

A First Course in Causal Inference: Chapters 14–16

Peng Ding (2024)

Lecture Notes

Sweta Karlekar

November 2023

Chapter 14: Using Propensity Scores in Regressions for Causal Effects

Recap

Recall from Chapter 11 the propensity score defined with potential outcomes:

$$e(X, Y(1), Y(0)) = \Pr\{Z = 1 \mid X, Y(1), Y(0)\}.$$

Under *strong ignorability*,

$$\Pr\{Z = 1 \mid X, Y(1), Y(0)\} = \Pr(Z = 1 \mid X).$$

Thus the propensity score reduces to

$$e(X) = \Pr(Z = 1 \mid X).$$

Uses:

- Dimension reduction tool.
- Stratify on propensity score.
- Propensity-score weighting.
- Covariate balance check.

Overview

1. Include propensity scores as covariates in regression.
2. Run regressions weighted by the inverse of the propensity score.

Propensity scores as covariates

Motivating example:

Effect of an optional tutoring program on end-of-year test scores

This is an observational study: students who opt in might systematically differ from students who do not.

X : age, gender, grade, socioeconomic status, initial test score,

Z : tutoring (0 or 1),

Y : final test score.

We want to control for students' propensity to enroll in the tutoring program.

Step 1: Create propensity scores

$$e(X) = \Pr(Z = 1 \mid X).$$

Step 2: Regress on treatment, controlling for propensity score

$$Y = \beta_0 + \beta_z Z + \beta_e e(X).$$

OLS:

$$(\hat{\beta}_0, \hat{\beta}_z, \hat{\beta}_e) = \arg \min_{\beta_0, \beta_z, \beta_e} \mathbb{E} \left[\left(Y - \beta_0 - \beta_z Z - \beta_e e(X) \right)^2 \right].$$

$$\hat{\tau} = \hat{\beta}_z.$$

Chapter 15: Matching in Observational Studies

Matching in an observational study is analogous to matching in experimental studies (Chapter 7).

Overview

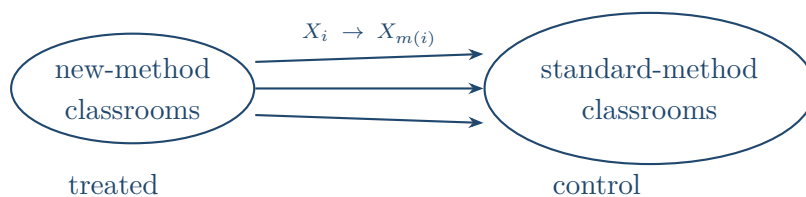
- Motivating example.
- Practical considerations.
- Define the matching estimator of average causal effect $\hat{\tau}^m$.
- Bias estimator; matching estimator for average causal effect $\hat{\tau}^{mbc}$.
- Variance estimator.
- Connection to the doubly robust estimator.
- Repeat for the average causal effect on treated units: $\hat{\tau}_T^m$ and $\hat{\tau}_T^{mbc}$.

Matching estimator for the average causal effect and for the treated units.

Motivating example

New teaching method $\rightarrow$ test scores.

We roll this out to a small number of classrooms in a district.



What do we know about these classrooms?

X : class size, student demographics, teacher experience, previous test scores, etc.

Ideal scenario: exact match on all covariates, $X_i = X_{m(i)}$.

We can then analyze the results as a *matched-pair experiment* (MPE) — extreme stratification (Chapter 7).

$\rightarrow$ Fisher randomization test or Neymanian approach.

But what about real-world settings?

$$X_i \approx X_{m(i)}, \quad m(i) = \arg \min_{k: Z_k=0} d(X_i, X_k),$$

where d could be Euclidean distance, Mahalanobis distance, etc.

Practical considerations

1. Sampling strategies

With replacement	Without replacement
Computationally convenient.	Discrete optimization.
Matches of higher quality.	Potentially matches of lower quality.
Reusing the same matched unit introduces dependence.	Matched units are more independent.
More complex data analysis (weighted analysis).	More straightforward paired t -test or regression analysis.

If the control pool is large, either method may give similar results.

With replacement: assign weights according to how many times a control unit is used, then use weighted t -tests or regression.

