Directed Acyclic Graphs (DAGs)

Prof. Mike Aguilar

https:/www.linkedin.com/in/mike-aguilar-econ

DAGs for Causality - Preview

- DAGs are a visual tool used to represent causal relationships
- (Directed): arrows point one way
- (Acyclic): no cycles of causality
- DAGs are most useful in observational studies (as opposed to random experiments) to identify and control for confounders
- Tip: Use subject matter expertise to draw a DAG before any data analysis

Outline

1. Motivation - Case Study

2. DAGs
   - Intro to DAGs
   - Good Controls
   - Bad Controls
   - Neutral Controls

Outline

1. Motivation - Case Study

2. DAGs
   - Intro to DAGs
   - Good Controls
   - Bad Controls
   - Neutral Controls

Does Insurance Improve Health?



Image created with Bing Image Creator
The following is adapted from the book Mastering 'Metrics

Outcomes, Treatment, Confounders

Q: Can you identify Outcome, Treatment, and Confounder in our case?

Outcomes, Treatment, Confounders

Q: Can you identify Outcome, Treatment, and Confounder in our case?

- Outcome Y: Health (e.g. NHIS health index score)
- Treatment D: Insurance (e.g. health insurance)
- Confounder X: Strength (e.g. proclivity to getting sick)
- Treated: Khuzdar
- Control: Maria

Outline

1. Motivation - Case Study

2. DAGs
   - Intro to DAGs
   - Good Controls
   - Bad Controls
   - Neutral Controls

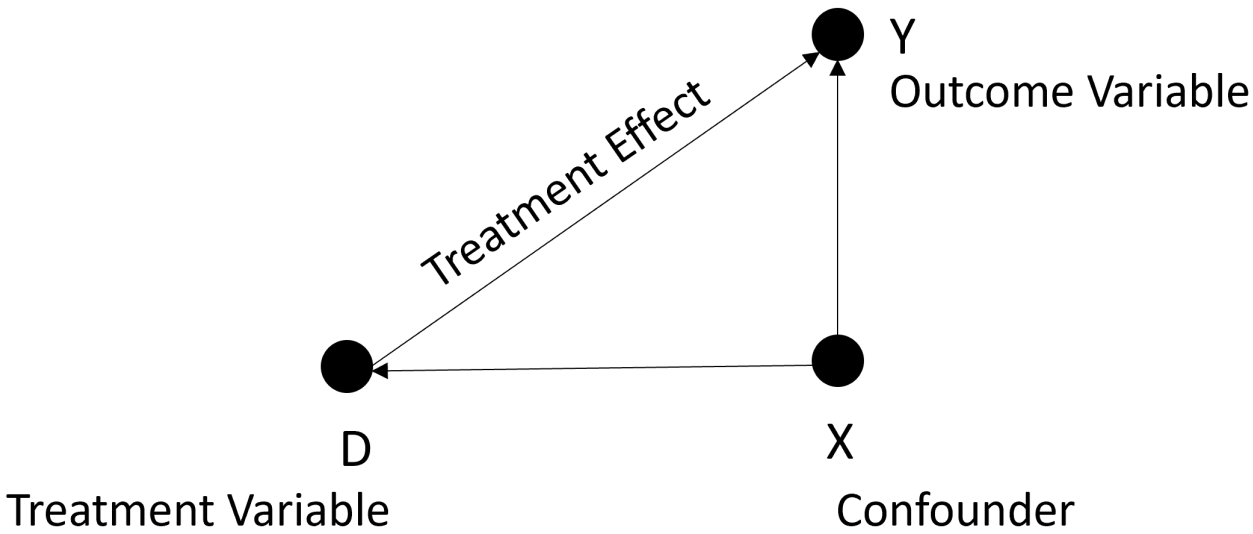
Outline

1. Motivation - Case Study

2. DAGs
   - Intro to DAGs
   - Good Controls
   - Bad Controls
   - Neutral Controls

Visualizing Causality

Directed Acyclic Graph (DAG)



