SOC 690S: Machine Learning in Causal Inference

Week 5: Causal Inference from Directed Acyclic Graphs

Wenhao Jiang
Department of Sociology, Fall 2025
Duke University

Week	Date	Topic	Problem sets Assign	Problem sets Due
1	Aug 26	Introduction: Motivation and Linear Regression		
2	Sep 2	Foundation: Machine Learning Basics		
3	Sep 9	Foundation: Machine Learning Advanced	1	
4	Sep 16	Foundation: Potential Outcome Framework		
5	Sep 23	Foundation: Directed Acyclic Graph (DAG)	2	1
6	Sep 30	Core: Instrumental Variable Estimation		
8	Oct 7	Core: PSM and Doubly Robust Estimation	3	2
7	Oct 14	*Fall break*		
9	Oct 21	**In-class midterm**		
10	Oct 28	Core: Regression Discontinuity Design	4	3
11	Nov 4	Core: Panel Data and Difference-in-Difference		
12	Nov 11	Advanced: Heterogeneous Treatment Effect	5	4
13	Nov 18	Advanced: Unstructured Data Feature Engineering		
14	Nov 25	Advanced: Causal Reasoning in Machine Learning		5
	Dec 2	**Take-home final**		

Causal Inference via Linear Structural Equations

Structural Equation Model for Causal Inference

- Early conceptions of causality, most notably by Sewall Wright (a geneticist), used a *structural* approach to link variables

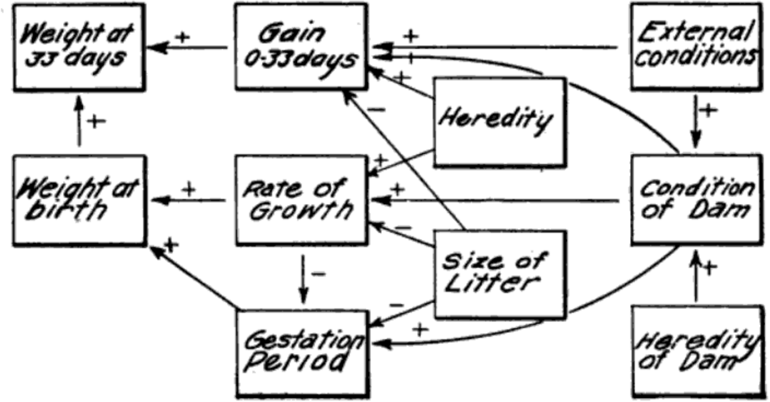
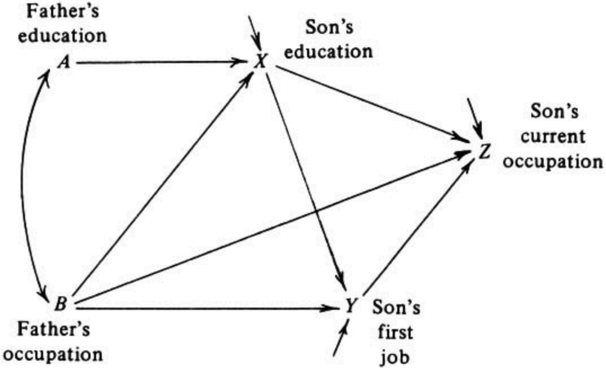


FIG. 1.—Diagram illustrating the interrelations among the factors which determine the weight of guinea pigs at birth and at weaning (33 days).

Structural Equation Model for Causal Inference

- This *structural* approach became known as *path analysis*, though most researchers did not interpret it under the framework of potential outcome
- Sociologists such as Peter M. Blau and Otis Dudley Duncan carried forward *path analysis* in the mid- and late-20th century



Structural Equation Model for Causal Inference

- Most applied researchers who used *path analysis* did not interpret it under the potential outcome framework for causal identification
- Judea Pearl introduced a formal graphical language (*directed acyclic graphs*, or DAG) in the 90s that linked *structural equation models* to precise rules for causal identification
- Pearl's framework provided clear conditions (*e.g.*, backdoor and frontdoor criteria) that determine when causal effects are identifiable from observational data
- We will start from a simple *structural equation model* to motivate the use of DAG

