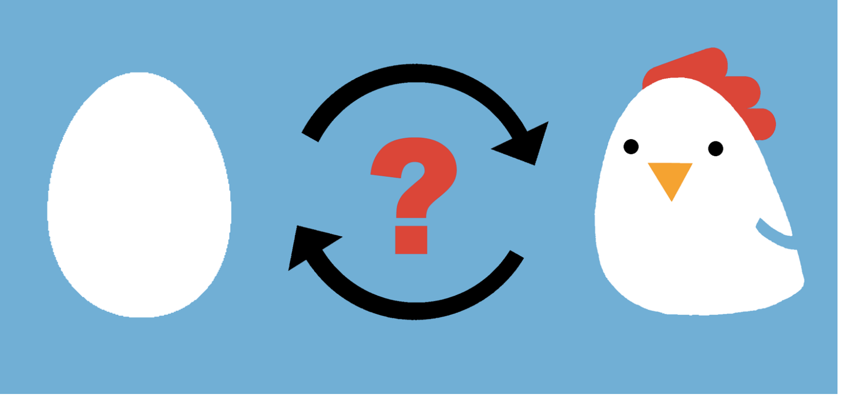


CBMS Conference -- Foundations of Causal Graphical Models and Structure Discovery

Lecture 4

Identification of Causal Effects & Counterfactual Inference

Instructor: Kun Zhang

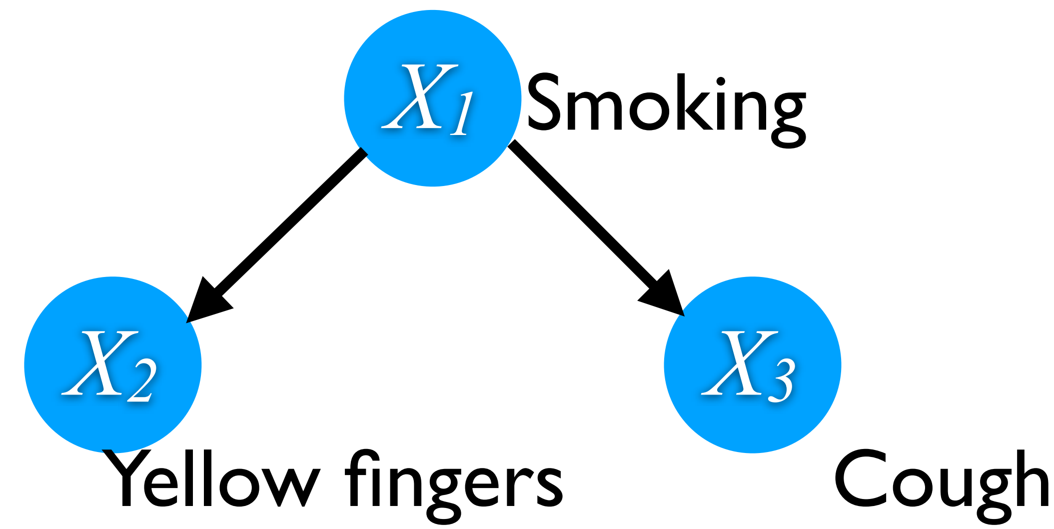
Carnegie Mellon University

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Identification of Causal Effects & Counterfactual Inference: Outline

- Problem definition
- Potential outcome framework
  - Propensity score
- Backdoor criterion and front door criterion
- Counterfactual inference

Three Types of Problems in Current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
…	…	…

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X3 \mid do\ (X2=1))$$

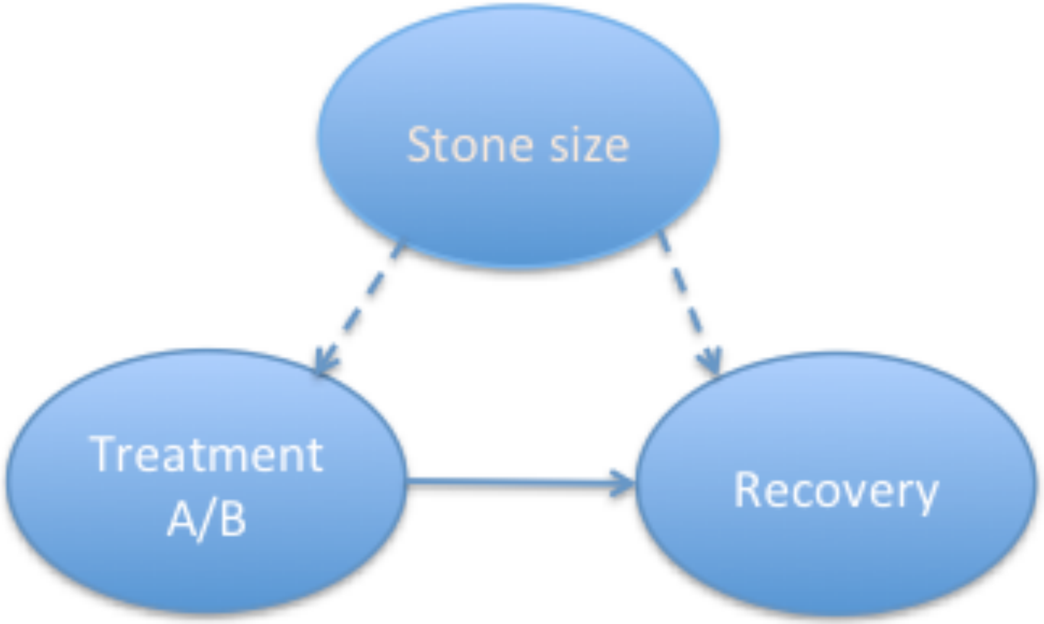
- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X3_{X2=1} \mid X2 = 0, X3 = 1)$$

Identification of Causal Effects

P(Recovery | do(Treatment=A)) ?

