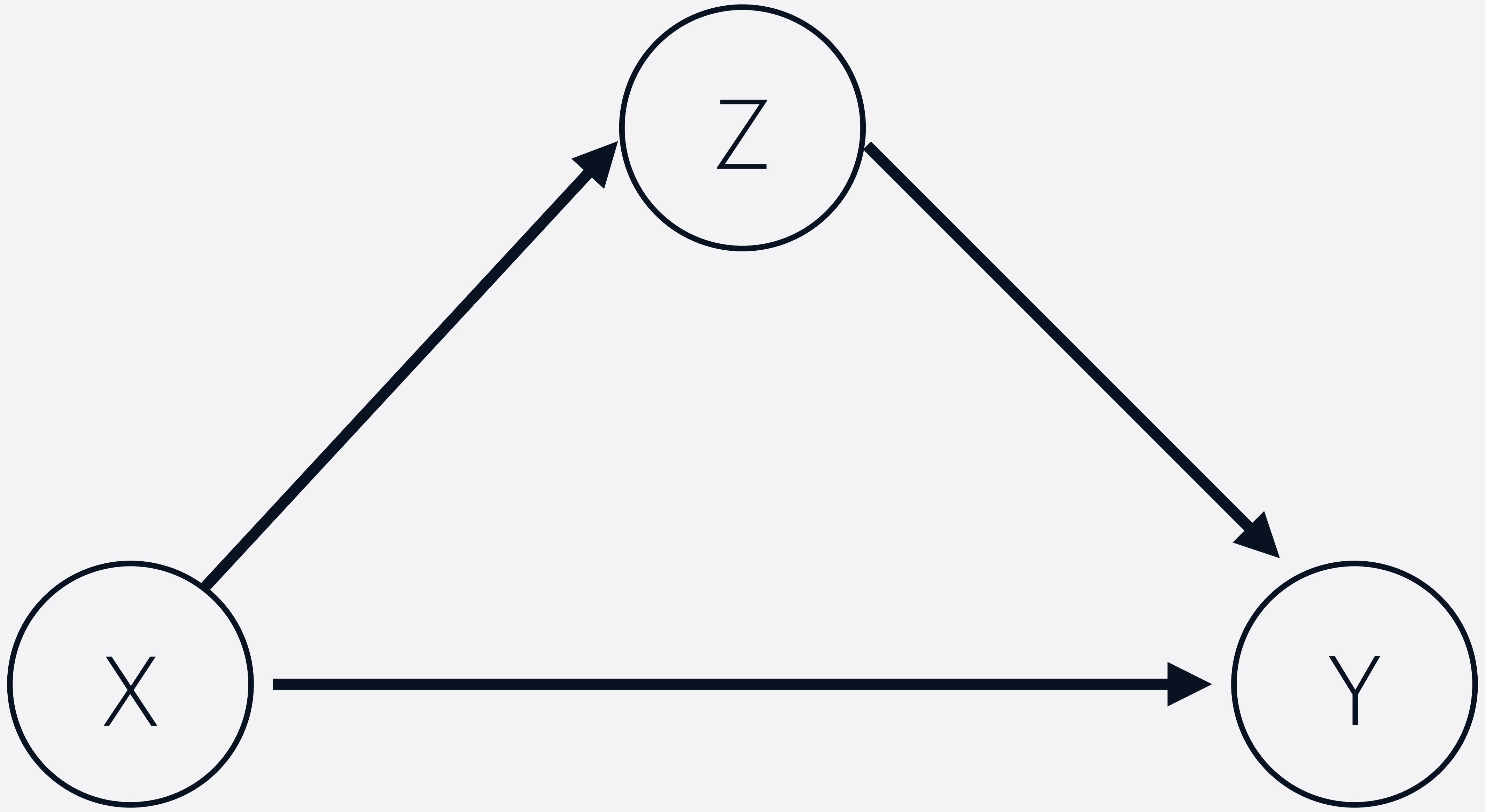
Thus: $u_i = \beta_2 z_i + \epsilon_i$

If $COR(X, Z) \neq 0$, then $COR(X, u) \neq 0$

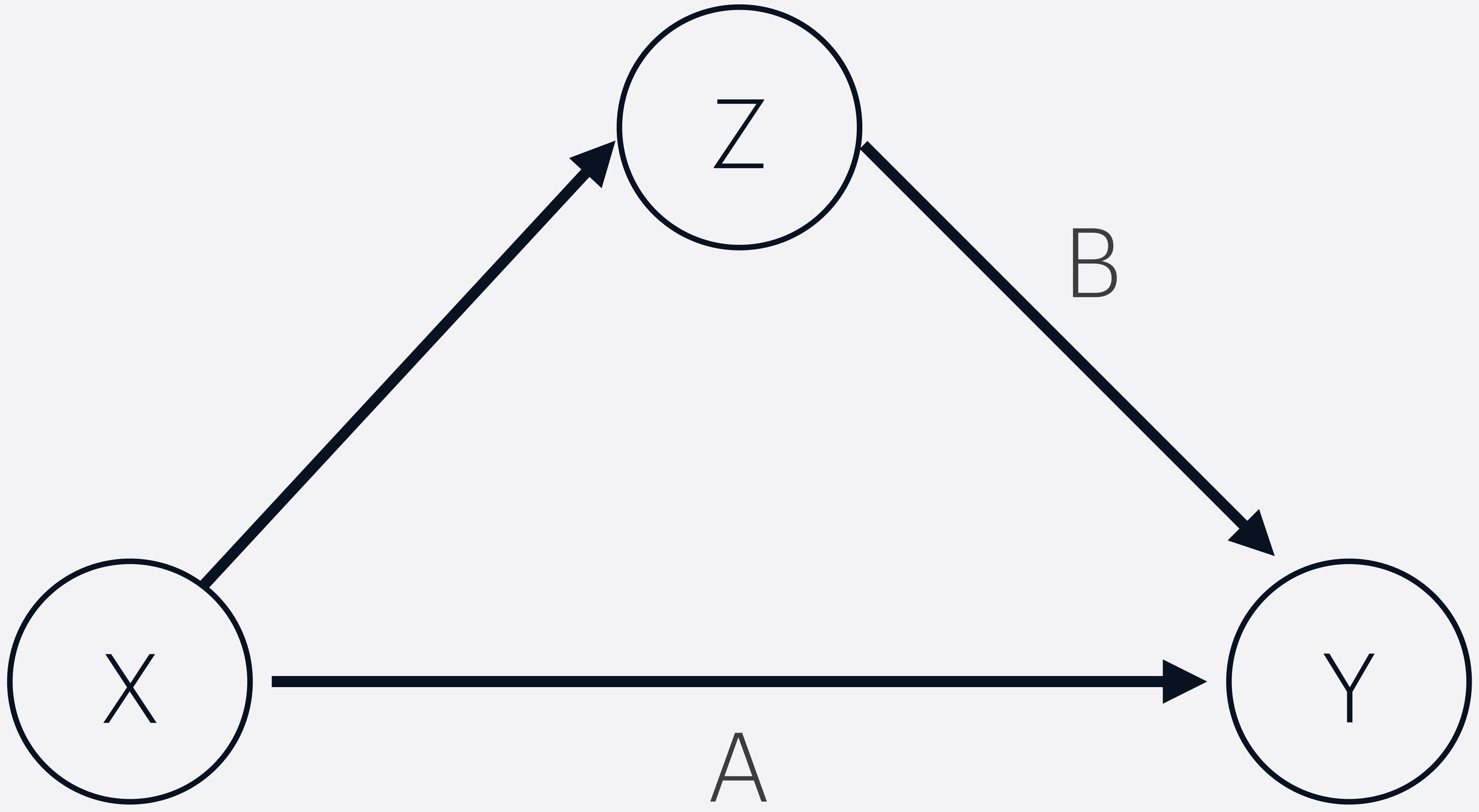