2. Covariate adjustment during data analysis

Residual covariate imbalance remains even after matching.

Earlier (Section 7.4), covariate adjustment was used to improve estimation efficiency. Here, the goal is to adjust for *bias*.

3. What if we observe a ton of covariates?

A treated unit i may be far from all control units in covariate space, especially if X is high-dimensional.

$$X_i \text{ far from } X_{m(i)}.$$

We may have to drop some treated units, which changes the population of interest.

One solution: match on the propensity score instead.

Matching estimator

$$\hat{\tau}^{\text{m}} \longrightarrow \text{matching estimator.}$$

Let's derive it. Let $\mathcal{J}_i$ contain the M matched units for unit i .

For a treated unit i : $Z_i = 1$

$$\underbrace{\hat{Y}_i(0)}_{\text{potential outcome estimate}} = \frac{1}{M} \sum_{k \in \mathcal{J}_i} Y_k.$$

For a control unit i : $Z_i = 0$

$$\hat{Y}_i(1) = \frac{1}{M} \sum_{k \in \mathcal{J}_i} Y_k.$$

Combining observed and matched potential outcomes,

$$\hat{\tau}^{\text{m}} = \frac{1}{n} \sum_{i=1}^n (\hat{Y}_i(1) - \hat{Y}_i(0)).$$

What if these estimates are poor because covariates of the matches do not agree?

Abadie & Imbens proposed a bias estimator. Subtract the estimated bias from the matching estimator:

$$\hat{\tau}^{\text{mbc}} = \hat{\tau}^{\text{m}} - \hat{B}.$$

We need to construct $\hat{B}$.

Bias

Let $\{\hat{\mu}_1(X_i), \hat{\mu}_0(X_i)\}$ be predicted potential outcomes (e.g., fitted by OLS).

For a treated unit i with $Z_i = 1$:

$$\hat{B}_i = \frac{1}{M} \sum_{k \in \mathcal{J}_i} \underbrace{\left(\hat{\mu}_0(X_i) - \hat{\mu}_0(X_k) \right)}_{\text{predicted control outcomes at different covariates}}.$$

The difference measures the discrepancy due to imperfect matching on the covariates.

For a control unit i with $Z_i = 0$:

$$\hat{B}_i = -\frac{1}{M} \sum_{k \in \mathcal{J}_i} \left(\hat{\mu}_1(X_i) - \hat{\mu}_1(X_k) \right).$$

The signs differ because the missing potential outcome differs between treated and control units.

There is some difference we were previously “docking” from the treatment effect.

Putting the two cases together:

$$\hat{B}_i = \frac{2Z_i - 1}{M} \sum_{k \in \mathcal{J}_i} \left\{ \hat{\mu}_{1-Z_i}(X_i) - \hat{\mu}_{1-Z_i}(X_k) \right\},$$

$$\hat{B} = \frac{1}{n} \sum_{i=1}^n \hat{B}_i.$$

$2Z_i - 1$ is $+1$ for $Z_i = 1$ and -1 for $Z_i = 0$.

To obtain a variance estimator, rewrite

$$\begin{aligned} \hat{\tau}^{\text{mbc}} &= \hat{\tau}^{\text{m}} - \hat{B} \\ &= \frac{1}{n} \sum_{i=1}^n \left(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{B}_i \right) \\ &= \frac{1}{n} \sum_{i=1}^n \hat{\psi}_i. \end{aligned}$$

Think of $\hat{\tau}^{\text{mbc}}$ as a sample average of the $\hat{\psi}_i$'s.

Variance estimator

$$\hat{V}^{\text{mbc}} = \frac{1}{n^2} \sum_{i=1}^n \left(\hat{\psi}_i - \hat{\tau}^{\text{mbc}} \right)^2.$$

$\hat{\sigma}^2 = \frac{1}{n} \sum_i (\hat{\psi}_i - \hat{\tau}^{\text{mbc}})^2$ is the variance of a single-observation contribution. Dividing by n gives the variance of the sample average.