- "Golden standard": randomized controlled experiments
  - **All the other factors** that influence the outcome variable are either fixed or vary at random, so any changes in the outcome variable must be due to the controlled variable



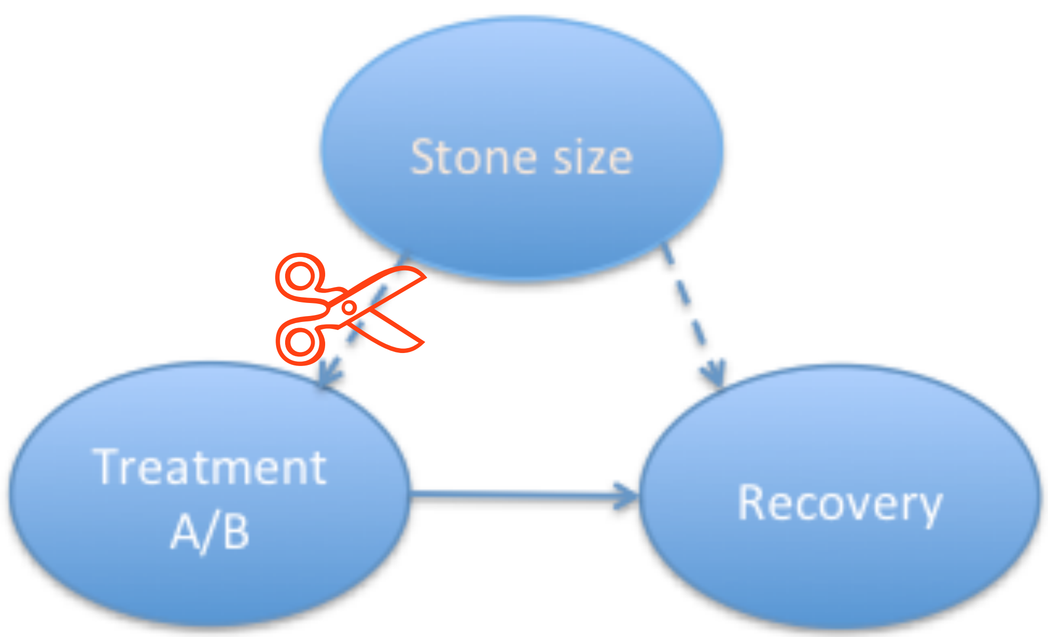
  - Usually expensive or impossible to do!

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

$$P(R|T) = \sum_S P(R|T,S)P(S|T)$$

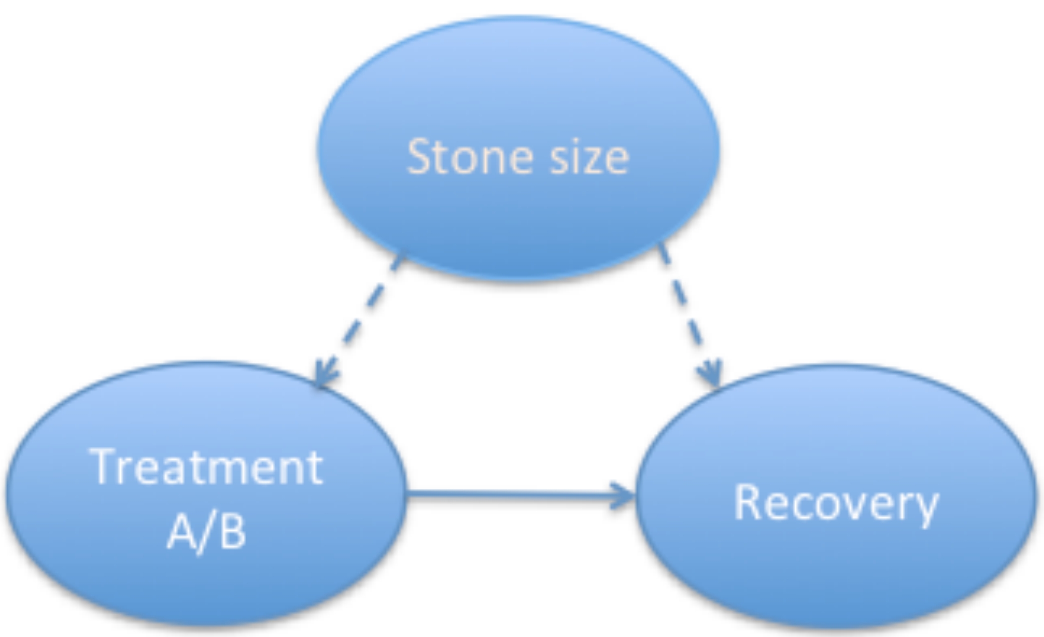
$$P(R\,|\,do(T)) = \sum_S P(R\,|\,T,S)P(S)$$