Omitted variable bias: Confounders



Omitted variable bias: Mediators



Omitted variable bias: Mediators

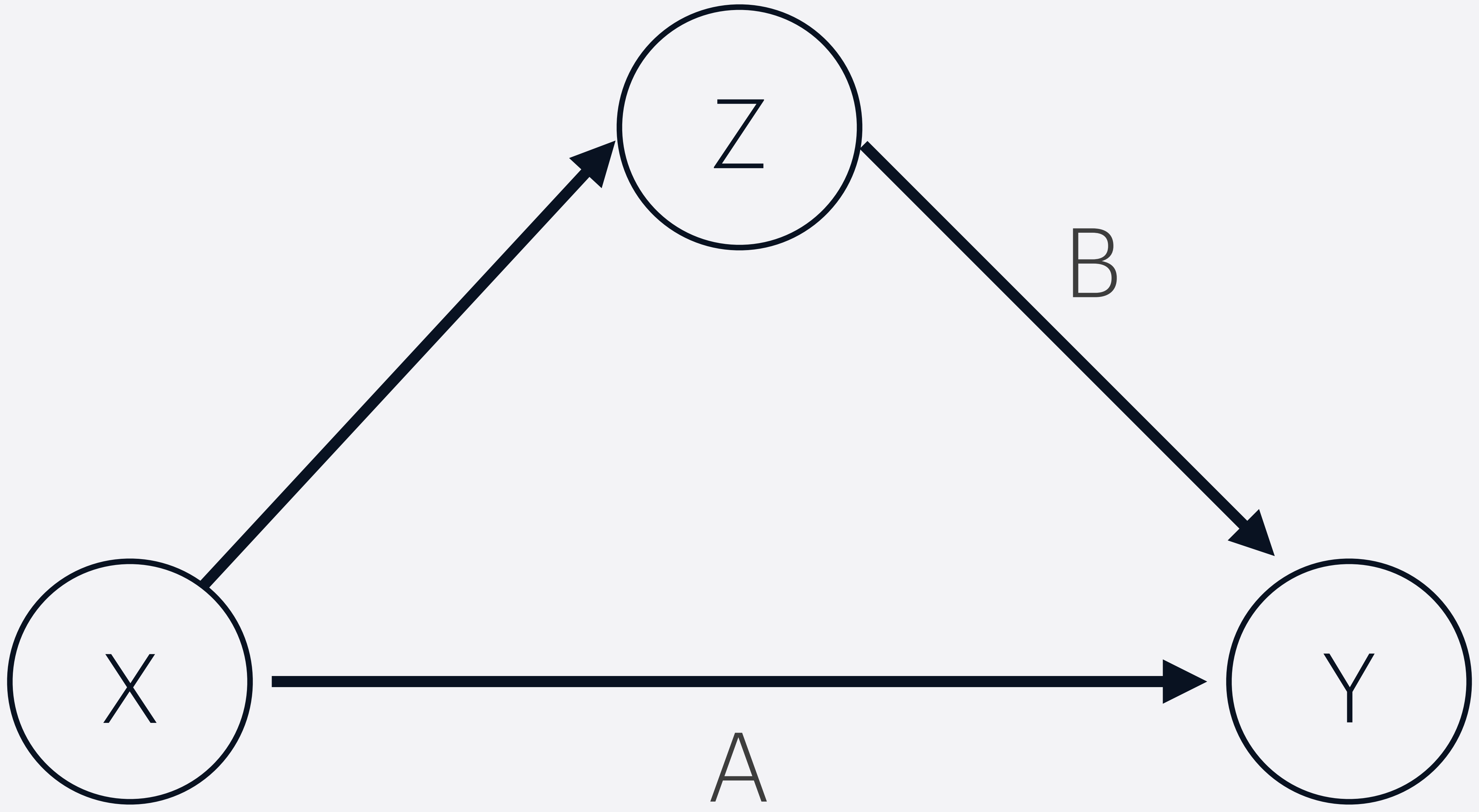