A Simple Triangular Structural Equation Model

- We start with a simple model of a household's demand for gasoline
- We model log-demand y given the log-price p

$$y(p) := \delta p$$

- where δ is the elasticity of demand. Demand is random across households, and we may model this randomness as U—a stochastic shock that describes variation of demand across households

$$Y(p) = \delta p + U, \quad E[U] = 0$$

- $Y(p)$ plays the *same* role as Rubin's potential outcome, *i.e.*,

$$E[Y(p_1) - Y(p_0)] = \delta(p_1 - p_0)$$

A Simple Triangular Structural Equation Model

- We may want to introduce covariates to capture other observable factors that may be associated with demand
- We may think there are observable parts of the stochastic shock (U), characterized by X, which help predict household demand; X may include family size, income, number of cars, or geographic location

$$U = X'\beta + \epsilon_Y$$

- where ϵ_Y is independent of X and has mean zero

$$Y(p) := \delta p + X'\beta + \epsilon_Y, \quad \epsilon_Y \perp\!\!\!\perp X$$

- This is a structural model of potential economic outcome; if log-price is set to p, then a household with X can be predicted to purchase $\delta p + X'\beta$ log-units of gasoline

A Simple Triangular Structural Equation Model

- In reality, the observed log-price P may depend on household characteristics X

$$P(X) := X'v + \epsilon_p, \quad \epsilon_p \perp\!\!\!\perp X$$

- where ϵ_p is independent of X and has mean zero
- For example, households located in different regions would experience different gas prices
- $\epsilon_p \perp\!\!\!\perp X$ assumes that household characteristics are determined well before gasoline prices faced by individual households are set

A Simple Triangular Structural Equation Model

- We also assume that households are otherwise *price-takers*, meaning the observed P is determined *outside* of the model conditioning on X
- For example, log-price P is independent of the stochastic shock of demand not captured by X

$$P \perp\!\!\!\perp \epsilon_Y \mid X$$

- This is the assumption of *conditional exogeneity*—the econometric analog of *conditional ignorability*

A Simple Triangular Structural Equation Model

- Under these assumptions and equations, we have a triangular structural equation model (TSEM)

$$
\begin{aligned}
Y &:= \delta P + X'\beta + \epsilon_Y \\
P &:= X'v + \epsilon_P \\
X &
\end{aligned}
$$

- ϵ_Y, ϵ_P, and X are mutually independent and determined *outside of the model*
- In SEM, they are called *exogeneous* variables
- Y and P are determined *within* the model and are called the *endogeneous* variables

A Simple Triangular Structural Equation Model

- What do we mean by the model being *structural*
- Each of the equation is assumed to model counterfactual scenarios

$$Y(p, x) := \delta p + x'\beta + \epsilon_Y$$
$$P(x) := x'v + \epsilon_P$$

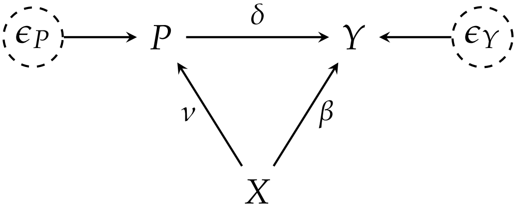
- The conceptual operation of “setting” the variables to their potential or counterfactual values is assumed to leave the structure intact
    - The structural parameters are supposed to be invariant to changes in the distribution of exogeneous variables—X, ϵ_Y, ϵ_P—that have been generated outside of the model
- We can therefore use these structural parameters to generate *counterfactual* predictions
- The structural parameters δ and v can be identified by linear regression

From SEM to DAG

From SEM to DAG

$$
\begin{aligned}
Y &:= \delta P + X'\beta + \epsilon_Y \\
P &:= X'v + \epsilon_P
\end{aligned}
$$

- This simple TSEM can be graphically depicted as a causal diagram
- Observed variables are shown as nodes, and causal paths are represented by directed arrows



From SEM to DAG

- The graph initiates with the *root nodes* $X, \epsilon_P, \epsilon_Y$
- The absence of links between the root nodes indicates orthogonality
- Understanding the orthogonality structure between nodes is an important input into identification of structural parameters through (linear) projection
- The nodes X and ϵ_P are *parents* of P; the nodes P, X, and ϵ_Y are *parents* of Y
- The node Y is a *collider*, as two or more other variables (two causal arrows in the graph) "collide" at that variable