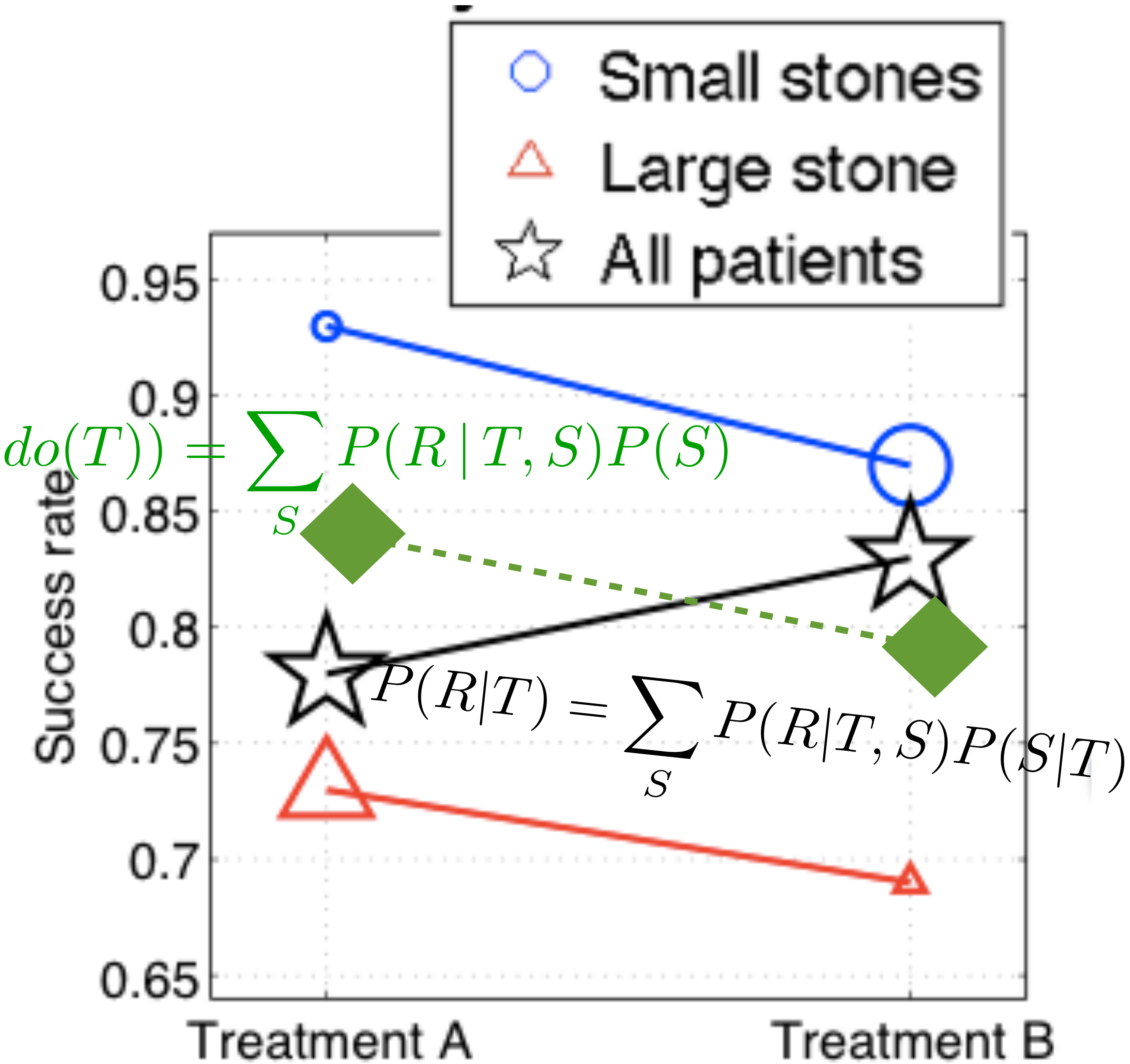


conditioning vs. manipulating

Identification of Causal Effects: Example

	Treatment A	Treatment B
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**





$$P(R \mid do(T)) = \sum_S P(R \mid T, S) P(S)$$

$$P(R|T) = \sum_S P(R|T,S) P(S|T)$$

conditioning vs. manipulating

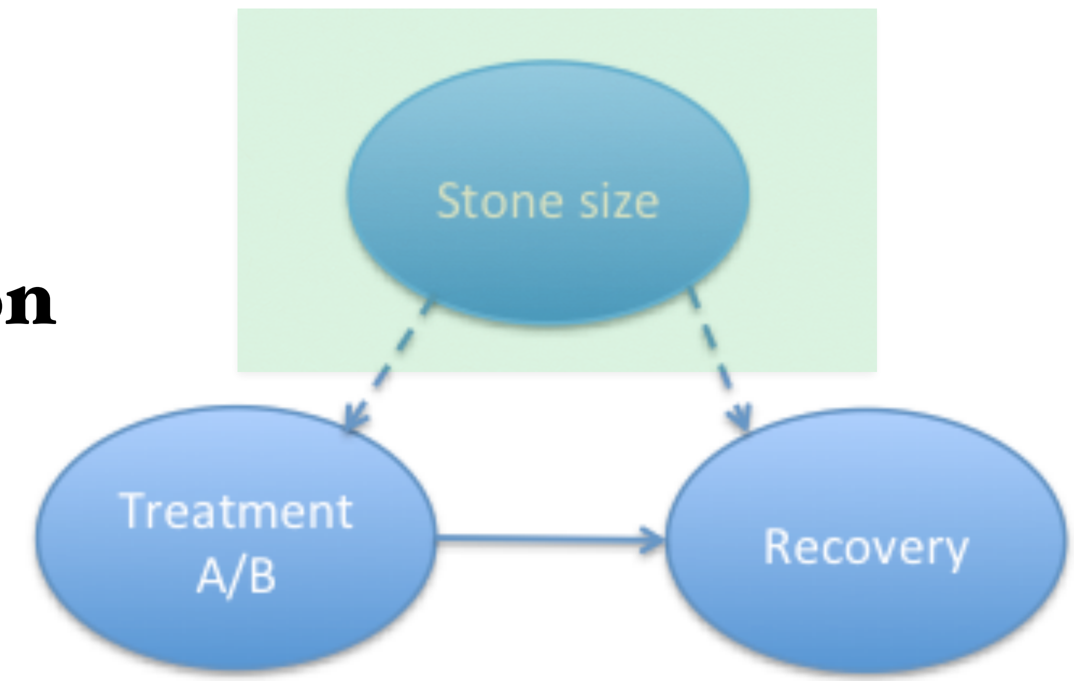
Identifiability of Parameters in Statistical Models

- Identifiability, in simple words, means that different values of a parameter must produce different probability distributions.
- Mathematically, a parameter θ is said to be identifiable if and only
$\theta \neq \theta' \Rightarrow P_\theta \neq P_{\theta'}$, or equivalently $P_\theta = P_{\theta'} \Rightarrow \theta = \theta'$
- Is the mean of a Gaussian distribution identifiable?

Identifiability of Causal Effects

Sometimes written as $P(y \mid \hat{x})$

- Is causal effect, denoted by $P(Y \mid do(X))$, identifiable given complete or partial causal knowledge?
  - Two models with **the same causal structure** and **the same distribution for the observed variables** give the same causal effect?