A: Direct effect

B: Indirect effect

A + B: Total effect

Omitted variable bias: Mediators



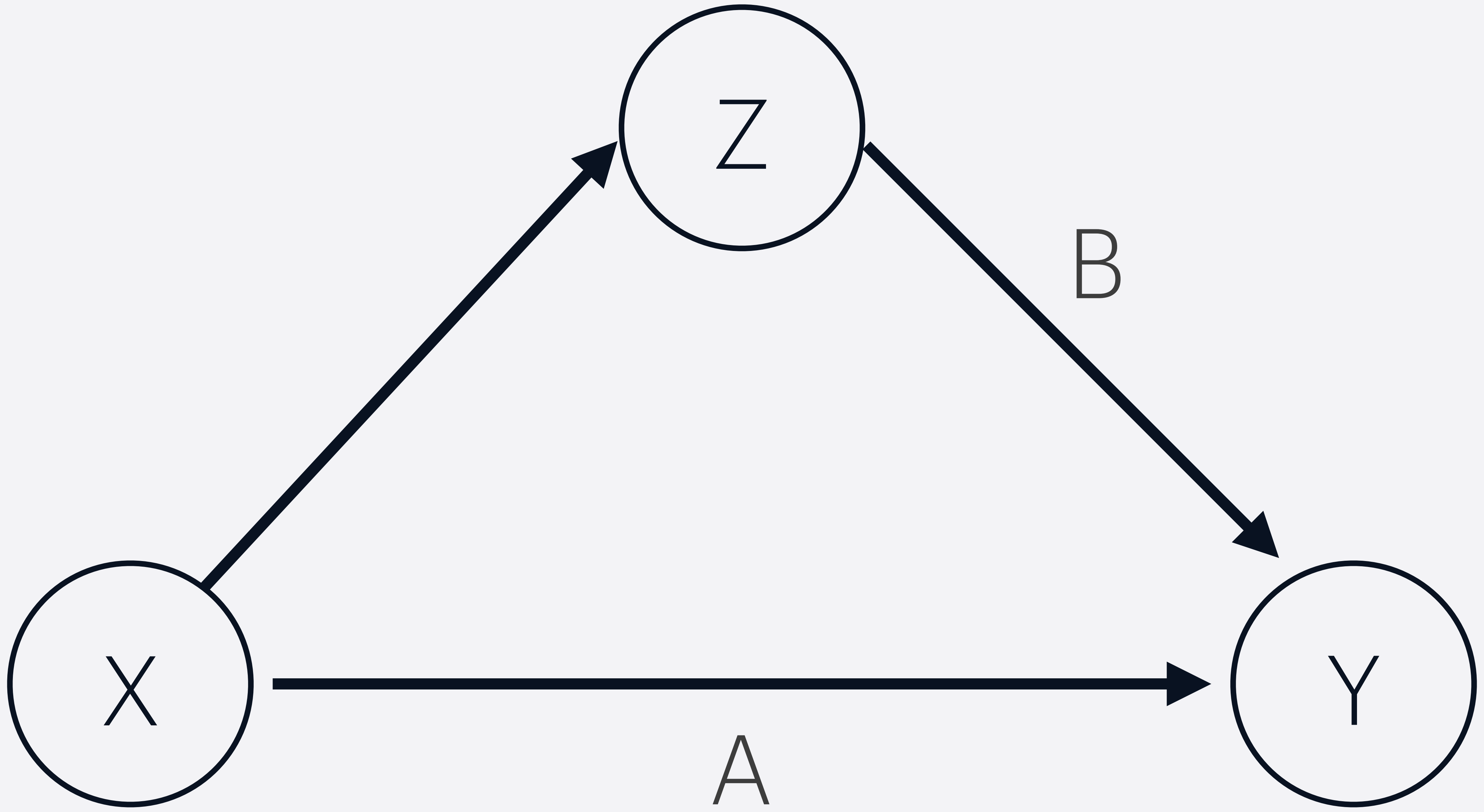
A: Direct effect

B: Indirect effect

A + B: Total effect

Whether you control for Z depends on if you want to estimate the direct effect, or the total effect.