From SEM to DAG

- The main effect of interest, δ, or the structural causal effect of P on Y, is identified after adjusting for X

$$Y(p) := \delta p + X'\beta + \epsilon_Y, \quad \epsilon_Y \perp\!\!\!\perp X, p$$

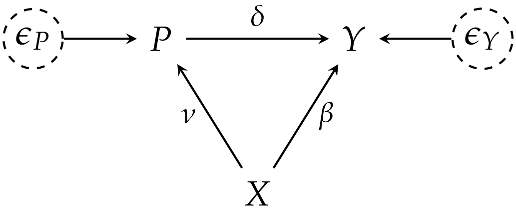
- In the language of the DAG, there are two paths connecting P and Y

$$P \to X \text{ and } P \leftarrow X \to Y$$

- The second path is called a *backdoor path*—there is a common cause for P and Y
- Figuratively speaking, controlling for X is said to be "closing the backdoor path" (or satisfying the *backdoor criterion*), shutting down the non-causal sources of statistical dependence between P and Y

From SEM to DAG

- Going back to the structural model; how do household characteristics impact gasoline demand
- The direct effect β via $X \rightarrow Y$
- The indirect effect $v\delta$ via $X \rightarrow P \rightarrow Y$
- The total effect of X on Y is $v\delta + \beta$, which can be identified by projection of Y on X
- There are no backdoor paths from X to Y; no adjustments are needed to identify the total effect of X on Y



From SEM to DAG

- We can also verify it by plugging in P

$$
\begin{aligned}
Y &= \delta(X'v + \epsilon_P) + X'\beta + \epsilon_Y \\
&= (v\delta + \beta)'X + (\epsilon_Y + \delta\epsilon_P) \\
\epsilon_Y + \delta\epsilon_P &\perp X
\end{aligned}
$$

- Therefore, $v\delta + \beta$ coincides with the projection coefficient in the projection of Y on X

From SEM to DAG

- We can also verify it by plugging in P

$$\begin{aligned} Y &= \delta(X'v + \epsilon_P) + X'\beta + \epsilon_Y \\ &= (v\delta + \beta)'X + (\epsilon_Y + \delta\epsilon_P) \end{aligned}$$

$$\epsilon_Y + \delta\epsilon_P \perp X$$

- Therefore, $v\delta + \beta$ coincides with the projection coefficient in the projection of Y on X
- Conditioning on P would allow us to identify the direct of X, *i.e.*, β, but would prevent us from identifying the total effect $v\delta + \beta$

The Backdoor Criterion

- A set of variables Z satisfies the *backdoor criterion* relative to an ordered pair of variables (D, Y) in a DAG if
  - No node in Z is a descendant of D; and
  - Z blocks every path between D and Y that contains an arrow into D (*i.e.*, every "backdoor path" from D to Y)

The Backdoor Criterion

- A set of variables Z satisfies the *backdoor criterion* relative to an ordered pair of variables (D, Y) in a DAG if
  - No node in Z is a descendant of D; and
  - Z blocks every path between D and Y that contains an arrow into D (*i.e.*, every "backdoor path" from D to Y)
- If Z satisfies the backdoor criterion, then the causal effect of D on Y is identifiable by conditioning on Z:

$$P(Y(d)) = P(Y \mid do(D = d)) = \int P(Y \mid D = d, Z = z) f(z) dz$$

- $do(D = d)$ means that we intervene in the system and set $D = d$ as a potential treatment
- The distribution of the potential outcome $Y(d)$ is obtained by averaging the observed distribution of Y among units with $D = d$ across strata of Z, weighted by the prevalence of each stratum

From Backdoor Criterion to Conditional Ignorability

- The *backdoor criterion* is closely linked with the potential outcome framework and the *conditional ignorability* assumption
- Under *conditional ignorability*, conditioning on covariates X (*i.e.*, within the same value or strata of X), treatments are as if randomly assigned

$$D \perp\!\!\!\perp Y(d) \mid X$$

- In reality, there could be a large array of covariates, and *"if 'good' is taken to mean 'best' fit, it is tempting to include anything in X_i that helps predict treatment."* (Wooldridge, 2005)
- The *backdoor criterion* provides the set of covariates to control for causal identification; its satisfaction is equivalent to fulfilling the *conditional ignorability* assumption