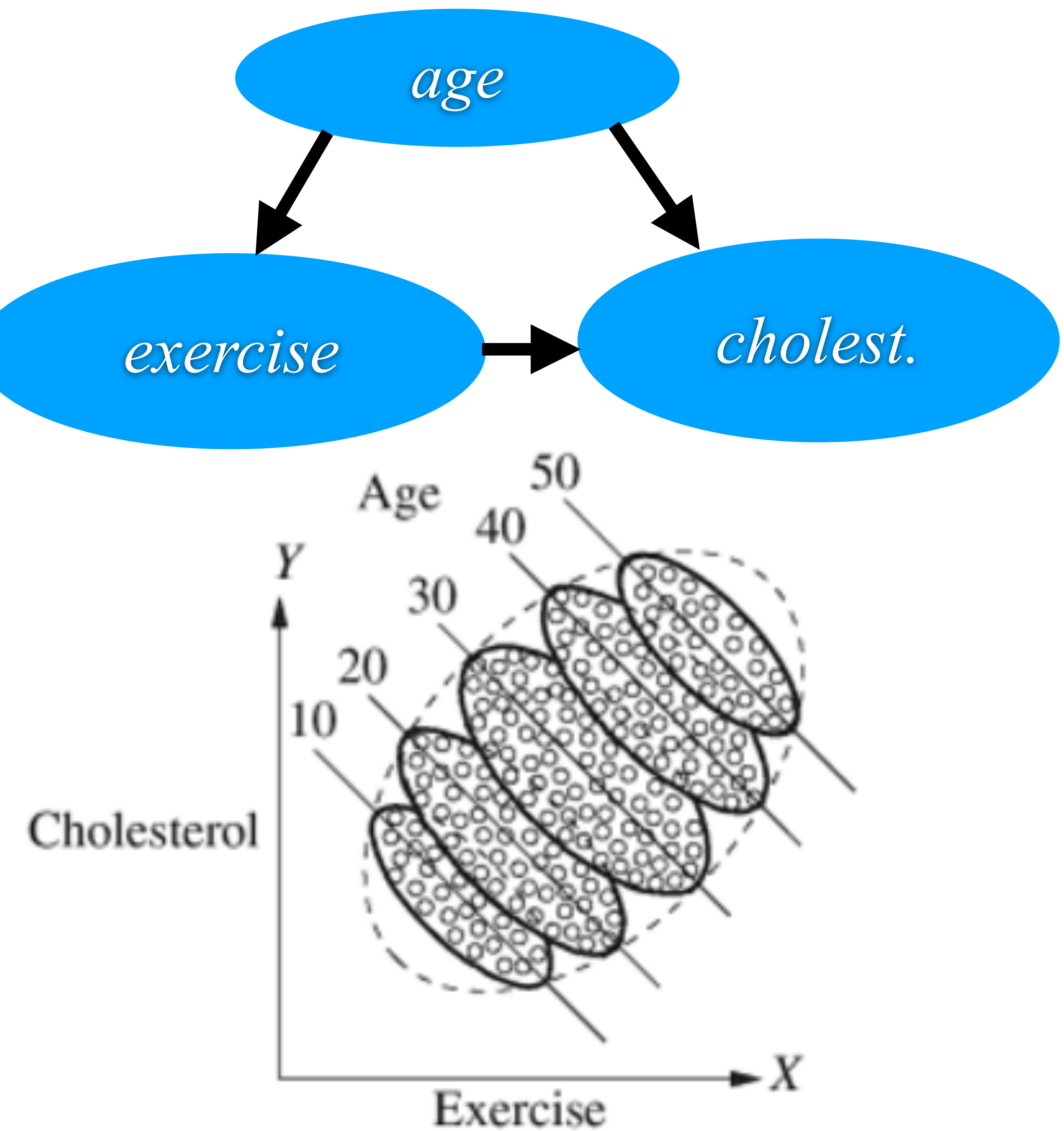
- How?

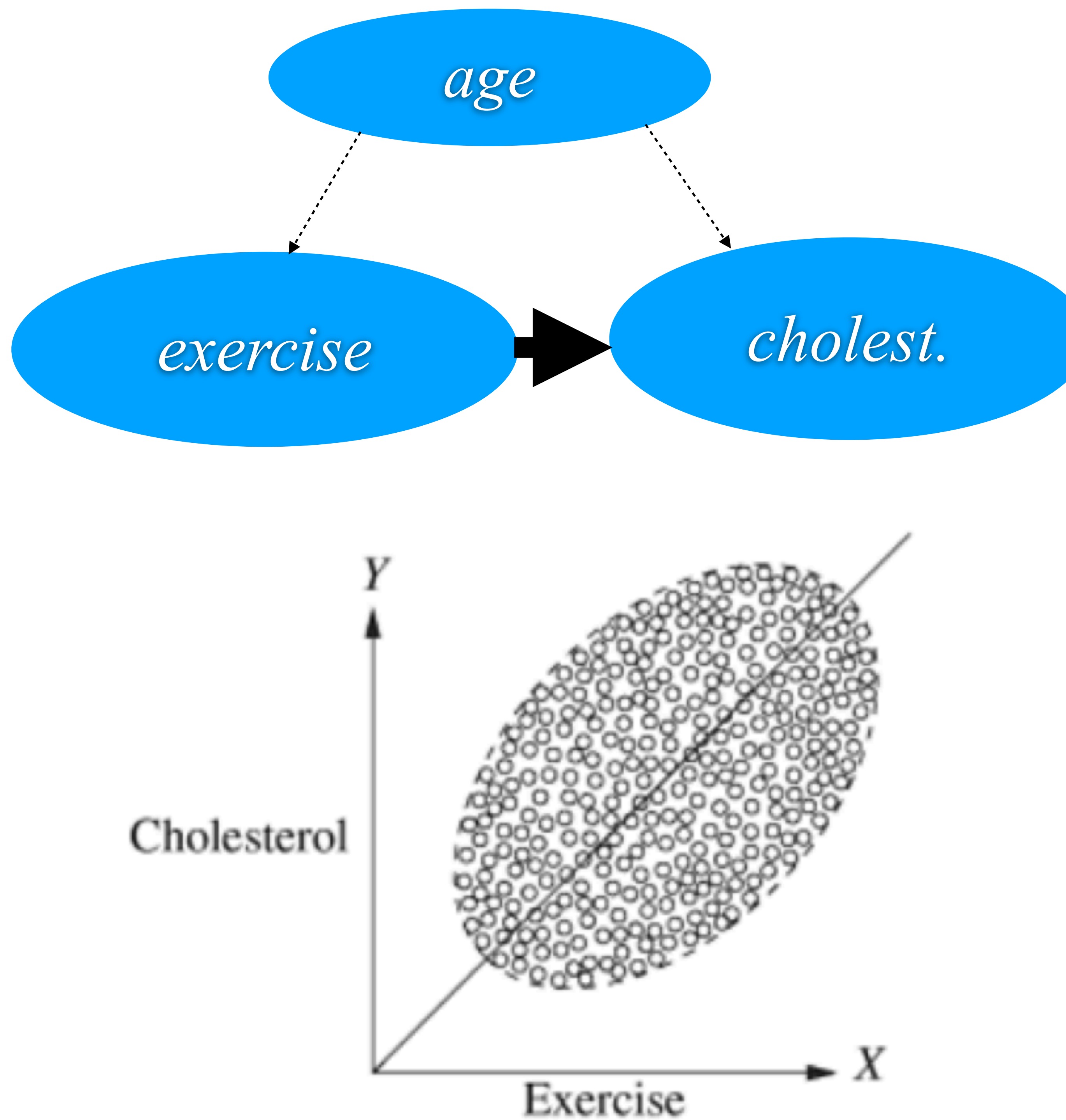
- Key issue: Controlling confounding effects

Examples: Average causal effect (ACE)...