Connection to doubly robust estimators

Recall fitted outcome regressions $\hat{\mu}_0, \hat{\mu}_1$ and residuals

$$\hat{R}_i = \begin{cases} Y_i - \hat{\mu}_1(X_i), & Z_i = 1, \\ Y_i - \hat{\mu}_0(X_i), & Z_i = 0. \end{cases}$$

Outcome-regression estimator:

$$\hat{\tau}^{\text{reg}} = \frac{1}{n} \sum_{i=1}^n \left(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) \right).$$

Doubly robust estimator (Chapter 12):

$$\hat{\tau}^{\text{dr}} = \hat{\tau}^{\text{reg}} + \frac{1}{n} \sum_{i=1}^n \left\{ \frac{Z_i \hat{R}_i}{e(X_i)} - \frac{(1 - Z_i) \hat{R}_i}{1 - e(X_i)} \right\}.$$

Outcome regression augmented by inverse propensity weighting of residuals.

A similar form for matching

The bias-corrected matching estimator can be written

$$\hat{\tau}^{\text{mbc}} = \frac{1}{n} \sum_{i=1}^n \left[\underbrace{\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)}_{\hat{\tau}^{\text{reg}} \text{ contribution}} + (2Z_i - 1)\hat{R}_i + \frac{(2Z_i - 1)K_i}{M}\hat{R}_i \right].$$

Here K_i is the number of times unit i is used as a match.

Grouping the residual terms, we can define

$$\hat{\psi}_i = \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + (2Z_i - 1) \left(1 + \frac{K_i}{M} \right) \underbrace{\left(Y_i - \hat{\mu}_{Z_i}(X_i) \right)}_{\hat{R}_i}.$$

Then

$$\underbrace{Z_i \left(1 + \frac{K_i}{M} \right)}_{\text{treated}} - \underbrace{(1 - Z_i) \left(1 + \frac{K_i}{M} \right)}_{\text{control}}$$

plays a role analogous to the propensity weighting in doubly robust estimators.

$$\begin{aligned} Z_i = 1 : \quad & 1 + \frac{K_i}{M} \approx \frac{1}{e(X_i)}, \\ Z_i = 0 : \quad & 1 + \frac{K_i}{M} \approx \frac{1}{1 - e(X_i)}. \end{aligned}$$

What does this mean intuitively?

For a treated unit i :

A small $e(X_i)$ means its propensity to be in the treated group is low.

There are likely many similar controls to match with.

When $e(X_i)$ is small,

$$\frac{1}{e(X_i)} \text{ becomes large,}$$

so K_i (the number of times the unit is matched) also becomes large for fixed M .

Lin et al. (2023) provide formal theory for why we should allow M to grow with the sample size.

Thus we can view matching as a method for estimating propensity-score weights, and bias-corrected matching as a doubly robust estimator.

Matching for the Treated Units

We can repeat the above for the *average causal effect on treated units*.

$$\tau_T = \mathbb{E}[Y(1) \mid Z = 1] - \mathbb{E}[Y(0) \mid Z = 1].$$

Matching estimator:

$$\hat{\tau}_T^m = \frac{1}{n_1} \sum_{i=1}^n Z_i (Y_i - \hat{Y}_i(0)),$$

where n_1 is the number of treated units.

Bias

$$\hat{B}_{T,i} = \frac{1}{M} \sum_{k \in \mathcal{J}_i} \{ \hat{\mu}_0(X_i) - \hat{\mu}_0(X_k) \}, \quad (Z_i = 1),$$

$$\hat{B}_T = \frac{1}{n_1} \sum_{i=1}^n Z_i \hat{B}_{T,i}.$$

Bias-corrected estimator

$$\hat{\tau}_T^{\text{mbc}} = \hat{\tau}_T^{\text{m}} - \hat{B}_T.$$

This can be rewritten using

$$\hat{\psi}_{T,i} = Z_i(Y_i - \hat{\mu}_0(X_i)) - (1 - Z_i)\frac{K_i}{M}(Y_i - \hat{\mu}_0(X_i)).$$

Variance estimator

$$\hat{V}_T^{\text{mbc}} = \frac{1}{n_1^2} \sum_{i=1}^n \left(\hat{\psi}_{T,i} - \hat{\tau}_T^{\text{mbc}} \frac{n_1}{n} \right)^2.$$

$\hat{\tau}_T^{\text{mbc}}$ behaves like the sum/average of the $\hat{\psi}_{T,i}$ terms, rescaled by n_1 .