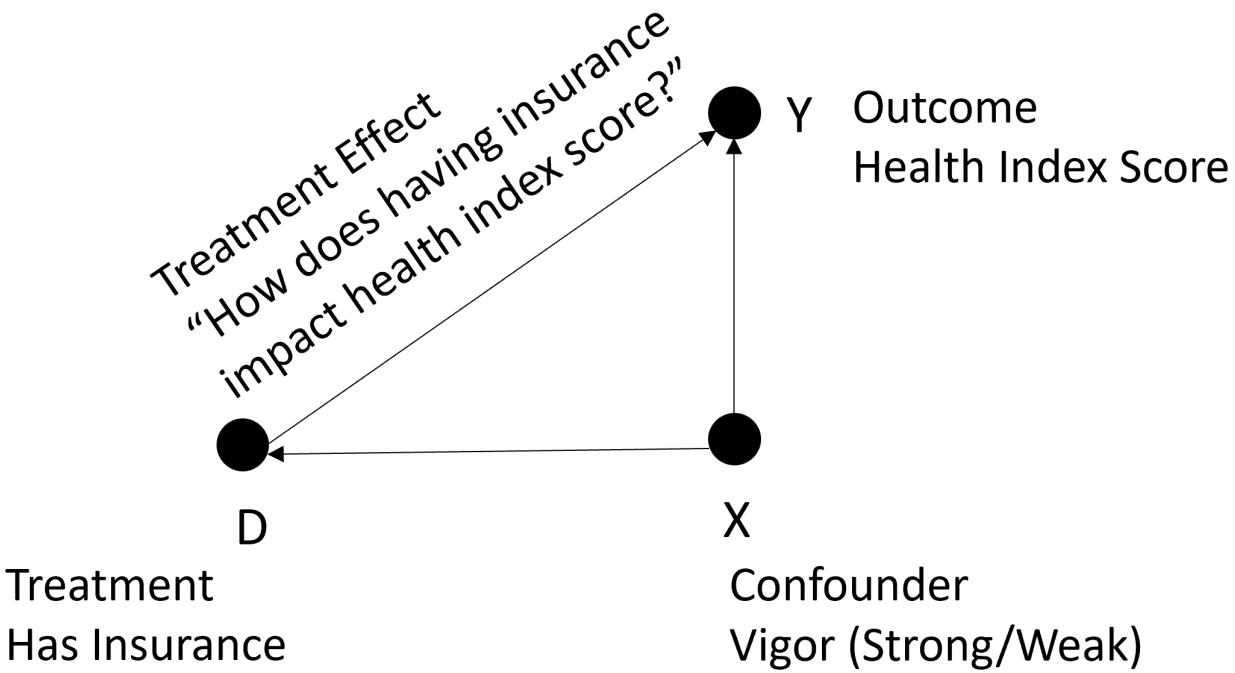
- Directed: Arrows point one way
- Acyclic: Rules out cycles of causality: $D \to Y \to X \to D \to \cdots$
- DAGs are inherently non parametric (i.e. arrows could represent linear or nonlinear relationships)
- DAGs are magnitude agnostic (i.e. arrows typically don't represent strength of relationships)

DAG for Insurance

Q: Can you draw the DAG and label the TE for our case?

DAG for Insurance

Q: Can you draw the DAG and label the TE for our case?



When are DAG's needed?

- In controlled experiments the analyst has control over the entire data generating process and then applies a treatment ("do" operation) to one variable. The DAG is a visualization of that causal pathway.

When are DAG's needed?

- In controlled experiments the analyst has control over the entire data generating process and then applies a treatment ("do" operation) to one variable. The DAG is a visualization of that causal pathway.
- In randomized experiments the analyst doesn't need DAG to estimate TE since covariate balance is ensured. Since most implementations of randomized experiments (ie RCTs) are possibly flawed sample approximations of the underlying DGP, a DAG might be helpful to evaluate validity.

When are DAG's needed?

- In controlled experiments the analyst has control over the entire data generating process and then applies a treatment ("do" operation) to one variable. The DAG is a visualization of that causal pathway.
- In randomized experiments the analyst doesn't need DAG to estimate TE since covariate balance is ensured. Since most implementations of randomized experiments (ie RCTs) are possibly flawed sample approximations of the underlying DGP, a DAG might be helpful to evaluate validity.
- In natural experiments, and observational settings more generally, a DAG is helpful to identify and control for confounders.
- Note that in controlled and random experiments the results help to verify a DAG, whereas in observational studies estimates of treatment effects are built assuming the DAG is true.