Key Issue: Controlling Confounding Bias

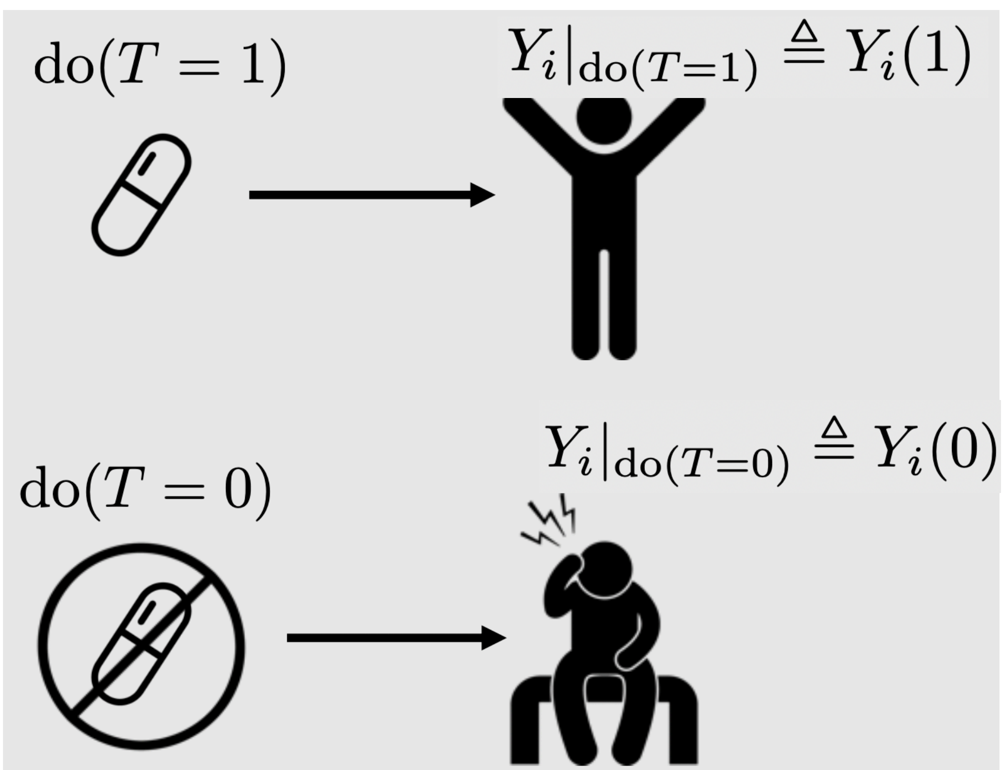
- Exercise-cholesterol study: identifiable if age is not observed?





