Session 8: Threats to inference

MGT 581 | Introduction to econometrics

Michaël Aklin

PASU Lab | EPFL

Last time...

- Performance of OLS
- Bias/unbiasdness
- Consistency
- Efficiency

Today:

- Threats to inference
- Directed acyclic graphs

Readings:

- Stock and Watson (2011), ch 9
- Verbeek (2018), ch 3.2, 5.1-5.2
- Pearl (2009), Morgan and Winship (2014) ch3-4, Angrist and Pischke (2008) ch1, 4, 6

Threats to inference

Overview

- Omitted variable bias
- DAG
- Randomized controlled trial

Omitted variable bias

Omitted variable formula

- Consider this toy example:
  - True model: $y = \lambda + \tau x + \gamma z + \mu$
  - Estimated model: $y = \alpha + \beta x + \varepsilon$
- What is the consequence of using the wrong model?
- Earlier, we had:

$$\hat{\beta} = \frac{s_{xy}}{s_{xx}} = \frac{Cov(X,Y)}{V(X)}$$

$$E(\hat{\beta}) = E\left(\beta + \frac{\sum(x_i - \bar{x})\varepsilon_i}{s_{xx}}\right)$$

- The last part disappeared if the exogeneity assumption is met

But instead of the last expression, we know have:

$$
\begin{aligned}
E(\hat{\beta}) &= E\left(\beta + \frac{\sum(x_i - \bar{x})(\gamma z_i + \mu_i)}{s_{xx}}\right) \\
&= \beta + \gamma \frac{s_{xz}}{s_{xx}}
\end{aligned}
$$

Omitted variable formula

$\frac{s_{xz}}{s_{xx}}$ can be interpreted as:

$$\begin{aligned} x &= \phi + \omega z + \rho \\ &\equiv \phi + z\frac{s_{xz}}{s_{xx}} + \rho \end{aligned}$$

Thus:

$$\begin{aligned} E(\hat{\beta}) &= \beta + \gamma\frac{s_{xz}}{s_{xx}} \\ &= \beta + \gamma\omega \end{aligned}$$

Therefore, the bias is:

$$E(\hat{\beta}) - \beta = \gamma\omega$$

- Thus: to get unbiased estimates, we need to specify our econometric model correctly
- But how? Which variables should be included? Which ones shouldn't?
- In data science: regularization methods (LASSO, etc)
- Unlike data science, we are generally interested in **causal** estimates
- One way to build models: **directed acyclic graphs**
- Note: many other sources of bias: measurement error, non-random missingness of responses, etc.

Introduction to DAG

Directed acyclic graphs

- DAGs: developed by Pearl (2009)
- Responds to vague guidance from econometric analysis for modeling
- Idea: build up a graphical representation of **data generating processes**
- Use this representation to identify what variables to include (or not) in the model to be estimated

SCM

Structural causal model

- A set of endogenous variables, with corresponding distribution
  - Y, X, Z, ... with $f_y(\cdot)$, $f_x(\cdot)$, $f_z(\cdot)$ etc.
- Exogenous disturbances, with corresponding distribution
  - $\mathbf{U} \equiv [U_y, U_X, U_Z, ...]$, eg $U_Z \sim N(0, \sigma^2)$
- Causal relations between these variables ($\rightarrow$ or $\leftarrow$)