Omitted variable bias: Mediators



A: Direct effect

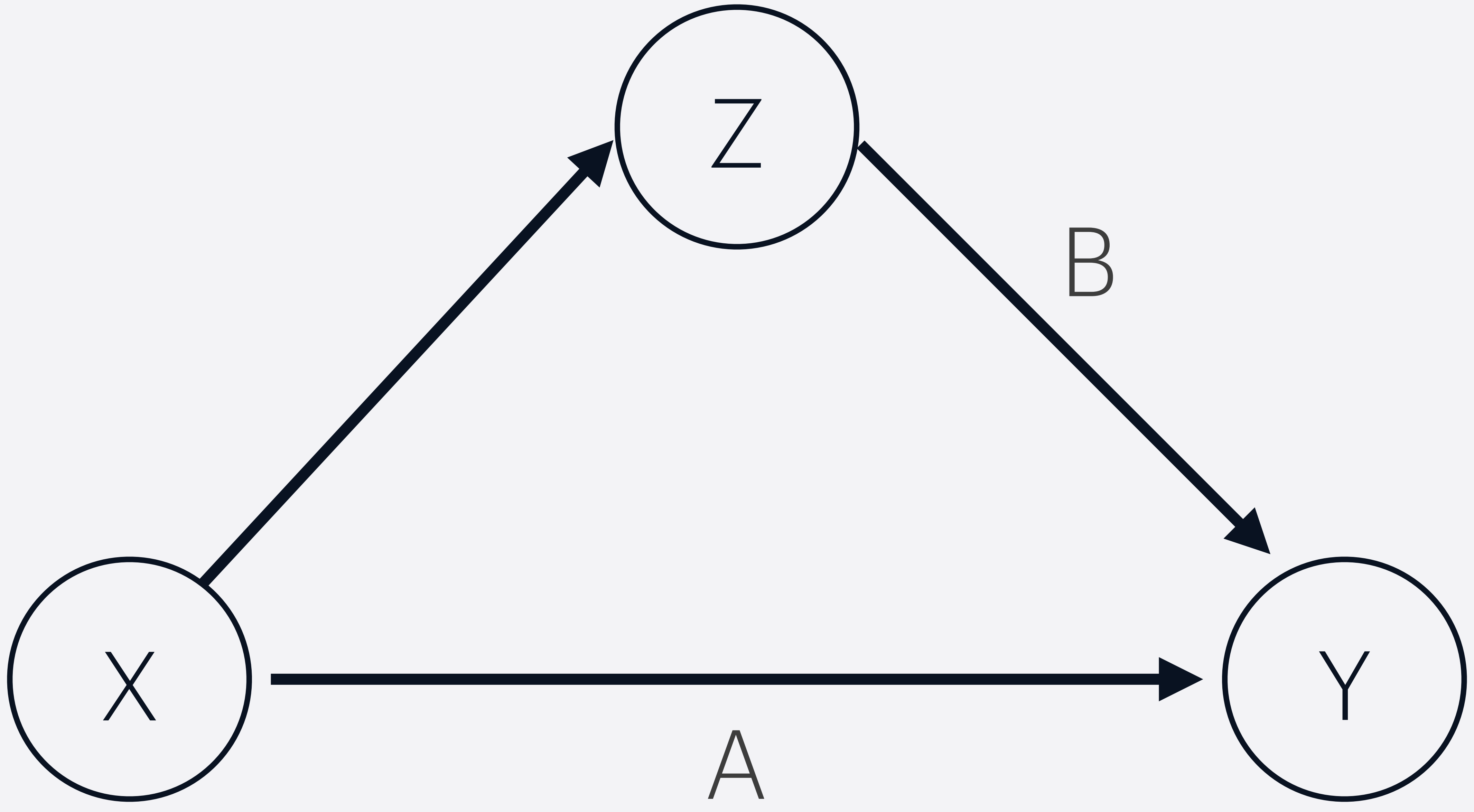
B: Indirect effect

A + B: Total effect

Whether you control for Z depends on if you want to estimate the direct effect, or the total effect.

What is your estimand?

Post treatment bias

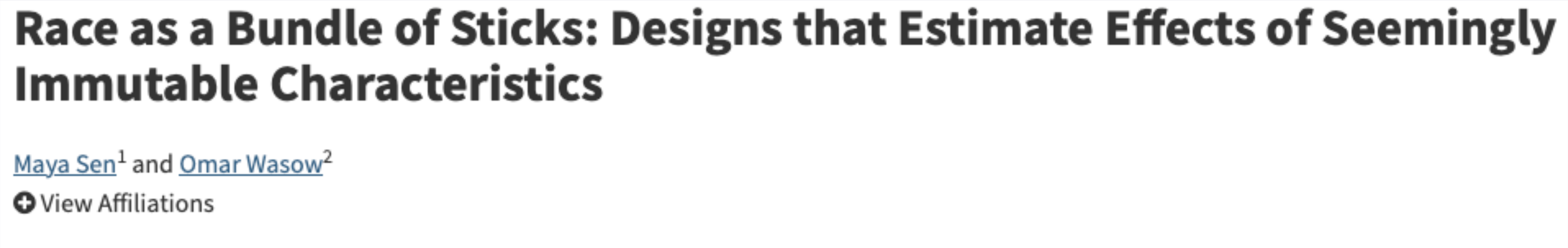


Controlling for a variable that is a consequence of, or affected by, the treatment when you want to estimate the total, rather than the direct effect.

Controlling for Z, when you want to estimate A + B.

Post treatment bias: examples

- Controlling for lung cancer when trying to estimate the effect of smoking on mortality
- Controlling for education or income when either is your treatment
- Controlling for… almost anything when trying to estimate the effect of race



Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics

Maya Sen[1] and Omar Wasow[2]

⊕ View Affiliations

Measurement Error

Suppose our true independent variable is x_i.

But we observe $x_i * = x_i + e_i$, where e_i is some random measurement error. Thus:

$x_i = x_i * -e_i$

$y_i = \beta_0 + \beta_1 x_i * + \epsilon_i$

$y_i = \beta_0 + \beta_1(x_i - e_i) + \epsilon_i$

$y_i = \beta_0 + \beta_1 x_i - \beta_1 e_i + \epsilon_i$

$y_i = \beta_0 + \beta_1 x_i + (\epsilon_i - \beta_1 e_i)$

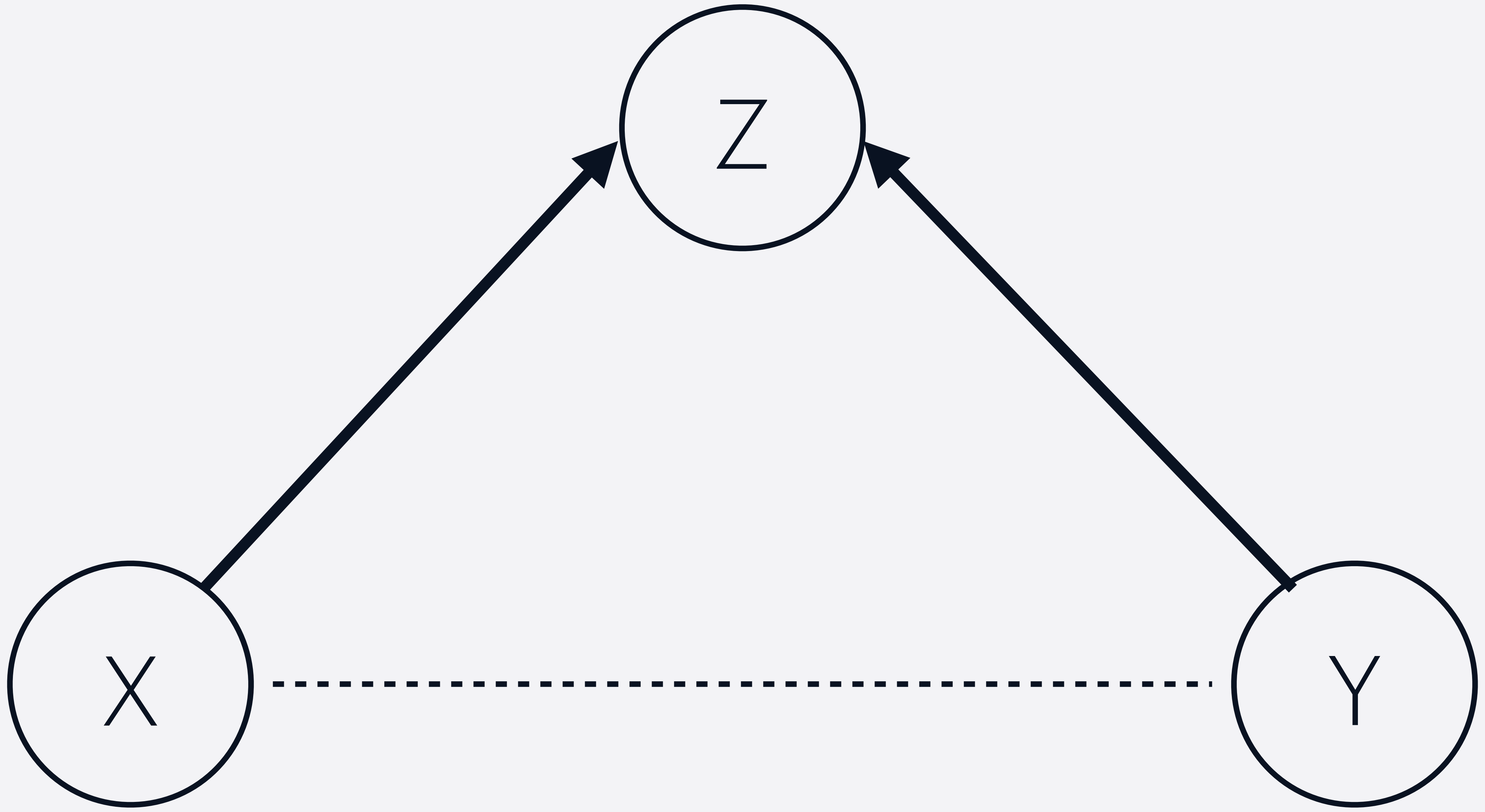
$y_i = \beta_0 + \beta_1 x_i + u_i$ where $u_i = \epsilon_i - \beta_1 e_i$ and $x_i = x_i * -e_i$

$COR(X, u) \neq 0$ because X and u are both a function of e

Simultaneity

A two-way causal relationship between the dependent and independent variable. Can be thought of as a special case of omitted variable bias.