Potential Outcome

- Causal inference: Inferring the effect of treatment/ policy on some outcome



Causal effect:
$Y_i(1)$ - $Y_i(0)$

T: observed treatment

Y: observed outcome

i: denote a specific subject or unit

$Y_i(1)$: **potential outcome if the patient had been treated**

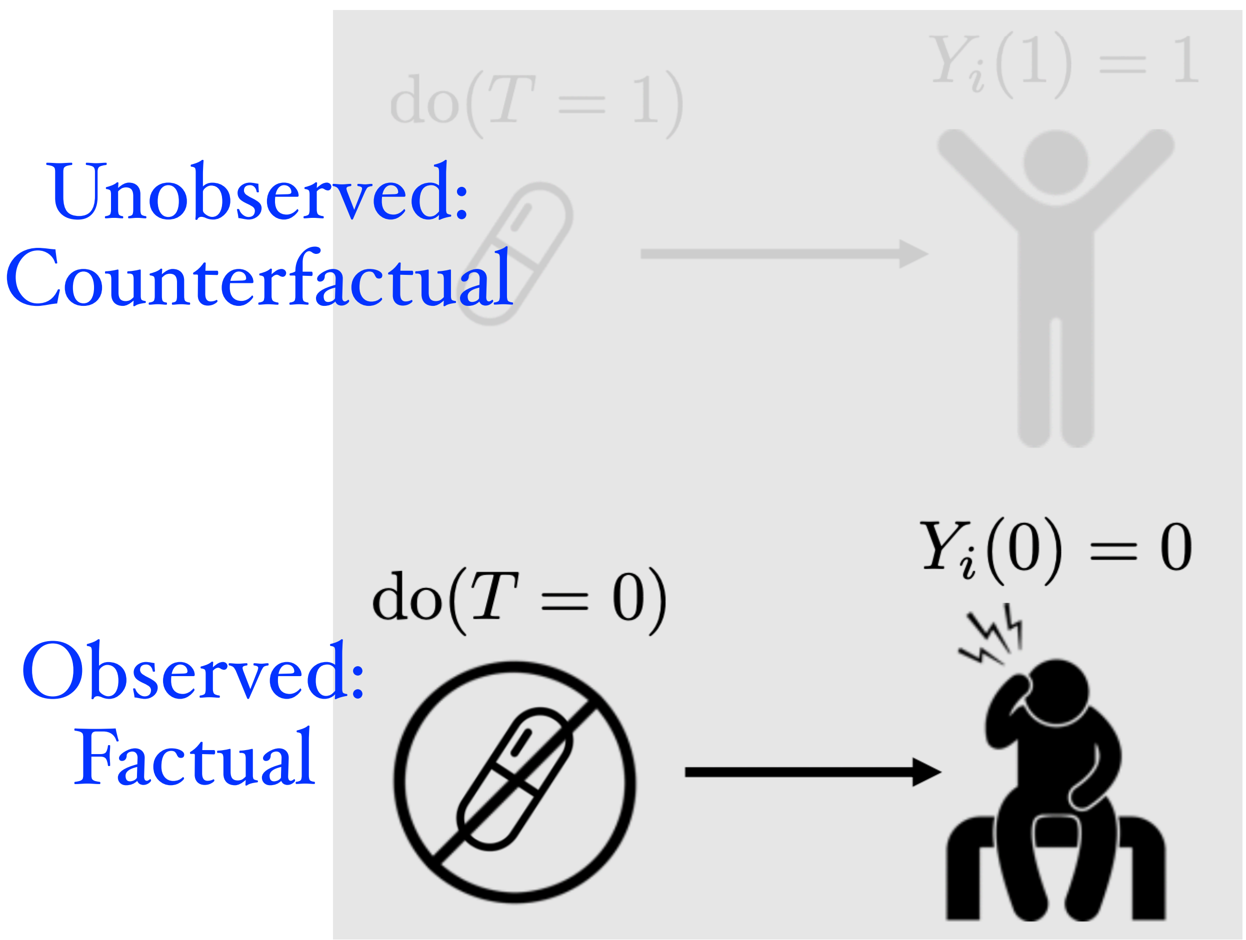
$$Y_{i|do(T=1)} \triangleq Y_i(1)$$

$Y_i(0)$: **potential outcome if the patient had not been treated**

$$Y_{i|do(T=0)} \triangleq Y_i(0)$$

Fundamental Problem of Causal Inference

- Missing data issue



T: observed treatment

Y: observed outcome

i: denote a specific subject or unit

$Y_i(1)$: potential outcome under treatment

$$Y_{i|do(T=1)} \triangleq Y_i(1)$$

$Y_i(0)$: potential outcome without treatment

$$Y_{i|do(T=0)} \triangleq Y_i(0)$$

Causal effect:

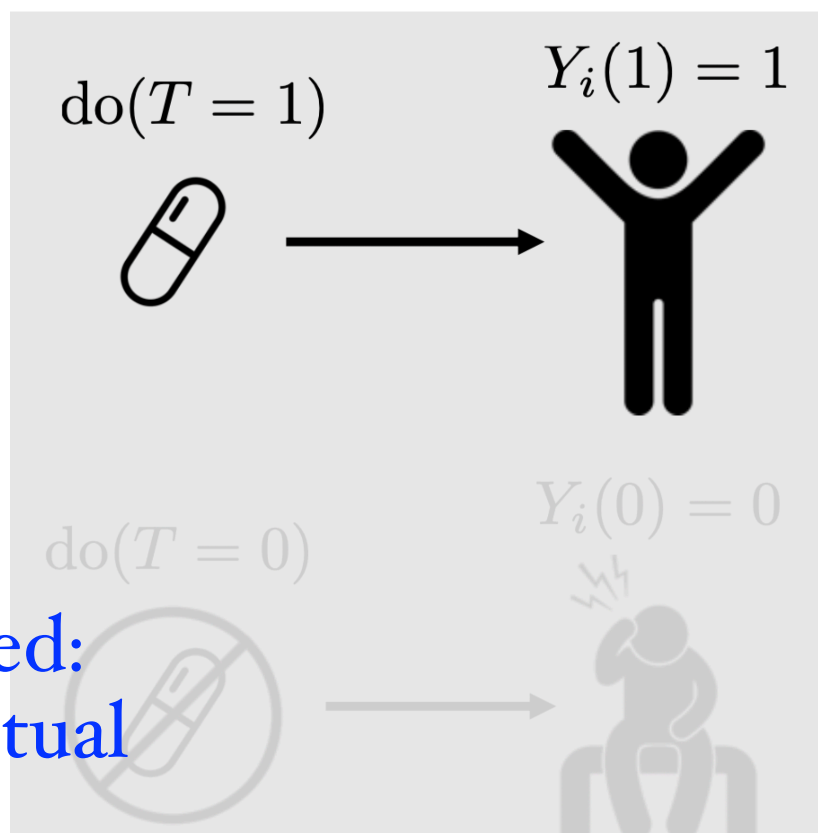
$Y_i(1) - Y_i(0)$

Fundamental Problem of Causal Inference

- Missing data issue

Observed: Factual

Unobserved: Counterfactual



T: observed treatment

Y: observed outcome

i: denote a specific subject or unit

$Y_i(1)$: potential outcome under treatment

$$Y_{i|do(T=1)} \triangleq Y_i(1)$$

$Y_i(0)$: potential outcome without treatment

$$Y_{i|do(T=0)} \triangleq Y_i(0)$$

Causal effect:
$Y_i(1)$ - $Y_i(0)$

Fundamental Problem of Causal Inference

- Missing data issue

i	T	Y	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?

T: observed treatment

Y: observed outcome

i: denote a specific subject or unit

$Y_i(1)$: potential outcome under treatment

$$Y_{i|do(T=1)} \triangleq Y_i(1)$$

$Y_i(0)$: potential outcome without treatment

$$Y_{i|do(T=0)} \triangleq Y_i(0)$$

Causal effect:
$Y_i(1) - Y_i(0)$

Potential Outcome Framework

- For a set of i.i.d. subjects $i = 1, \cdots, n$, we observed a tuple (X_i, T_i, Y_i), comprised of
  - A **feature vector** $X_i \in \mathbb{R}^p$
  - A **treatment assignment** $T_i \in \{0,1\}$
  - A **response** $Y_i \in \mathbb{R}$

- $Y_i(1)$ and $Y_i(0)$ are **potential outcomes** in that they represent the outcomes for individual i had they received the treatment or control respectively.

- Missing data issue: we only get to see Y_i, with

$$Y_i = Y_i(T_i) = Y_i(0)(1 - T_i) + Y_i(1)T_i$$

Potential Outcome Framework

- Our first goal is to estimate the **average treatment effect (ATE)**

$$\tau = E[Y_i(1) - Y_i(0)]$$

- However, we cannot find $Y_i(1) - Y_i(0)$ because of the unobserved potential outcome

- Then what assumptions do we need in order to estimate ATE from observational data?