Example

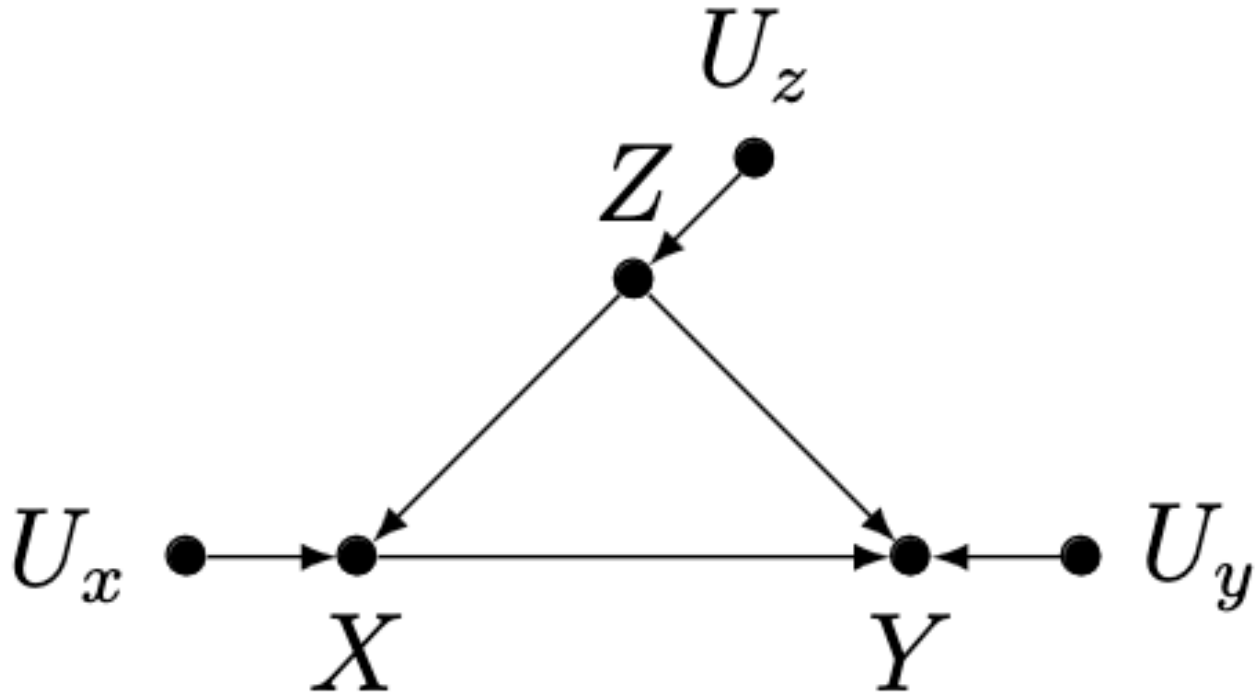


Figure 1: Directed acyclic graph: example.

- $X = f_x(Z, U_x)$
- $Y = f_y(X, Z, U_y)$
- Regressing Y on X: bias from omitted Z.

Example

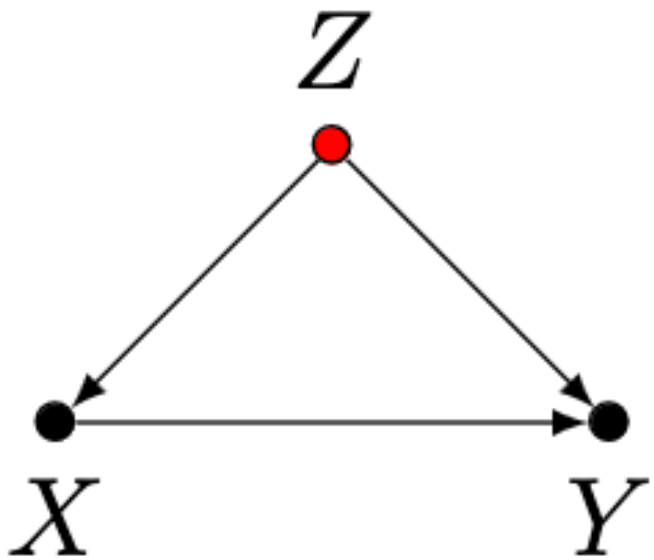


Figure 2: Directed acyclic graph: example (often with implicit disturbance).

Model building

- What to include and exclude?
- Fundamental principal: identify paths from your treatment X to outcome Y.
- What you need is...
  - (1) To block open **backdoor** paths
  - (2) **Not** to open blocked paths

Example

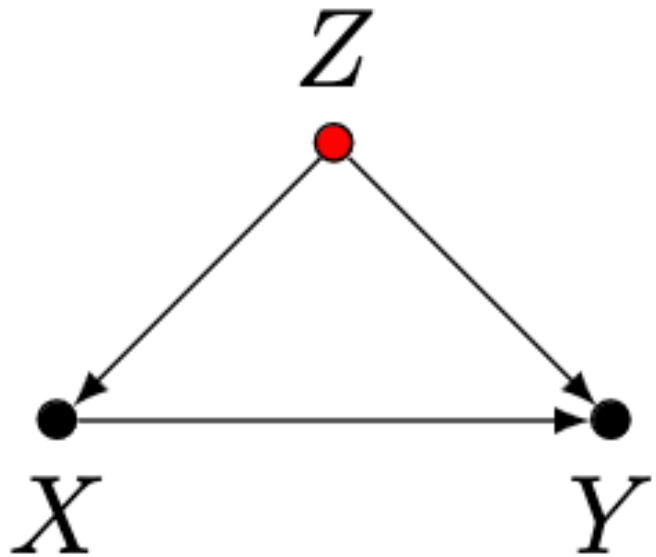


Figure 3: Directed acyclic graph: example (without disturbance.

Two paths:

- $X \to Y$
- $X \leftarrow Z \to Y$ (backdoor)
- For causal estimate of X on Y: need to **condition/adjust/control** for Z

Types of causal relations

- So should we just control for all variables? **No**!
- Only paths that need to be blocked are **backdoor** (arrows pointing at X)
- Three types of variables/paths
- **Forks** ("common cause"). Z is a fork in $X \leftarrow Z \rightarrow Y$
- **Mediators**. Z is a mediator in $X \rightarrow Z \rightarrow Y$
- **Collider** ("common effect"). Z is a collider in $X \rightarrow Z \leftarrow Y$

- Key: you must block backdoor paths that include forks and mediators.
- These generate confounding (or omitted variable bias).
- You do so by **conditioning** (or **controlling** or **adjusting**) for forks or mediators in the backdoor path.
- However: don't condition for colliders. Adjusting for them **opens** paths and creates correlation where there is none.