Confounders vs Controls

- Confounder: variable indicated within a DAG that impacts the relationship between treatment and outcome (i.e. impacts the treatment effect)

Confounders vs Controls

- Confounder: variable indicated within a DAG that impacts the relationship between treatment and outcome (i.e. impacts the treatment effect)
- Control: variable included within a statistical model to accommodate for the confounding variable's influence on the treatment effect

Confounders vs Controls

- Confounder: variable indicated within a DAG that impacts the relationship between treatment and outcome (i.e. impacts the treatment effect)
- Control: variable included within a statistical model to accommodate for the confounding variable's influence on the treatment effect

Q: Can you illustrate the difference between confounder and control within our case study?

Confounders vs Controls

- Confounder: variable indicated within a DAG that impacts the relationship between treatment and outcome (i.e. impacts the treatment effect)
- Control: variable included within a statistical model to accommodate for the confounding variable's influence on the treatment effect

Q: Can you illustrate the difference between confounder and control within our case study?

A: We might include vigor (innate strength/weakness) in our statistical model to "control" for the "confounding" relationship it has on the treatment effect of insurance upon health

Good vs Bad Controls

Not all controls should be included within your statistical model

Good vs Bad Controls

Not all controls should be included within your statistical model

- Good Control: A Control variable whose inclusion in the statistical model improves the estimate of the treatment effect.

Good vs Bad Controls

Not all controls should be included within your statistical model

- Good Control: A Control variable whose inclusion in the statistical model improves the estimate of the treatment effect.
- Bad Control: A Control variable whose inclusion in the statistical model worsens the estimate of the treatment effect.

Why are Bad Controls Bad?

- For ease of illustration, we will use linear DGPs and discuss controls within the context of OLS. The concepts of DAGs, Good & Bad Controls extend to other settings.

Why are Bad Controls Bad?

- For ease of illustration, we will use linear DGPs and discuss controls within the context of OLS. The concepts of DAGs, Good & Bad Controls extend to other settings.
- Bad Controls violate the orthogonality condition within a linear regression model (i.e. the regressors must be unrelated to the errors). By including Good Controls within the regression model the orthogonality condition is restored.

Why are Bad Controls Bad?

- For ease of illustration, we will use linear DGPs and discuss controls within the context of OLS. The concepts of DAGs, Good & Bad Controls extend to other settings.
- Bad Controls violate the orthogonality condition within a linear regression model (i.e. the regressors must be unrelated to the errors). By including Good Controls within the regression model the orthogonality condition is restored.
- The presence of Bad Controls implies that we shouldn't blindly throw all our data into a model. (WARNING: Avoid the ML-temptation to use all of your variables and use dimension reduction)

DAG Language and Rules

Using the style and notation from Cinelli, Forney, Pearl 2022

WARNING: The notation used here is slightly different than the rest of our notes. This is done only so that you can connect to the fantastic graphs produced by Cinelli et al.

●	A vertex representing an observed variable
○	A vertex representing an unobserved variable
Y	Outcome variable
X	Treatment variable
Z	Confounding variable
	Other variables used (U,W,M, etc..) represent various other relationships (e.g. mediator) that are beyond our scope.
→	Arrows represent the direction of causal relationships. If no relationship, then no arrow. Arrows go one way (directed). Arrows don't create a closed loop (e.g. X-> Y-> Z->X->…)

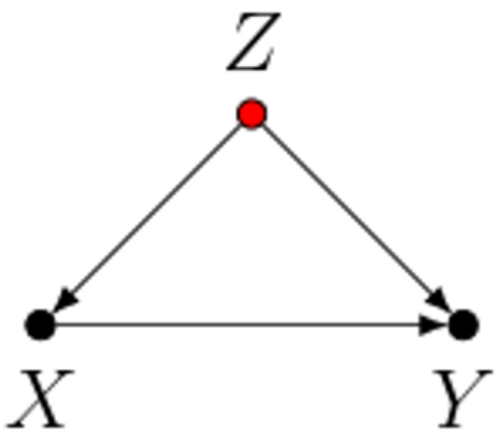
Outline

1. Motivation - Case Study

2. DAGs
   - Intro to DAGs
   - Good Controls
   - Bad Controls
   - Neutral Controls

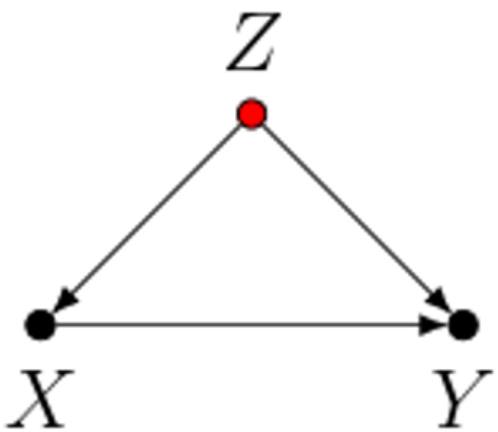
Confounders are Good Controls

Confounder (Z) influences treatment (X) and outcome (Y)
We should include Z (confounder) in our regression of Y on X