Assumptions in the Potential-Outcome Framework

Assumptions that make the ATE be estimated from observational data

- Ignorability: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i$

  Conditional ignorability: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$

- Positivity: $0 < P(T = 1 \mid X = x) < 1$

- No interference: $Y_i(t_1, \cdots, t_{i-1}, t_i, t_{i+1}, \cdots, t_n) = Y_i(t_i)$

- Consistency: $T = t \Longrightarrow Y = Y(t)$

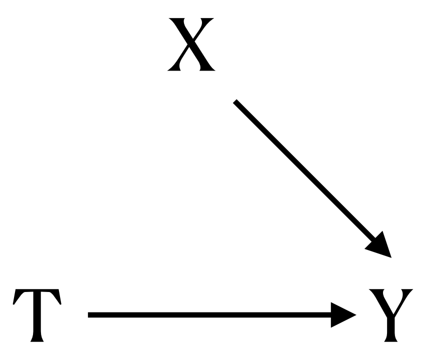
} Stable Treatment Value Assumption (SUTVA)

Assumption 1: Ignorability

- The ignorability assumption: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i$

  That is, the potential outcomes of subjects had they been treated or not does not depend on whether they have really been (observable) treated or not

- Corresponding graphical model: there is no other path from T to Y, except the direct edge



- ATE $= E[Y(1)] - E[Y(0)]$

$$= E[Y(1) \mid T = 1] - E[Y(0) \mid T = 0] \quad \textit{(Ignorability)}$$

$$\boxed{= E[Y \mid T = 1] - E[Y \mid T = 0]} \quad \textit{(Consistency)}$$

Only contains observable moments

Assumption 1: Ignorability

- The ignorability assumption: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i$

$$E[Y(1)] - E[Y(0)] = E[Y(1) \,|\, T = 1] - E[Y(0) \,|\, T = 0]$$

$$= E[Y \,|\, T = 1] - E[Y \,|\, T = 0]$$

$$= 100.52 - 100.59$$

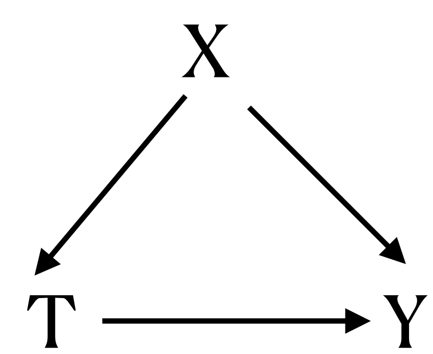
$Y_i(0)$	$Y_i(1)$	τ_i
154.68	—	—
135.67	—	—
—	117.68	—
—	95.08	—
—	146.73	—
117.89	—	—
—	75.59	—
—	65.68	—
100.07	—	—
—	82.30	—
...	...	...
110.59	100.52	—

Assumption 1: Conditional ignorability

- The conditional ignorability assumption: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$

  That is, given the covariates, the potential outcomes of subjects had they been treated or not does not depend on whether they have really been (observable) treated or not

- Corresponding graphical model: X blocks all paths from T to Y, except the direct edge



- Conditional average treatment effect:

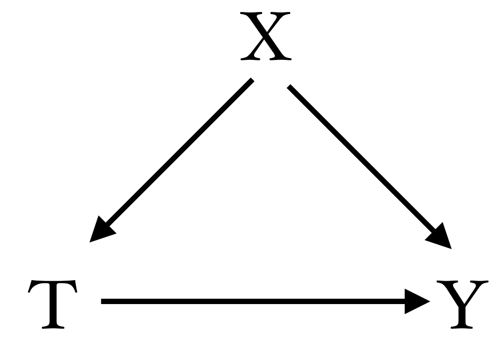
$$CATE = E[Y(1) - [Y(0) \mid X]$$
$$= E[Y(1) \mid X] - E[Y(0) \mid X]$$
$$= E[Y(1) \mid T = 1, X] - E[Y(0) \mid T = 0, X] \quad \textit{(Conditional ignorability)}$$
$$= \boxed{E[Y \mid T = 1, X] - E[Y \mid T = 0, X]} \quad \textit{(Consistency)}$$

Only contains observable moments

From CATE to ATE

- Adjustment formula to identifying ATE

$$
\begin{aligned}
ATE &= E[Y_i(1) - Y_i(0)] \\
&= E_X E[Y_i(1) - Y_i(0) \mid X_i] \\
&= E_X\big[E[Y_i \mid T_i = 1, X_i] - E[Y_i \mid T_i = 0, X_i]\big]
\end{aligned}
$$



Assumption 2: Positivity

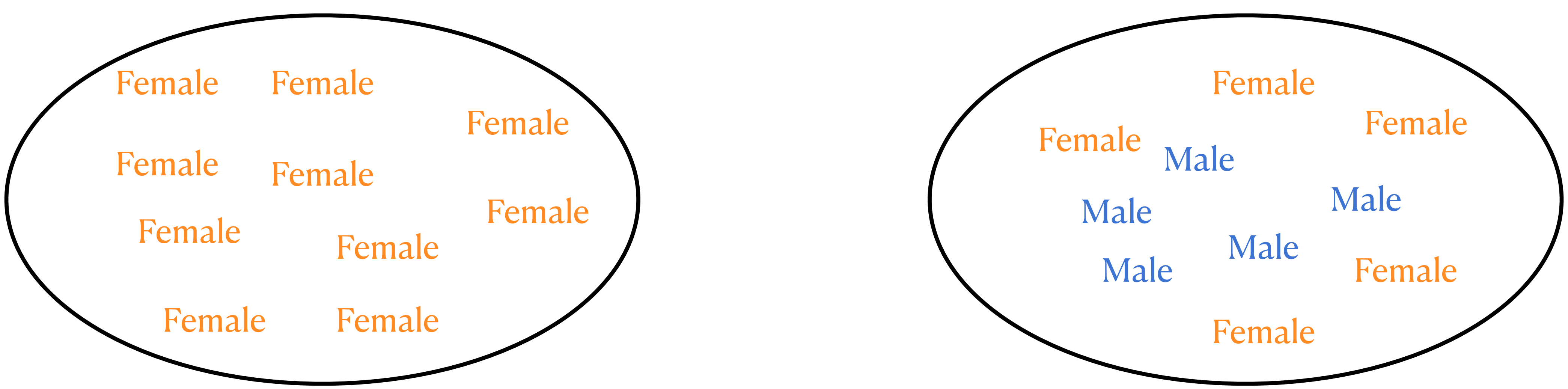
- The positivity assumption

  For all values of covariates x present in the population of interest (i.e., x such that $P(X = x) > 0$),