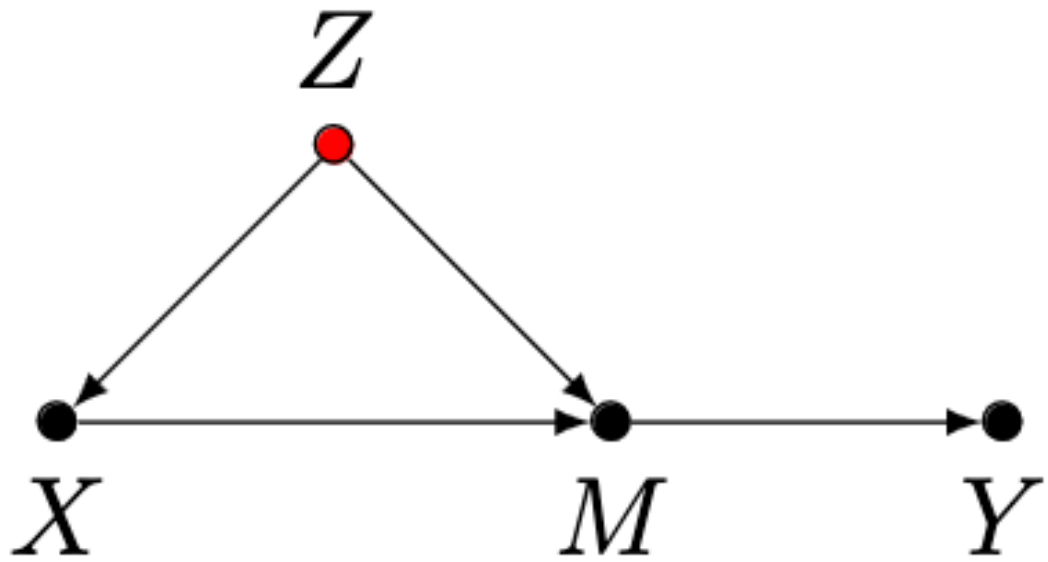


Figure 4: Control for Z?

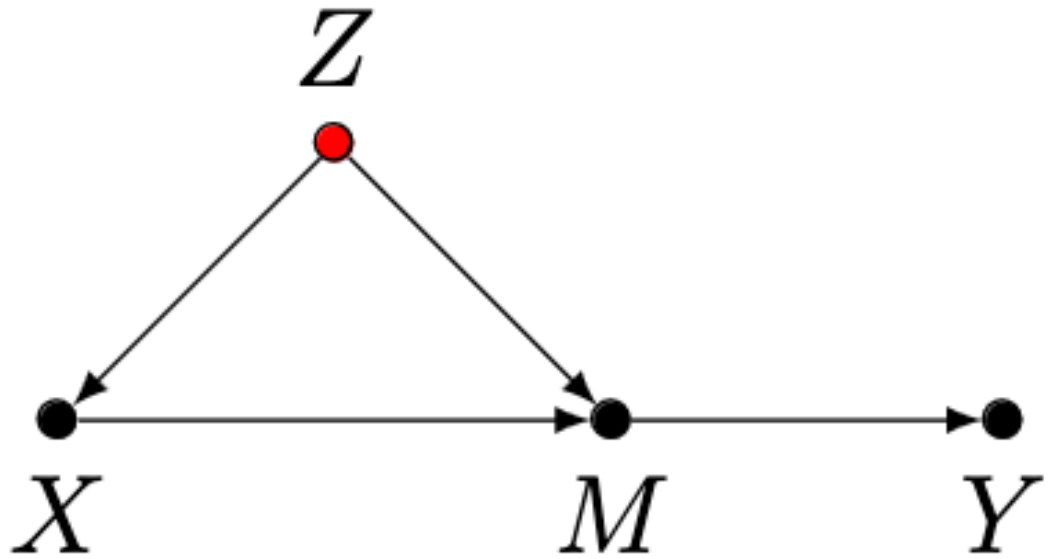


Figure 5: Control for Z?

Yes! Two paths:

- $X \to M \to Y$
- $X \leftarrow Z \to M \to Y$ (backdoor). Needs to be blocked!

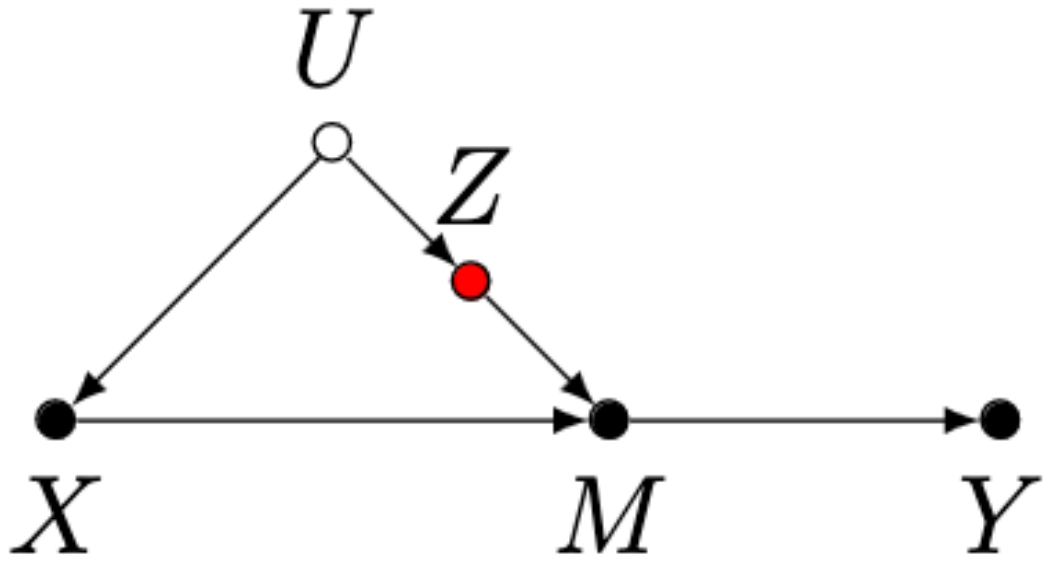


Figure 6: Control for Z?