Confounders are Good Controls

Confounder (Z) influences treatment (X) and outcome (Y)
We should include Z (confounder) in our regression of Y on X

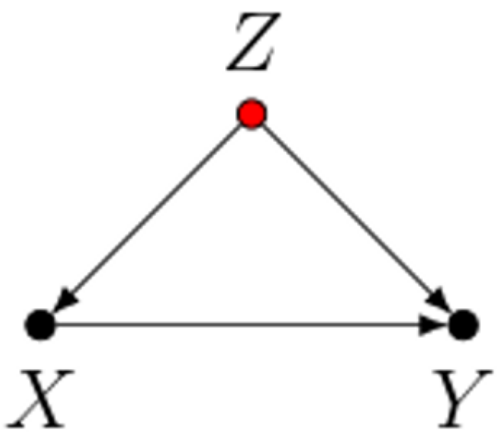


Data Generating Process (DGP)

$$Y = \alpha + \beta X + \gamma Z$$

Confounders are Good Controls

Confounder (Z) influences treatment (X) and outcome (Y)
We should include Z (confounder) in our regression of Y on X



Data Generating Process (DGP)

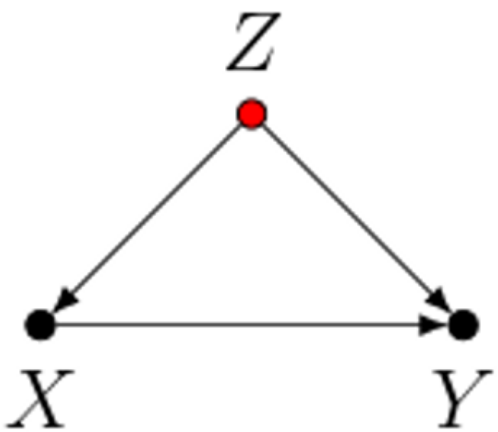
$$Y = \alpha + \beta X + \gamma Z$$

Misspecified model violates orthogonality

$$Y = \alpha + \beta X(Z) + \eta(Z)$$

Confounders are Good Controls

Confounder (Z) influences treatment (X) and outcome (Y)
We should include Z (confounder) in our regression of Y on X



Data Generating Process (DGP)

$$Y = \alpha + \beta X + \gamma Z$$

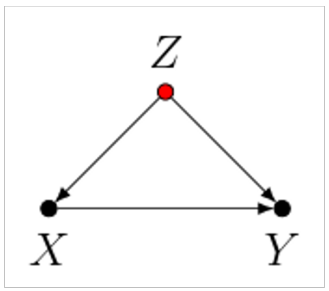
Misspecified model violates orthogonality

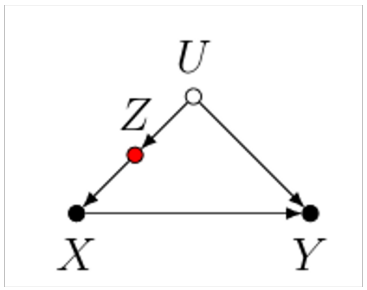
$$Y = \alpha + \beta X(Z) + \eta(Z)$$

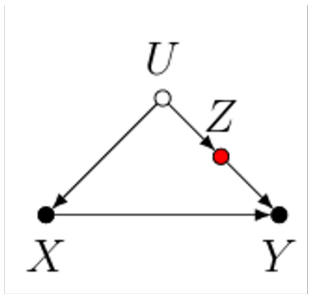
Controlling for Z rectifies

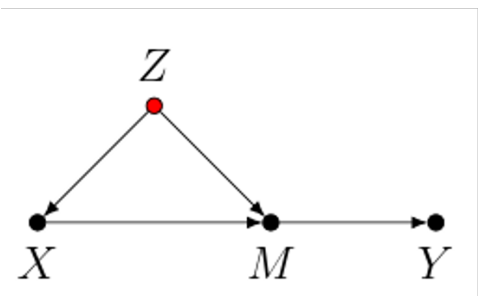
$$Y = \alpha + \beta X(Z) + \gamma Z + \eta$$