From Backdoor Criterion to Conditional Ignorability

- *The First Law of Causal Inference*: DAGs (and corresponding SEM) is equivalent to the potential outcome framework
- In the above case (replacing P by D as a typical notation for treatment), conditioning on X satisfies the *backdoor criterion*

$$Y(d) \perp\!\!\!\perp D \mid X$$
$$E[Y(d) \mid X] = E[Y(d) \mid D = d, X]$$

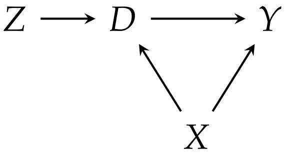
- This is also known as *d-sepration*; conditioning on X *d-separates* the actual treatment D and potential or counterfactual outcome $Y(d)$

The Backdoor Criterion

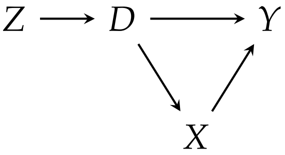
$$\begin{tikzcd}
Z \arrow[r] & D \arrow[rr] & & Y \\
& & X \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

Scenario 1

The Backdoor Criterion



Scenario 1



Scenario 2

The Backdoor Criterion

$$\begin{tikzcd}
Z \arrow[r] & D \arrow[rr] & & Y \\
& & X \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

Scenario 1

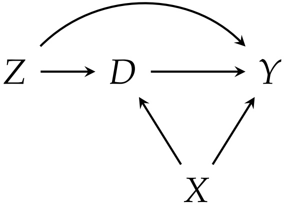
$$\begin{tikzcd}
Z \arrow[r] & D \arrow[rr] \arrow[dr] & & Y \\
& & X \arrow[ur] &
\end{tikzcd}$$

Scenario 2

$$\begin{tikzcd}
Z \arrow[r] & D \arrow[rr] \arrow[dr] & & Y \arrow[dl] \\
& & X &
\end{tikzcd}$$

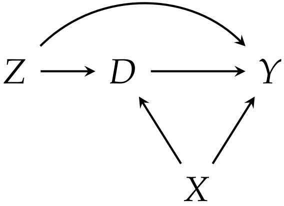
Scenario 3
collider X blocks the backdoor

The Backdoor Criterion

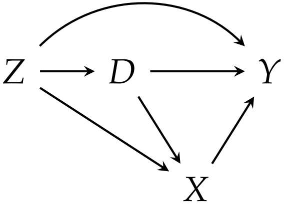


Scenario 4

The Backdoor Criterion

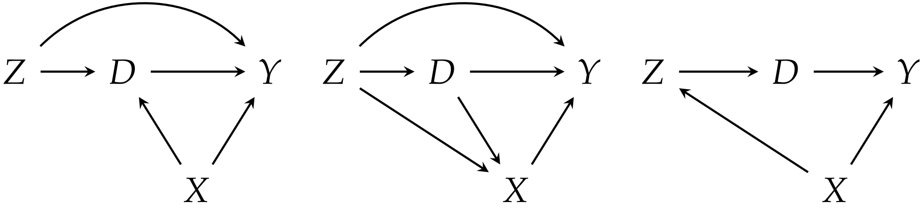


Scenario 4



Scenario 5

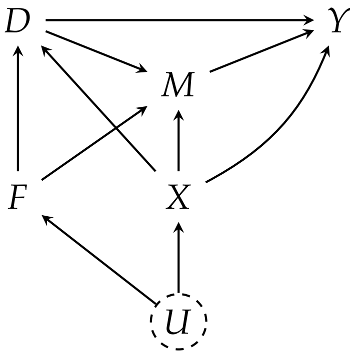
The Backdoor Criterion



Scenario 4 Scenario 5 Scenario 6

The Impact of 401(k) Eligibility on Financial Outcome

- 401(k) eligbility (D) might affect an individual's net financial assets (Y) both directly and indirectly through the employer's matching contribution (M)
- Worker-level (X) and firm-level (F) characteristics and latent factors (U) may additionally structure the model