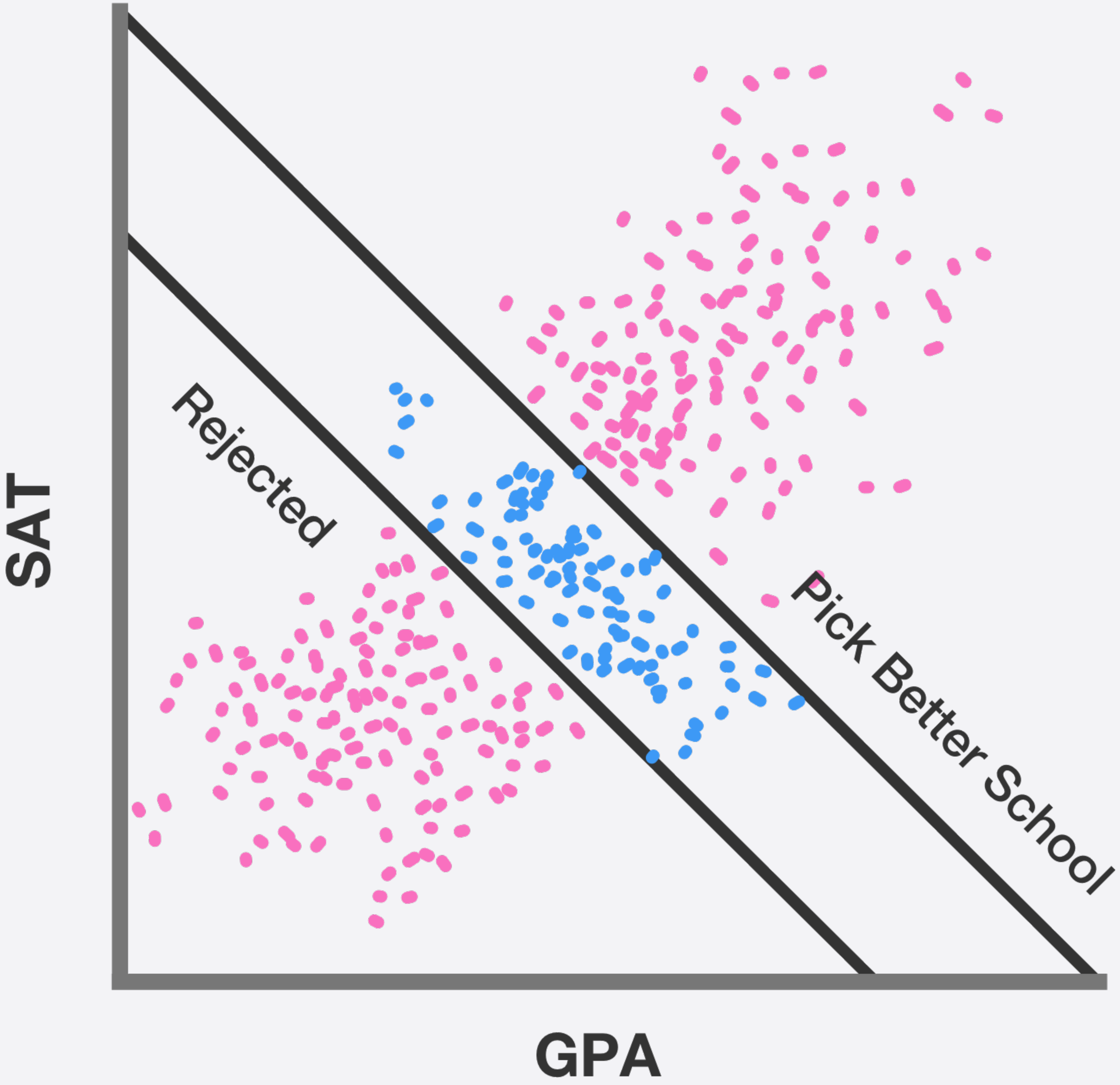
Colliders



Collider: A variable that is causally influenced by two or more variables

Collider bias: Conditioning on a collider via regression, sampling, or treatment application

Collider Bias (Berkson's Paradox)



Selection Bias

Systematic error due to study participants or data not being representative of the target population. Examples:

- Sampling bias
- Survivorship bias
- Nonresponse bias

Selection bias is often equivalent to conditioning on a collider