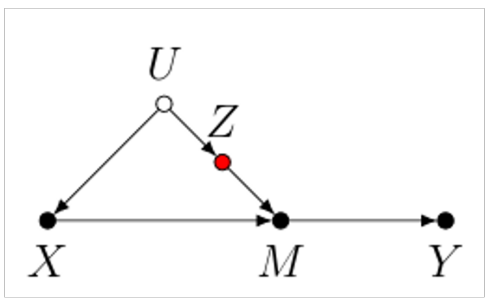
Good Controls

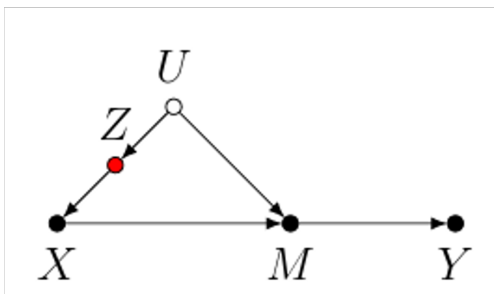












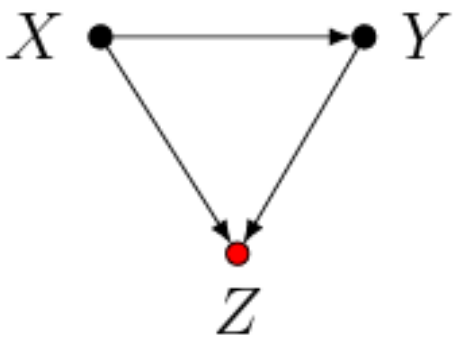
Outline

1. Motivation - Case Study

2. DAGs
   - Intro to DAGs
   - Good Controls
   - Bad Controls
   - Neutral Controls

Colliders are Bad Controls

Collider (Z) is influenced by treatment (X) and outcome (Y)
We should NOT include Z in our regression of Y on X

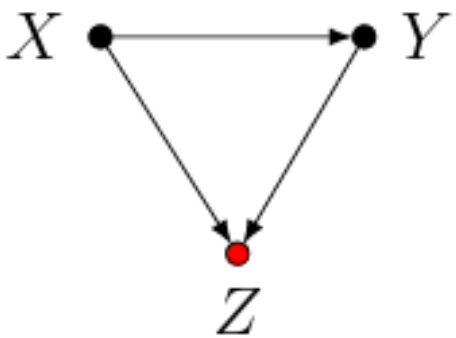


Data Generating Process (DGP)

$$Y = \alpha + \beta X$$
$$Z = \theta_0 + \theta_1 X + \theta_2 Y$$

Colliders are Bad Controls

Collider (Z) is influenced by treatment (X) and outcome (Y)
We should NOT include Z in our regression of Y on X



Data Generating Process (DGP)

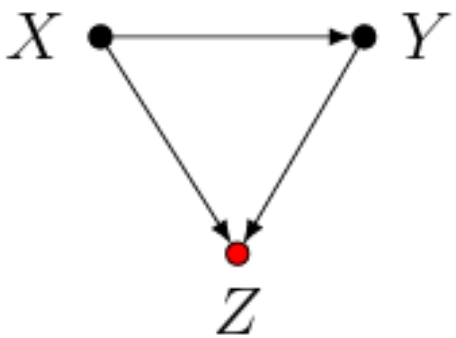
$$Y = \alpha + \beta X$$
$$Z = \theta_0 + \theta_1 X + \theta_2 Y$$

Misspecified model violates orthogonality

$$Y = \alpha + \beta X + \gamma Z + \eta$$
$$Y = \alpha + \beta X + \gamma(\theta_0 + \theta_1 X + \theta_2 Y) + \eta$$

Colliders are Bad Controls

Collider (Z) is influenced by treatment (X) and outcome (Y)
We should NOT include Z in our regression of Y on X