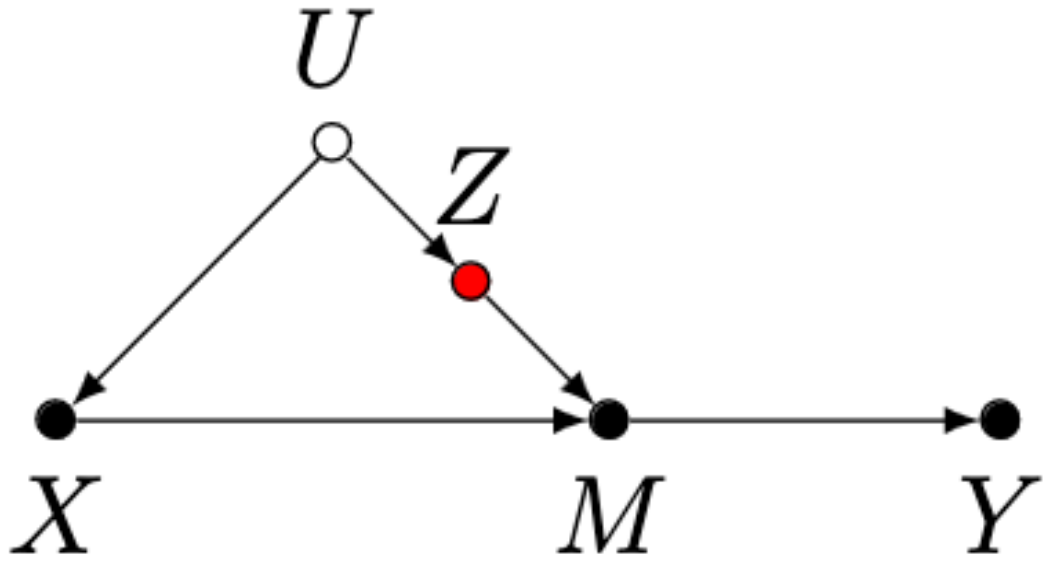


Figure 7: Control for Z? (Note: white circles means unobservable variable.)

Yes!

- $X \to M \to Y$
- $X \leftarrow U \to Z \to M \to Y$ (backdoor). Needs to be blocked!

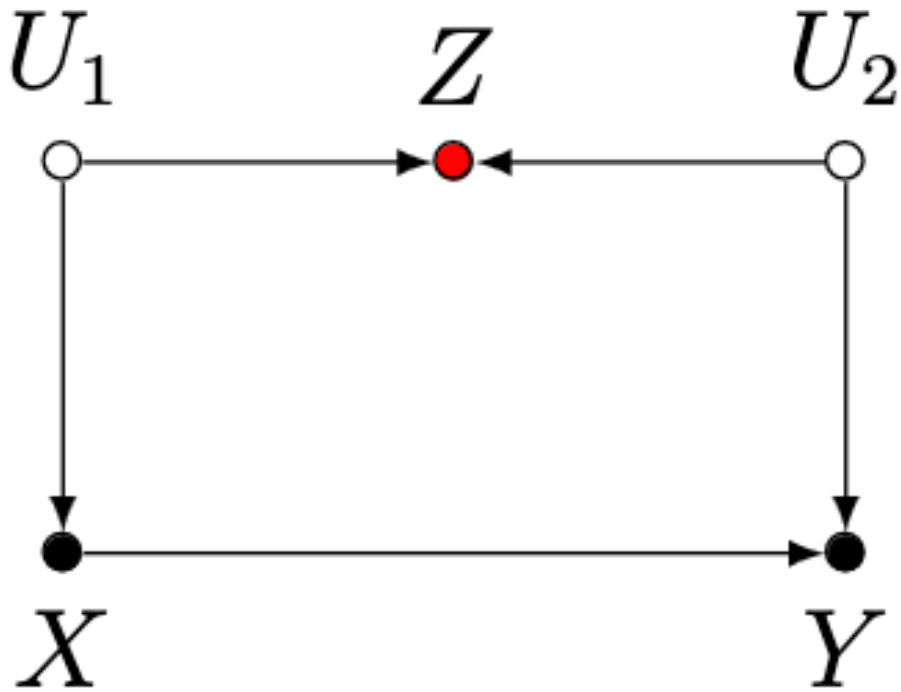


Figure 8: Control for Z?