Conditional Ignorability from DAG

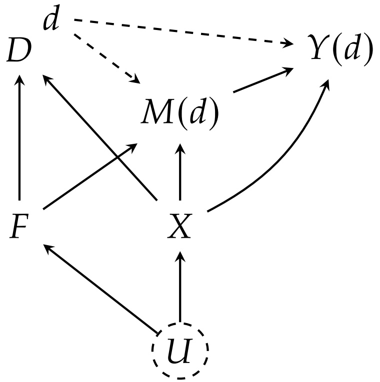
- In the case of 401(k) eligibility on financial outcome, conditioning on F and X satisfies the *backdoor criterion*

$$Y(d) \perp\!\!\!\perp D \mid F, X$$
$$E[Y(d) \mid F, X] = E[Y(d) \mid D = d, F, X]$$

- This is also known as *d-sepration*; conditioning on F and X *d-separates* the actual treatment D and potential or counterfactual outcome $Y(d)$

d-Separation from DAG and Conditional Ignorability

- The *actual* realization of treatment D is still a function of F and X
- D is *independent* of the potential outcome $Y(d)$ given F and X



When Conditioning Can Go Wrong: Collider Bias

When Conditioning Can Go Wrong: Collider Bias

- Consider the following SEM

$$\begin{aligned} T &:= \epsilon_T \\ B &:= \epsilon_B \\ C &:= T + B + \epsilon_C \end{aligned}$$

- where ϵ_T, ϵ_B, and ϵ_C are independent $\mathcal{N}(0,1)$ shocks; $E[T] = 0$, $E[T \mid B = b] = 0$

$$\begin{tikzcd} B \arrow[dr] & & T \arrow[dl] \\ & C & \end{tikzcd}$$

When Conditioning Can Go Wrong: Collider Bias

- Conditioning on *collider* C will create spurious dependence between B and T; $E[T \mid B, C] \neq 0$
- Remember $T = C - B - \epsilon_C$, $E[T \mid B, C]$ is the predicted value of T after linear projection of T on $C - B$

$$
\begin{aligned}
E[T \mid B, C] &= \frac{Cov(T, (C-B))}{V(C-B)}(C-B) \\
&= \frac{Cov(T, T+\epsilon_C)}{V(T+\epsilon_C)}(C-B) \\
&= \frac{1}{2}(C-B)
\end{aligned}
$$

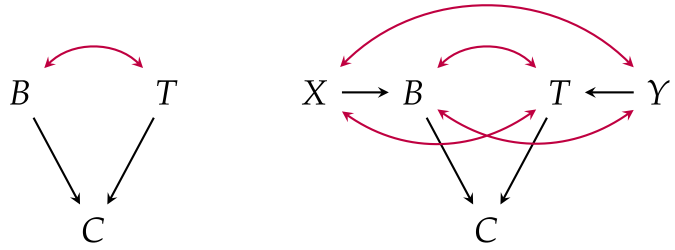
When Conditioning Can Go Wrong: Collider Bias

$$
\begin{aligned}
E[T \mid B, C] &= E\left[E[T \mid B = b, C]\right] \\
&= E\left[\frac{1}{2}(C - b)\right] \\
&= -\frac{b}{2}
\end{aligned}
$$

- Controlling for C, the predictive effect of B on T is $1/2$; this is *not* a causal effect (spurious)
- This is the *collider bias*, which is known as a form of sample selection bias (Heckman selection bias)

When Conditioning Can Go Wrong: Collider Bias

- Conditioning on a collider C opens associations among its parents *and all their ancestors*



When Conditioning Can Go Wrong: Collider Bias

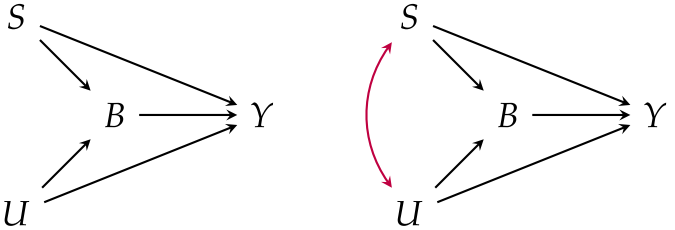
- In some cases, regression on a collider can be useful for *predictive* tasks
- Suppose the preceding SEM provides a simplified version of actors and actresses in Hollywood
- T denotes talent, C celebrity (success or popularity), and B bonhomie (approachability or friendliness
- Now if we condition on C (say the person remains in Hollywood $C > 0$), B and T will be *negatively* correlated
- For those without bonhomie characters, they have to be very talented to remain in Hollywood; a prediction of talent is possible within Hollywood