Connection to a doubly robust estimator for the treated

Outcome-regression estimator:

$$\hat{\tau}_T^{\text{reg}} = \frac{1}{n_1} \sum_{i=1}^n Z_i (Y_i - \hat{\mu}_0(X_i)).$$

Doubly robust estimator:

$$\hat{\tau}_T^{\text{dr}} = \hat{\tau}_T^{\text{reg}} - \frac{1}{n_1} \sum_{i=1}^n \frac{e(X_i)}{1 - e(X_i)} (1 - Z_i) \hat{R}_i.$$

Converting bias-corrected matching into a similar form,

$$\hat{\tau}_T^{\text{mbc}} = \hat{\tau}_T^{\text{reg}} - \frac{1}{n_1} \sum_{i=1}^n \frac{K_i}{M} (1 - Z_i) \hat{R}_i.$$

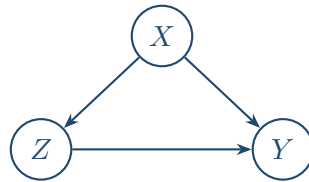
$$\boxed{\frac{K_i}{M} \approx \frac{e(X_i)}{1 - e(X_i)}} \quad \text{for a matched control unit.}$$

Chapter 16: Causal Diagrams

Overview

- Basics of a causal diagram.
- Motivating example.
- Assessing the unconfoundedness assumption:
 - Negative outcomes.
 - Negative exposures.
- Over-adjustment for confounding:
 - M-bias.
 - Z-bias.
- Example for review.

Basics of a Causal Diagram



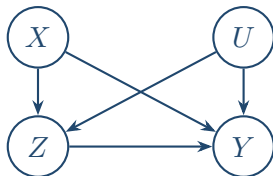
$$\begin{aligned}X &= f_X(\varepsilon_X), \\Z &= f_Z(X, \varepsilon_Z), \\Y(z) &= f_Y(X, z, \varepsilon_Y(z)).\end{aligned}$$

Assume $\varepsilon_Z \perp\!\!\!\perp \varepsilon_Y(z)$, for $z = 0, 1$.

$$\boxed{Z \perp\!\!\!\perp Y(z) \mid X.}$$

Unconfoundedness holds as long as we can observe and condition on all variables that affect both Z and Y .

What if we have an unmeasured confounder U ?



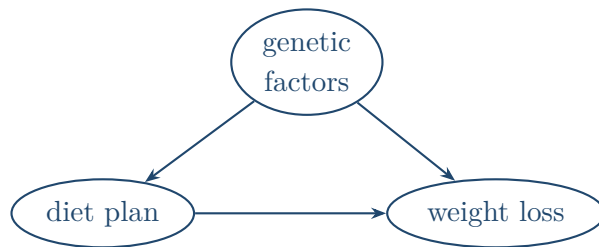
$$\begin{aligned}X &= f_X(\varepsilon_X), \\U &= f_U(\varepsilon_U), \\Z &= f_Z(X, U, \varepsilon_Z), \\Y(z) &= f_Y(X, U, z, \varepsilon_Y(z)).\end{aligned}$$

$$Z \not\perp\!\!\!\perp Y(z) \mid X.$$

U is unmeasured, so conditioning on X alone does not block the confounding.

Motivating example: Diet plan $\rightarrow$ weight loss

Suppose we think we have observed all the relevant variables.

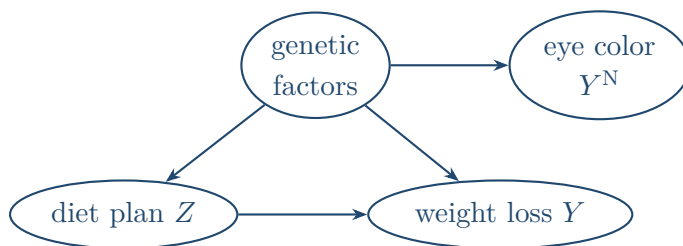


Have we considered all relevant genetic factors?