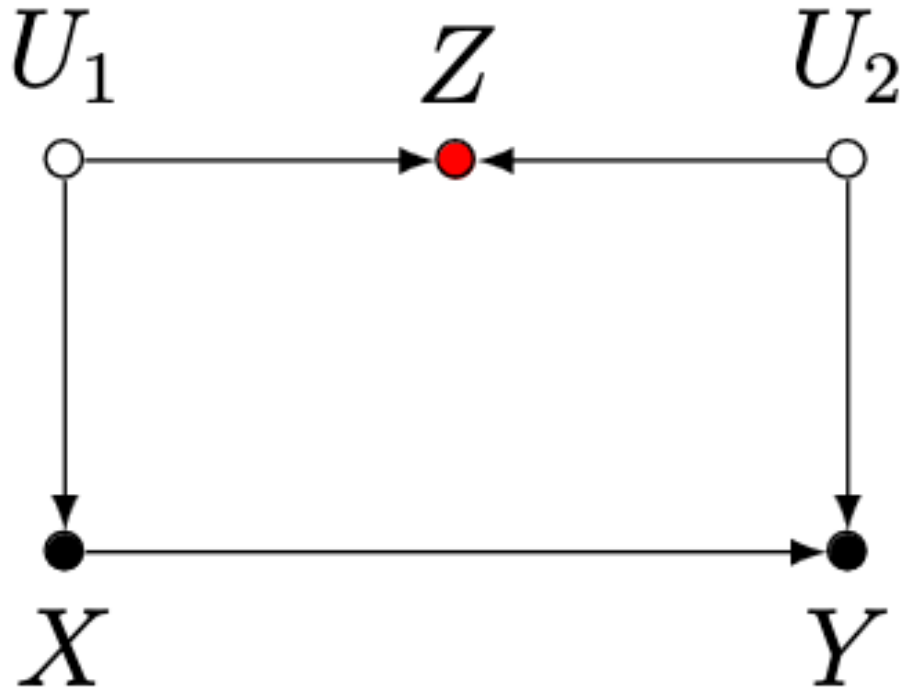


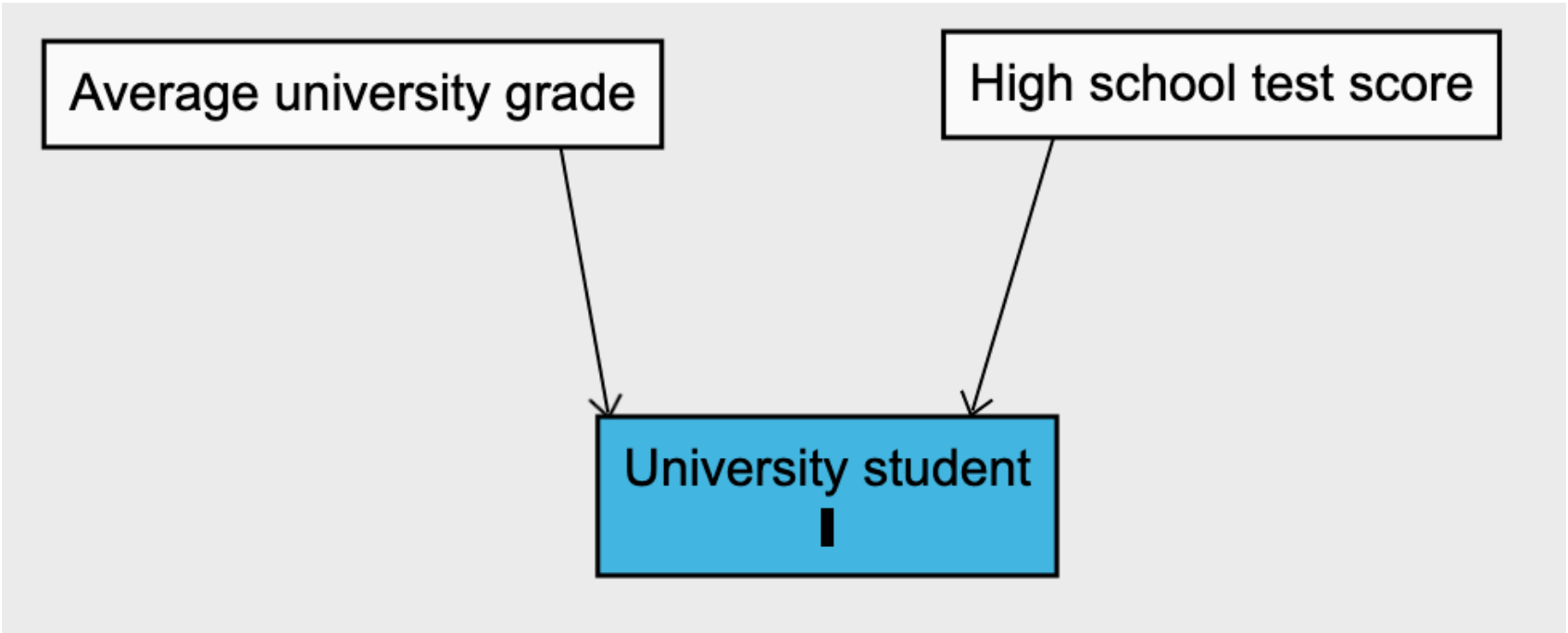
Figure 9: Control for Z?

No.

- $X \to Y$
- $X \leftarrow U_1 \to Z \leftarrow U_2 \to Y$. Z is a collider.