Data Generating Process (DGP)

$$Y = \alpha + \beta X$$
$$Z = \theta_0 + \theta_1 X + \theta_2 Y$$

Misspecified model violates orthogonality

$$Y = \alpha + \beta X + \gamma Z + \eta$$
$$Y = \alpha + \beta X + \gamma(\theta_0 + \theta_1 X + \theta_2 Y) + \eta$$

It is better not to control for Z

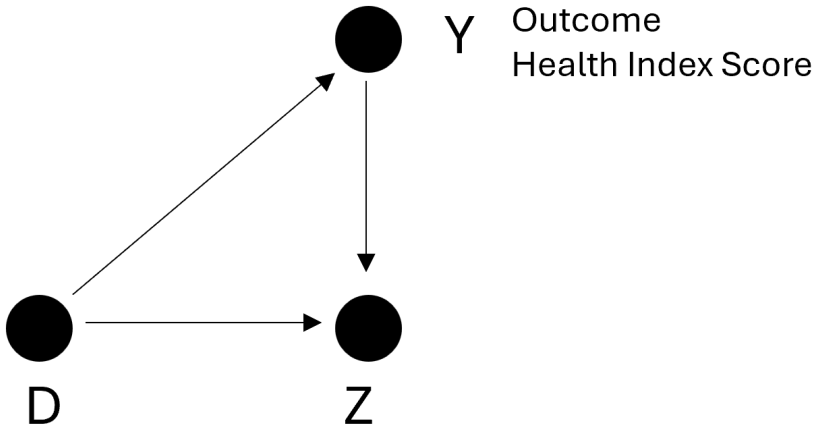
$$Y = \alpha + \beta X + \eta$$

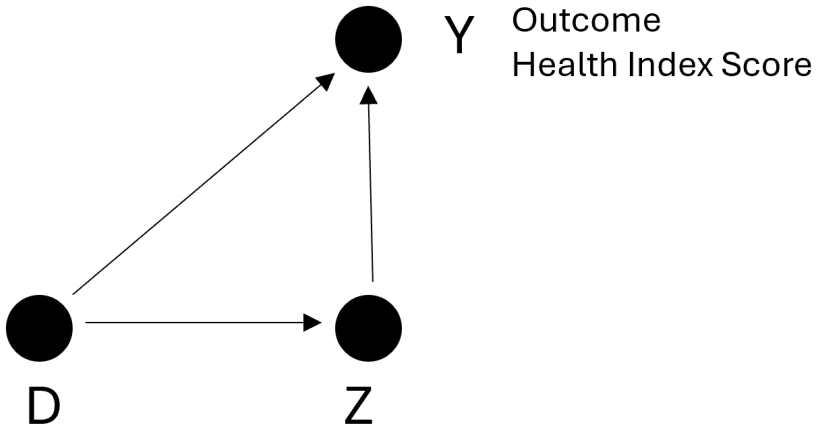
Bad Controls: Colliders and Mediators

Q: Can you explain the role of these Colliders and Mediators within our Case Study?

Bad Controls: Colliders and Mediators

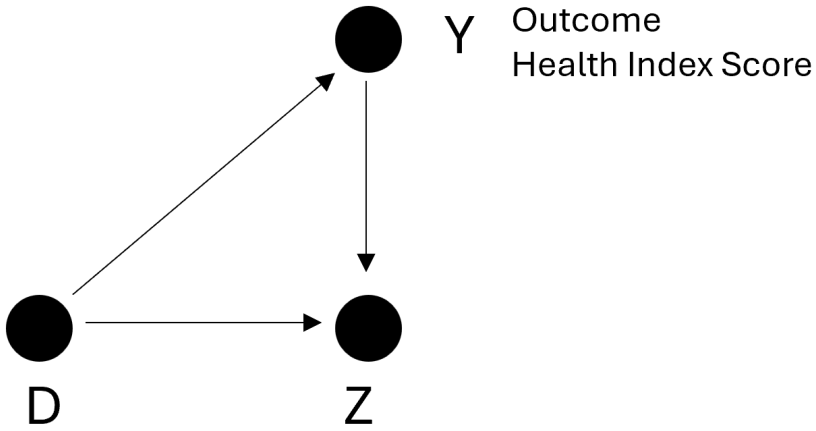
Q: Can you explain the role of these Colliders and Mediators within our Case Study?