Collider Bias: Birth-Weight Paradox

- In a study conducted in 1991 in the US, it was found that infants born to smokers have:
  - Higher risk of low birth-weight (LBW)
  - Higher infant mortality
- However, among infants with LBW, mortality is *lower* for infants of smokers than for infants of non-smokers
- The naive interpretation is that smoking may be protective conditional on having LBW
- However, the more plausible explanation is LBW is a *collider*

Collider Bias: Birth-Weight Paradox

- U could be unobserved competing risks that can cause LBW and higher mortality



Collider Bias: Birth-Weight Paradox

$$
\begin{aligned}
Y &:= S + B + \kappa U + \epsilon_Y \\
B &:= S + U + \epsilon_B \\
S &:= \epsilon_S \\
U &:= \epsilon_U
\end{aligned}
$$

- where ϵ_U, ϵ_Y, ϵ_S, and ϵ_B are independent $\mathcal{N}(0,1)$ shocks
- If we project Y on S, we recover the correct positive causal effect of 2
- However, when we project Y on S and B, we learn a CEF of the form

$$E[Y \mid S, B] = S + B + (1 - \kappa/2)S + (1 + \kappa/2)B$$

- If the competing risks U increase infant mortality a lot, *i.e.*, $\kappa \gg 1$, the project recovers an erroneous large negative effect $1 - \kappa/2$ of smoking on morality

Collider Bias and Pearl's Classic Example

- We want to estimate the causal effect of D on Y, *i.e.*, the mapping $d \mapsto Y(d)$

$$
\begin{tikzcd}
Z_2 \arrow[r] \arrow[dr] & X_3 \arrow[r] & Y \\
 & X_2 \arrow[ur] \arrow[dr] & M \arrow[u] \\
Z_1 \arrow[r] \arrow[ur] & X_1 \arrow[r] & D \arrow[u]
\end{tikzcd}
$$

Collider Bias and Pearl's Classic Example

- We want to estimate the causal effect of D on Y, *i.e.*, the mapping $d \mapsto Y(d)$

$$
\begin{tikzcd}
Z_2 \arrow[r] \arrow[dr] & X_3 \arrow[r] & Y \\
& X_2 \arrow[ur] \arrow[dr] & M \arrow[u] \\
Z_1 \arrow[r] \arrow[ur] & X_1 \arrow[r] & D \arrow[u]
\end{tikzcd}
\qquad
\begin{tikzcd}
Z_2 \arrow[r] \arrow[dr] \arrow[dd, leftrightarrow, bend right] & X_3 \arrow[r] & Y(d) & \\
& X_2 \arrow[ur] \arrow[dr] & & M(d) \arrow[ul] \\
Z_1 \arrow[r] \arrow[ur] & X_1 \arrow[r] & D & d \arrow[u]
\end{tikzcd}
$$

Deeper Dive in DAG, Good and Bad Controls

Good and Bad Controls from DAG

- Sometimes we have causes of only treatment or only outcome
- In Scenario 1, including Z can reduce estimation variance; in scenario 2, including Z may increase estimation variance

$$\begin{tikzcd} Z \arrow[rr, bend right] & D \arrow[r] & Y \end{tikzcd} \qquad \begin{tikzcd} Z \arrow[r] & D \arrow[r] & Y \end{tikzcd} \qquad \begin{tikzcd} Z \arrow[r] & D \arrow[rr] & & Y \\ & & U \arrow[ul] \arrow[ur] & \end{tikzcd}$$

Scenario 1 | Scenario 2 | Scenario 3

Good and Bad Controls: Single Cause

- In scenario 3, adjusting for Z can exacerbate the bias stemming from unobserved confounding
- Controlling for Z removes exogeneous variation in the treatment D that is useful for identifying the causal effect but leaves the confounded variation
- The resulting estimated effect may be essentially driven by the unobserved confounder and be heavily biased