Collider bias

- Collider bias is one of the less intuitive problems identified by Pearl et al.
- Why should we not adjust for this backdoor path?
- Typical example: **sample selection**
- Consider the following example
  - Is there a causal effect of standardized test scores on university performance?
  - Estimating *Average Grade* $= \alpha + \beta$*High school test score* **among uni students**
  - $\beta < 0$: HS test scores reduce uni grades!
- Or...



- You are conditioning on a collider. Why is it bad?
- Can create correlation where there is none!
- Assume that high school and university test scores are uncorrelated *in the population*, but universities select students based on either (good HS students or students with high potential)

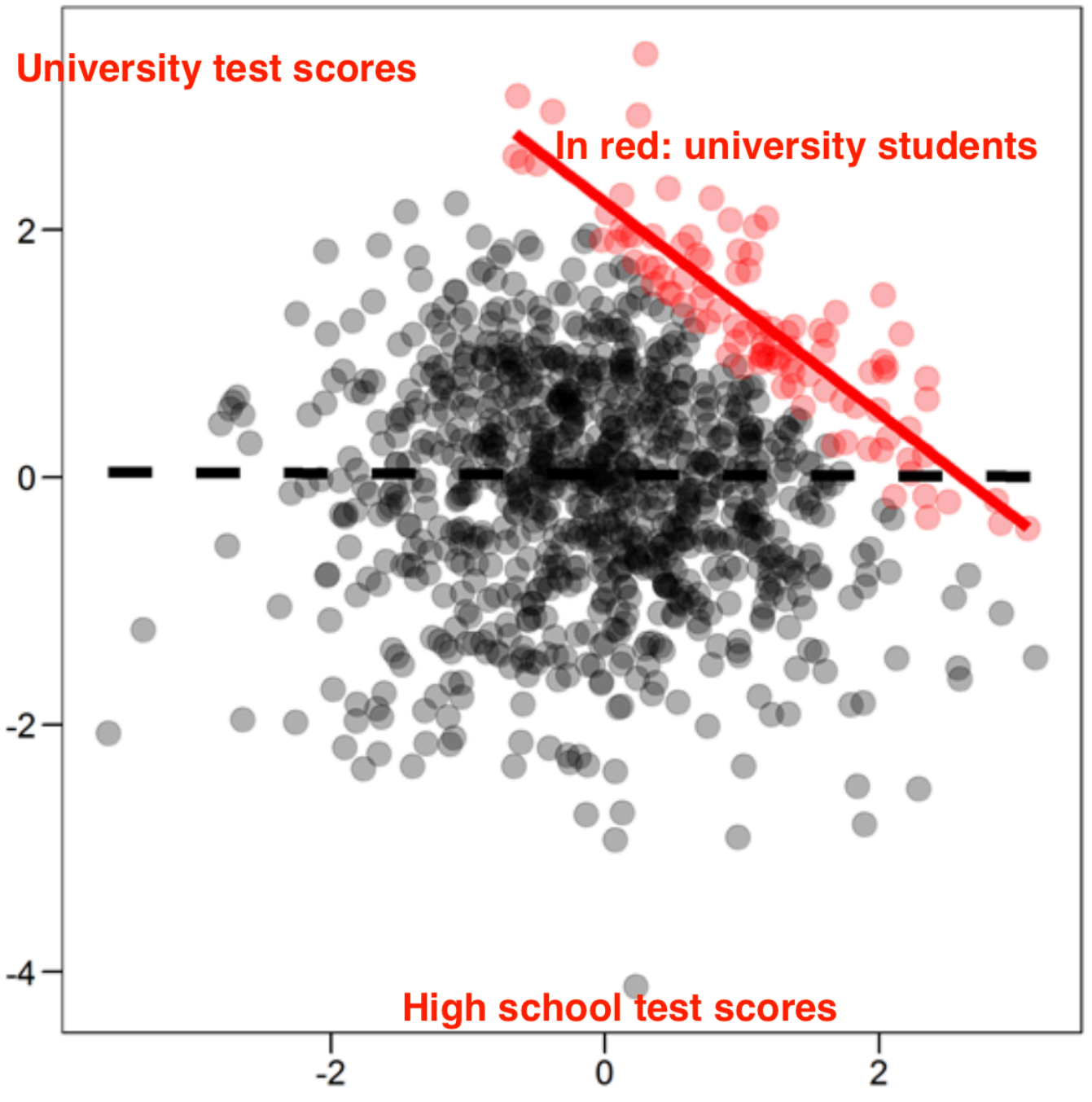


Figure 10: Conditioning on a collider

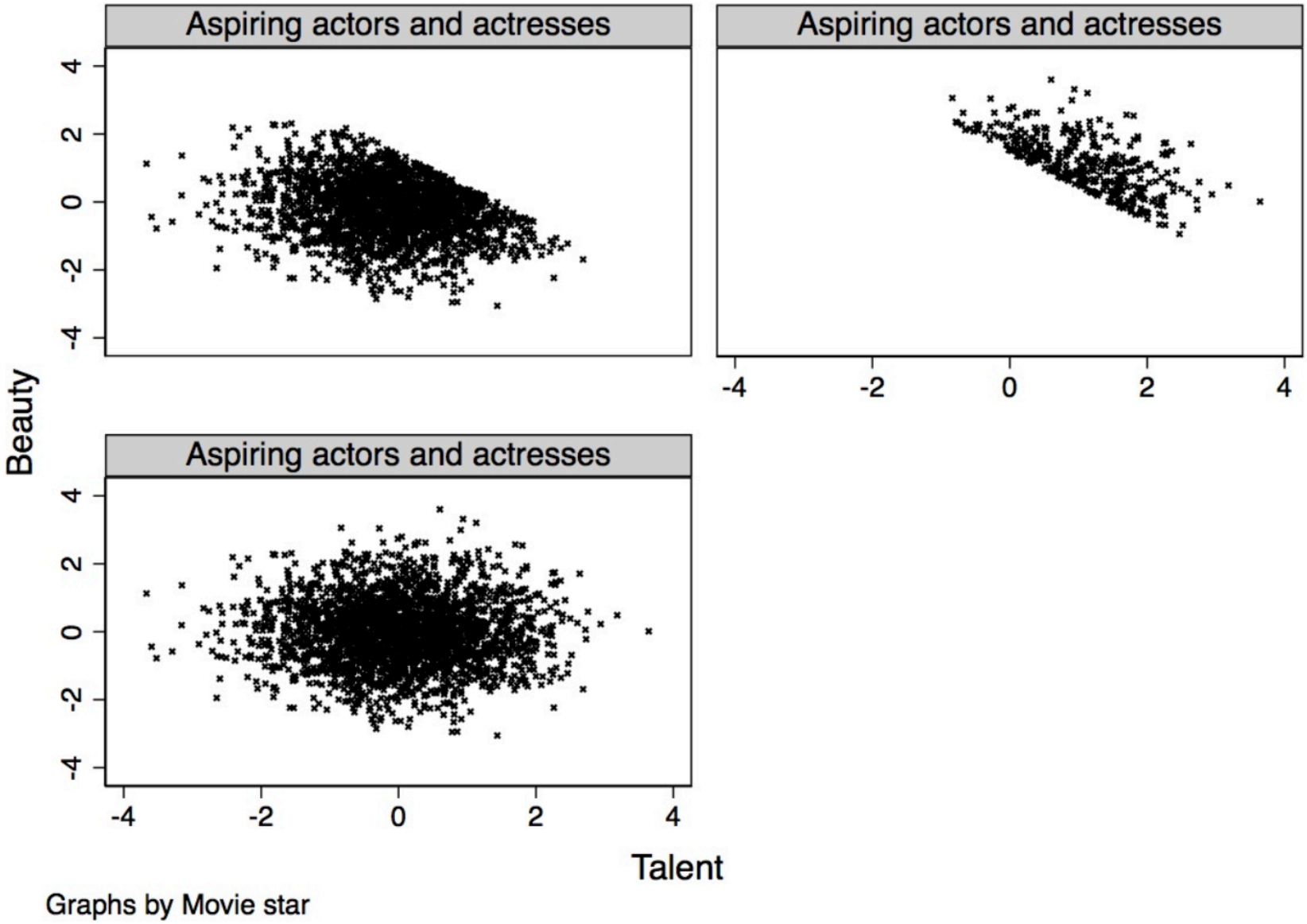


Figure 11: Is there a beauty/talent tradeoff? Don't just look at actors/actresses!

- So: we **should** condition/adjust/control for forks in an open back-door path
- We **shouldn't** condition/adjust/control for a collider in the back-door path
- What about the following cases?

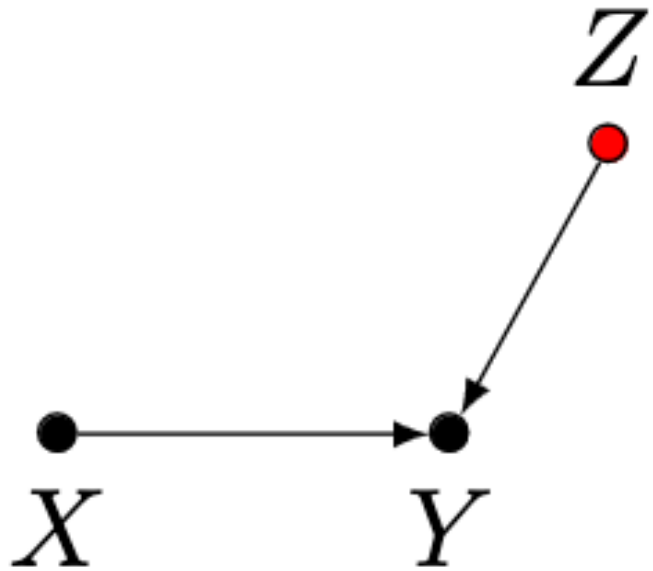


Figure 12: What about here?