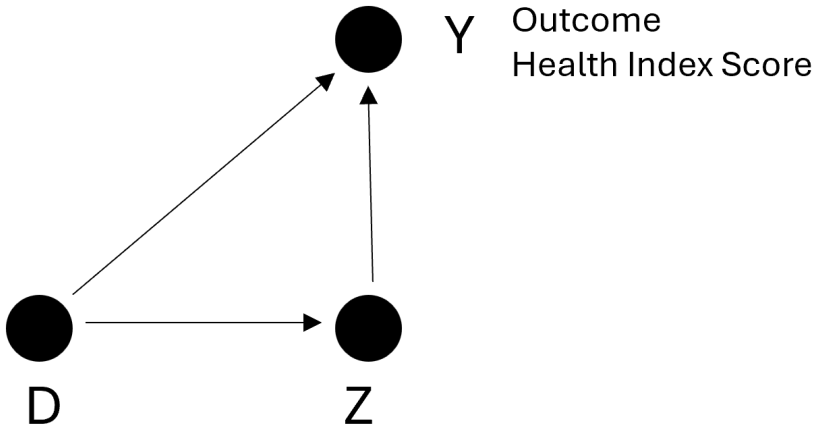




Bad Controls: Colliders and Mediators

Q: Can you explain the role of these Colliders and Mediators within our Case Study?



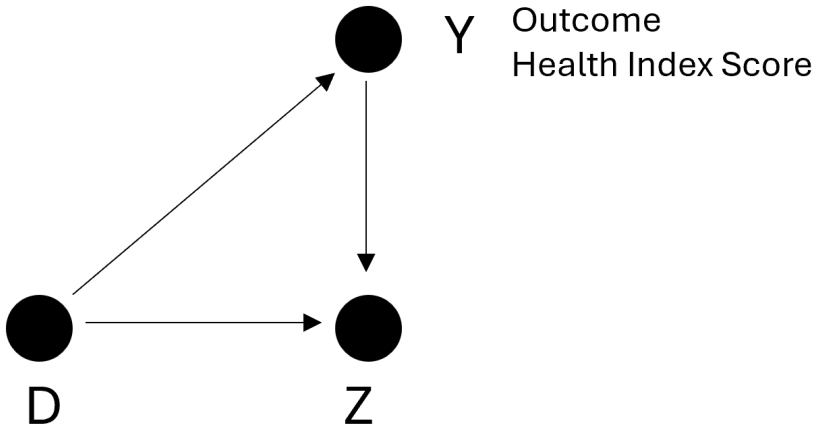


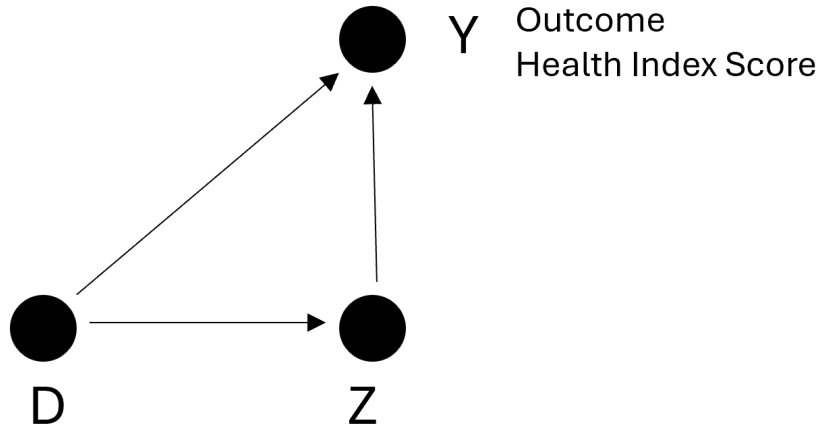
A:

- Collider: If someone has insurance, they might be more inclined to use health care services. If someone is less healthy, they might use more health care services.

Bad Controls: Colliders and Mediators

Q: Can you explain the role of these Colliders and Mediators within our Case Study?