We cannot directly test unconfoundedness. Instead, look for evidence against it.

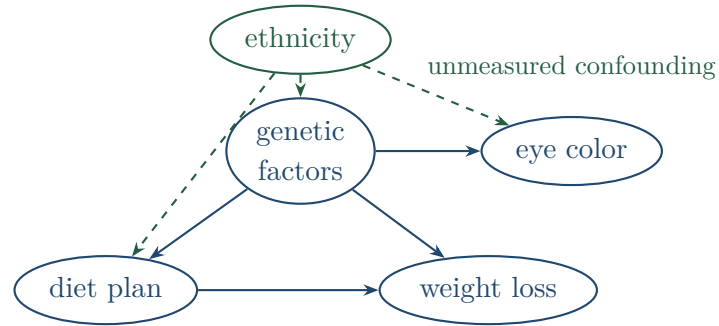
1. Negative outcomes

A negative outcome Y^N is affected by the confounder but *not* by the treatment.



An observed association between treatment and a negative outcome can indicate remaining confounding.

Unmeasured confounding in the example



If we believe that

$$Z \perp\!\!\!\perp Y(z) \mid X,$$

then we would also expect

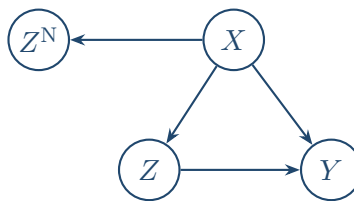
$$Z \perp\!\!\!\perp Y^N(z) \mid X$$

for an appropriately chosen negative outcome.

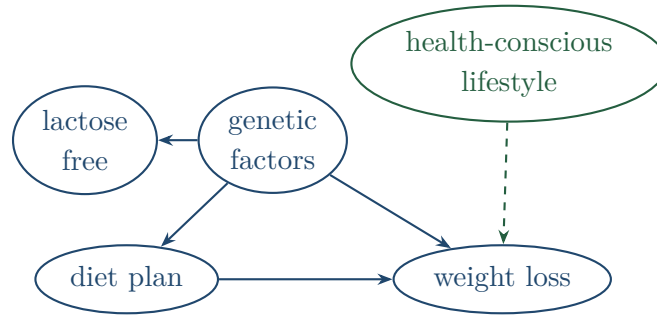
$$\tau(Z \rightarrow Y^N) = \mathbb{E}[Y^N(1) - Y^N(0)] = 0.$$

2. Negative exposures

A negative exposure Z^N is affected by the confounders but does not affect the outcome.



A negative-exposure example



If we believe $Z \perp\!\!\!\perp Y(z) \mid X$, then for a valid negative exposure we would expect

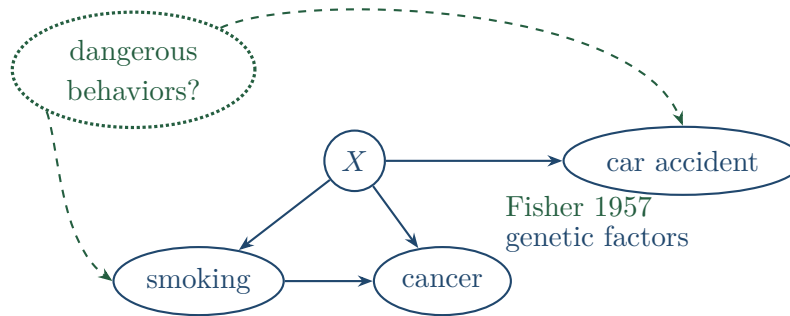
$$Z^N \perp\!\!\!\perp Y(z) \mid X, \quad \tau(Z^N \rightarrow Y) = 0.$$

Practical considerations

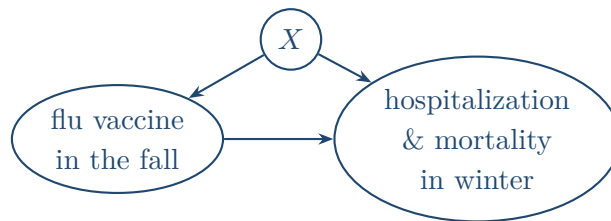
1. These tests can strengthen or weaken evidence for unconfoundedness; they cannot prove the assumption.
2. It is nontrivial to find convincing negative outcomes or negative exposures. One must understand the causal problem well.