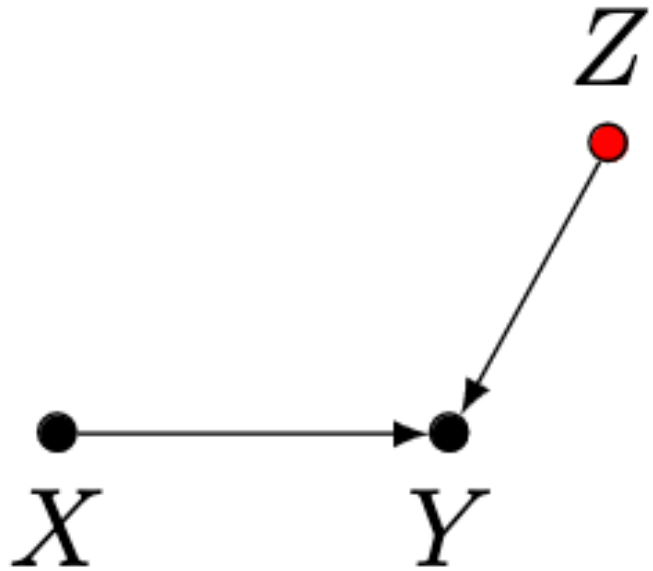


Figure 13: What about here?

Adjusting for Z won't affect estimates of X on Y. However, there is another reason to include Z.

(This is why we sometimes control for variables even in randomized controlled trials.)

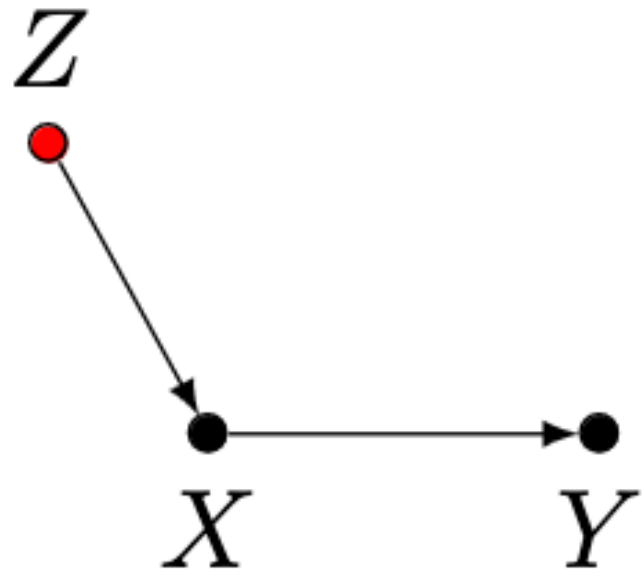


Figure 14: What about here?

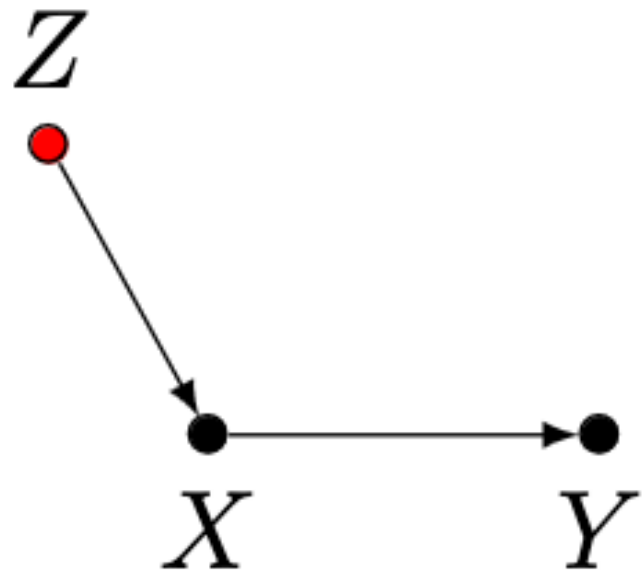


Figure 15: What about here?

Z will not reduce residuals. No reason to include it.

Figure 16: Mediator. Should we include it?

Figure 17: Mediator. Should we include it?

No. This is not a back-door path. No bias.

Adjusting for Z will "rob" X of its effect on Y.

In fact: risk of misleading results. Eg: gender wage discrimination.

As a general rule (with caveats):

- **Do** adjust for a fork in the back-door path.
  - A fork variable will introduce confounding.
  - Adjusting for it blocks the path from X to Y.
- **Don't** adjust for a mediator between X and Y.
  - A mediator will contain some of the effect of X on Y.
  - If $X \to M \to Y$: don't control for M.
- **Don't** adjust for a collider in the backdoor (or one of its descendant).
  - A collider does not introduce confounding.
  - But adjusting on it does! It opens a path.

Conclusion

- Key: it's really hard to build up a causal model
- DAGs help us decide what and what **not** to include
- Yet in many realistic settings: will be too hard to do with confidence
- Is there a way to bypass model-building? Yes, sometimes!
- Design for causal identification: **randomized controlled trial** (RCT) and **quasi-experimental methods**

Questions?

References

Angrist, Joshua, and Steffan J. Pischke. 2008. *Mostly Harmless Econometics*. Princeton: Princeton University Press.

Morgan, Steven L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd Edition. Cambridge: Cambridge University Press.

Pearl, Judea. 2009. *Causality*. 2nd Edition. New York: Cambridge University Press.

Stock, James H., and Mark W. Watson. 2011. *Introduction to Econometrics, 3rd Edition*. Pearson.

Verbeek, Marno. 2018. *A Guide to Modern Econometrics 5th Edition*. Wiley.