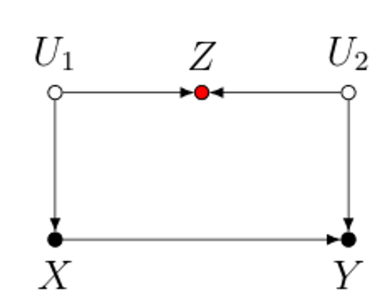


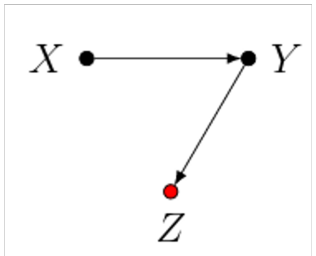


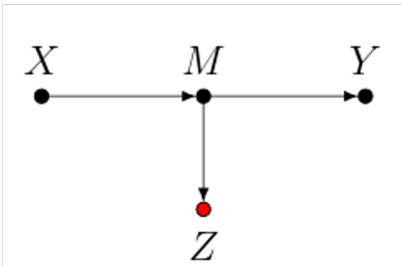
A:

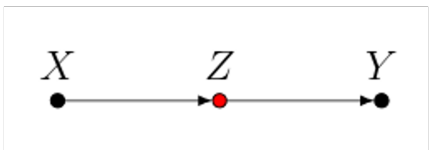
- Collider: If someone has insurance, they might be more inclined to use health care services. If someone is less healthy, they might use more health care services.
- Mediator: If someone has insurance, they might have better access to preventative care (e.g. wellness checks), which could impact their health separately from the direct impact of insurance upon health.

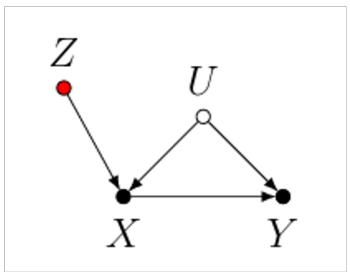
Bad Controls

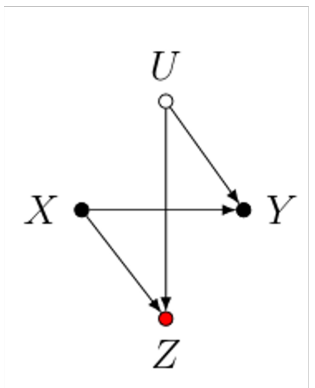


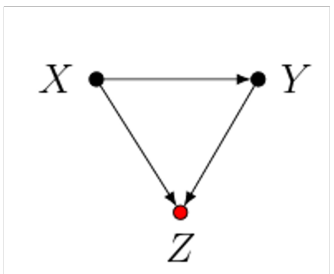


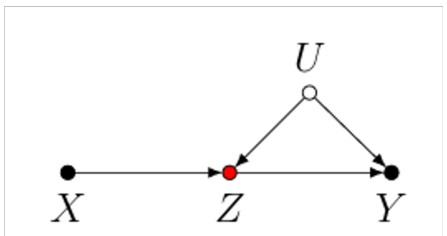










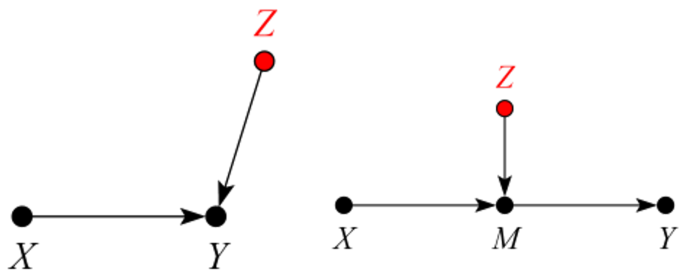


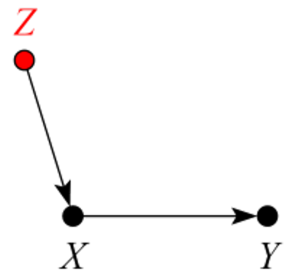
Outline

1. Motivation - Case Study

2. DAGs
   - Intro to DAGs
   - Good Controls
   - Bad Controls
   - Neutral Controls

Neutral Controls (no impact on bias)





May be good for precision by lowering

May be bad for precision by lowering

$$se(\hat{\beta}) = \hat{\sigma}_e \times \frac{1}{\sqrt{N}} \times \frac{1}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2}} \times \frac{1}{\sqrt{(1 - R_X^2)}}$$

In the formula, X is a vector of all regressors, including X and Z

DAGs for Causality - Summary

- DAGs are a visual tool used to represent causal relationships
- (Directed): arrows point one way
- (Acyclic): no cycles of causality
- DAGs are most useful in observational studies (as opposed to random experiments) to identify and control for confounders
- Tip: Use subject matter expertise to draw a DAG before any data analysis