Examples of Negative Controls

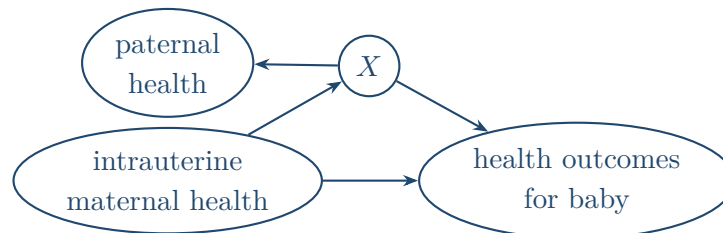
Smoking, cancer, and car accidents



Flu vaccination in the fall and hospitalization in winter



Paternal health, maternal health, and birth outcomes



Problems with Over-Adjusting

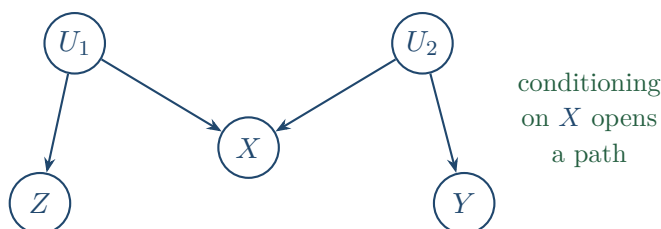
Rosenbaum (2002b) and Rubin (2007) motivate adjustment for observed pretreatment covariates.

Based on what we have seen, should we always control for all observed pretreatment covariates?

Pearl disagrees.

1. M-bias.
2. Z-bias.

M-bias



X is a collider on $U_1 \rightarrow X \leftarrow U_2$. Without conditioning on X , the noncausal path is blocked. Conditioning on X opens it.

Without conditioning on X ,

$$Z \perp\!\!\!\perp Y, \quad \tau_{\text{PF}} = \mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0) = 0.$$

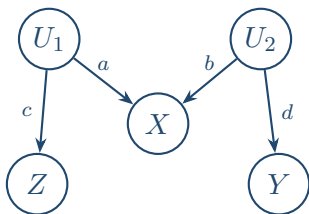
The simple estimator returns an unbiased result in the no-effect example.

However, conditioning on X can induce dependence:

$$Z \not\perp\!\!\!\perp Y \mid X,$$

$$\int \left[\mathbb{E}(Y \mid Z = 1, X = x) - \mathbb{E}(Y \mid Z = 0, X = x) \right] f(x) dx \neq 0.$$

Consider the following case:



$$X = aU_1 + bU_2 + \varepsilon_X,$$

$$Z = cU_1 + \varepsilon_Z,$$

$$Y = Y(z) = dU_2 + \varepsilon_Y.$$

$$U_1, U_2 \stackrel{\text{iid}}{\sim} N(0, 1).$$

Z and Y have distinct causes but conditioning on their common-effect X creates an association.

When $Z \perp\!\!\!\perp Y$, $\text{cor}(Z, Y) = 0$.

After conditioning on X , the partial correlation becomes (recall Chapter 1)

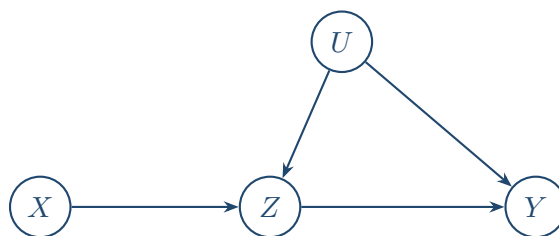
$$\rho_{ZY|X} = \frac{\rho_{ZY} - \rho_{ZX}\rho_{YX}}{\sqrt{1 - \rho_{ZX}^2}\sqrt{1 - \rho_{YX}^2}}.$$

Since $\rho_{ZY} = 0$,

$$\rho_{ZY|X} = -\frac{\rho_{ZX}\rho_{YX}}{\sqrt{1 - \rho_{ZX}^2}\sqrt{1 - \rho_{YX}^2}} \propto -abcd \quad (\text{sign / dependence}).$$

Spurious association arises from conditioning on the collider.