Good and Bad Controls: Single Cause

- In scenario 3, adjusting for Z can exacerbate the bias stemming from unobserved confounding
- Controlling for Z removes exogeneous variation in the treatment D that is useful for identifying the causal effect but leaves the confounded variation
- The resulting estimated effect may be essentially driven by the unobserved confounder and be heavily biased
- Indeed, variables like Z are known as *instrumental* variables
- These variables can be thought as inducing natural experiments that can be leveraged for causal identification in the presence of unobserved confounding
- Importantly, instruments *should not* be used in an identification by adjustment strategy

Good and Bad Controls: Post-Treatment Variables

- Explicitly adjusting for post-treatment variables is almost always a bad idea
- In many cases, post-treatment variables are included implicitly (and should be thought carefully) through *e.g.*, data and sample collection and variable definition
- For example, when estimating the effect of education on wages using data on *employed* individuals, we are implicitly conditioning on *employment*, which is a post-treatment variable and can lead to selection bias

Good and Bad Controls: Post-Treatment Variables

$$\begin{tikzcd} D \arrow[r] \arrow[rr, bend right] & M \arrow[r] & Y \end{tikzcd}$$

Bad Control

$$\begin{tikzcd} D \arrow[r] \arrow[rr, bend right] & M \arrow[r] \arrow[d] & Y \\ & Z & \end{tikzcd}$$

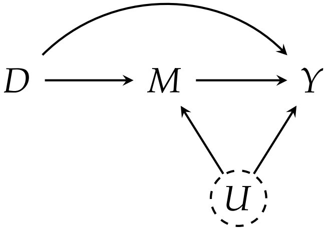
Bad Control

$$\begin{tikzcd} D \arrow[r] \arrow[rr, bend right] & M \arrow[r] & Y \\ & Z \arrow[u] & \end{tikzcd}$$

Neutral Control

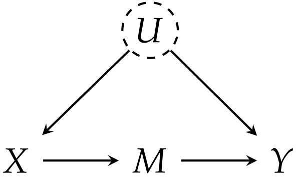
Good and Bad Controls: Post-Treatment Variables

- M is a bad control even for the *controlled direct effect*
- Outcome of the outcome is also a bad control



Front-Door Criterion

The Front-Door Criterion



- The frontdoor criterion has the following setup
  - All directed paths from X to Y go through M
  - There is *no* unblocked backdoor paths from X to M
  - All backdoor paths from M to Y are blocked by X

The Front-Door Criterion

- The causal effect of X on Y is then given by

$$
\begin{aligned}
P(Y(x)) = P(Y \mid do(X = x)) \\
&= \sum_m P(Y \mid do(X = x), M = m)\, P(M = m \mid do(X = x)) \\
&= \sum_m P(Y \mid do(M = m))\, P(M = m \mid X = x) \\
&= \sum_m P(M = m \mid X = x) \left[\sum_{x'} P(Y \mid M = m, X = x')\, P(X = x') \right]
\end{aligned}
$$

The Front-Door Criterion: The APC Problem

- Standard APC regression:

$$Y_{it} = \alpha + \beta_A A_i + \beta_P P_t + \beta_C C_i + \epsilon_{it}$$

  - Y_{it}: outcome for individual i in period t.
  - A_i: age of individual i
  - P_t: calendar period t
  - C_i: birth cohort of i ($C = P - A$)
- Because $C = P - A$, the three predictors are perfectly collinear; parameters $\beta_A, \beta_P, \beta_C$ are not separately identified

Mechanism-Based Identification using The Front-Door Criterion

- Without very strong assumption in APC identification, effect of historical period (X) on attitudes (Y) is unidentifiable
- However, we may use front-door criterion by finding *all* mechanisms where X affects Y
  - Period (P) affects exposure to new information, institutions, or policies (M)
  - These mechanisms then shape individual outcomes Y

$$P(Y \mid do(X)) = \sum_{m} P(M = m \mid X) \sum_{p'} P(Y \mid M = m, X = x')\, P(X = x')$$

Critiques of Mechanism-Based Identification (Front-Door)

- Complete mediation is unrealistic; it requires that all effects of X on Y pass through the observed mechanism M; in practice, period or policy may affect outcomes via multiple unmeasured channels
- No hidden confounding is a strong assumption; we assume $X \to M$ is unconfounded and that $M \to Y$ can be identified by adjusting for X