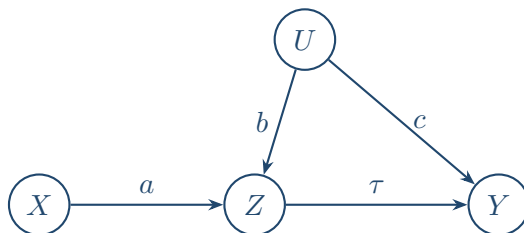
Z-bias



The covariate X is associated with treatment Z , not with outcome Y except through Z . Conditioning on X can amplify the remaining confounding through U .

Intuition

Z is a function of X , U , and random error. When we condition on X , we make Z less random relative to U — U plays a larger role in the residual variation of Z .



Consider the structural equations:

$$\begin{aligned} Z &= aX + bU + \varepsilon_Z, \\ Y(z) &= \tau z + cU + \varepsilon_Y, \\ X, U &\stackrel{\text{iid}}{\sim} N(0, 1). \end{aligned}$$

Assume the independent errors have variance 1 in the following calculation.

Without conditioning on X

The regression coefficient for OLS of Y on Z is

$$\begin{aligned}\tau_{\text{unadj}} &= \frac{\text{cov}(Z, Y)}{\text{var}(Z)} \\ &= \frac{\text{cov}(Z, \tau Z + cU)}{\text{var}(Z)} \\ &= \frac{\tau \text{var}(Z) + c \text{cov}(Z, U)}{\text{var}(Z)} \\ &= \tau + c \frac{\text{cov}(aX + bU + \varepsilon_Z, U)}{\text{var}(Z)} \\ &= \tau + \frac{c}{\text{var}(Z)} [a \text{cov}(X, U) + b \text{var}(U)].\end{aligned}$$

Using independence of X and U and unit variances,

$$\text{cov}(Z, U) = b, \quad \text{var}(Z) = a^2 + b^2 + 1.$$

Thus,

$$\tau_{\text{unadj}} = \tau + \underbrace{\frac{cb}{a^2 + b^2 + 1}}_{\text{bias term}}.$$

Conditioning on X

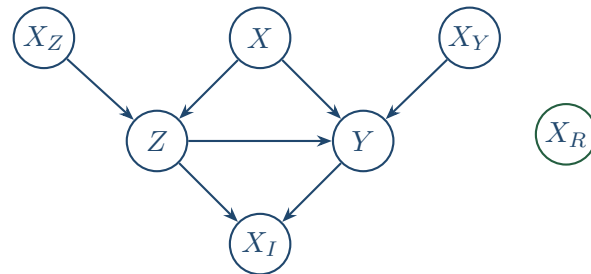
Regressing Z on X removes the aX term, so the remaining variation in Z is $bU + \varepsilon_Z$:

$$\tau_{\text{adj}} = \tau + \underbrace{\frac{cb}{b^2 + 1}}_{\text{new bias term}}.$$

The absolute bias increases when $a \neq 0$ and $cb \neq 0$, because conditioning removes variation associated with X without removing confounding from U .

If you know instrumental variables, X plays the role of an instrument in this diagram. This is also called instrumental-variable bias.

Difficulties in Unconfoundedness in Observational Studies



Z : smoking (treatment)

Y : lung cancer

X : genetic predisposition to nicotine addiction

X_Z : peer pressure

X_Y : genetic predisposition to lung cancer

X_I : other respiratory diseases

X_R : car accident

Which Covariates Should We Adjust For?

X : Affects both treatment and outcome, so we should control for it.

X_R : Random noise. Including it does not bias the outcome estimate, but introduces unnecessary variability in finite samples.

X_Z : With no unmeasured confounding, adjustment does not introduce bias; with unmeasured confounding, it can introduce Z -bias.

X_Y : Not required to remove confounding, but may improve precision because it predicts the outcome.

X_I : A post-treatment variable: do not adjust for it to estimate the total effect of treatment on the outcome.