$$0 < \underline{P(T = 1 \mid X = x)} < 1$$

- A case where the positivity assumption violates

T = 0 Suppose X = {Female, Male} T = 1



Assumption 3: No Interference

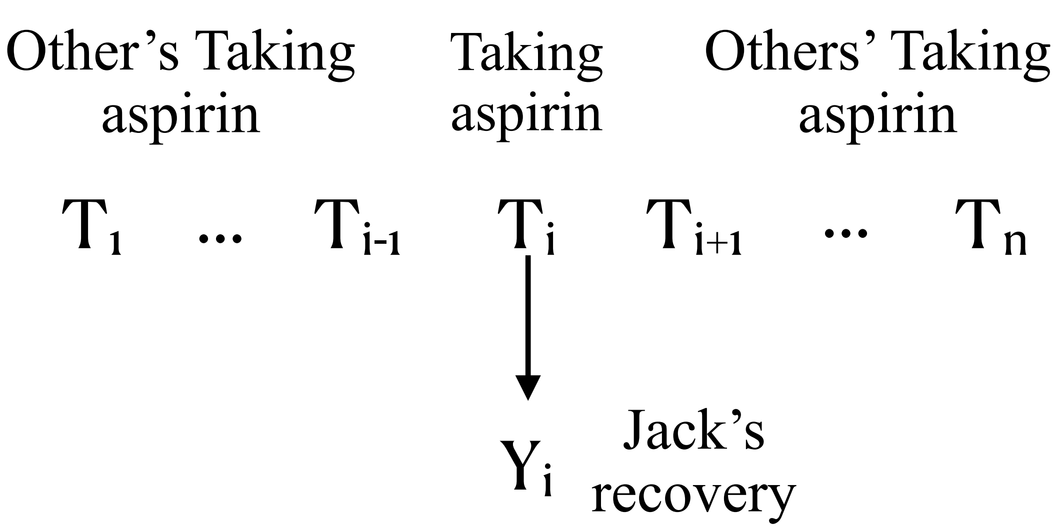
- The no interference assumption: treatments of other units do not affect one's potential outcome, so

$$Y_i(t_1, \cdots, t_{i-1}, t_i, t_{i+1}, \cdots, t_n) = Y_i(t_i)$$

That is, unit i's potential outcome is only a function of its own treatment, but will not be affected by other units' treatment

- *A case where the assumption holds:*

Jack's recovery is not affected by others' taking aspirin.



- *Violation:*

Job training for too many people may flood the market with qualified job applicants (interference)

Assumption 4: Consistency

- The consistency assumption: the potential outcome under treatment T=t, Y(t), is equal to the observed outcome if the actual treatment received is T=t, i.e.,

$$T = t \implies Y = Y(t), \text{ for all } t$$

That is, the observed outcome is equal to the potential outcome Y(t), when the actual treatment received is T =t; there is no variation in treatment

	$T = 1$ "I get a dog"	$T = 0$ "I don't get a dog"

$(T = 1) \implies Y = 1$ (I'm happy)

$(T = 1) \implies Y = 0$ (I'm not happy)

} Consistency assumption violated

(Adapted from Brady Neal, 2020)

Recall the Assumptions

Assumptions that make the ATE be estimated from observational data

- Ignorability: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$

  Conditional ignorability: $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$

- Positivity: $0 < P(T = 1 \mid X = x) < 1$

- No interference: $Y_i(t_1, \cdots, t_{i-1}, t_i, t_{i+1}, \cdots, t_n) = Y_i(t_i)$

- Consistency: $T = t \implies Y = Y(t)$

} Stable Unit Treatment Value Assumption (SUTVA)

Stable Unit Treatment Value Assumption (SUVTA): No interference assumption + Consistency assumption

SUVTA allows to write potential outcome for the i^{th} person in terms of only that person's treatments

Derivation of ATE

No interference: ↙ ↘

$$ATE = E[Y(1) - Y(0)] = E[Y(1)] - E[Y(0)]$$ (Linearity of expectation)

$$= E_X[E[Y(1) \mid X] - E[Y(0) \mid X]]$$ (Law of iterated expectations)

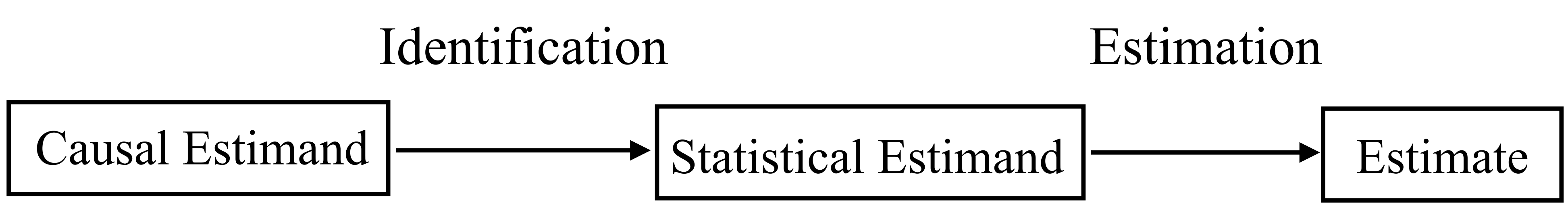
$$= E_X[E[Y(1) \mid T = 1, X] - E[Y(0) \mid T = 0, X]]$$ (Ignorability and Positivity)

$$= E_X[E[Y \mid T = 1, X] - E[Y \mid T = 0, X]]$$ (Consistency)

Estimands, Estimates, and Estimation

- Estimand: any quantity we want to estimate
  - Causal estimand (e.g. $E[Y(1) - Y(0)]$
  - Statistical estimand (e.g. $E_X[E[Y \mid T = 1, X] - E[Y \mid T = 0, X]]$)
- Estimate: approximation of some estimand, using data
- Estimation: process for getting from data + estimatand to estimate

The Identification-Estimation Flowchart



Example: Effect of Sodium Intake on Blood Pressure

Data (Epidemiological example taken from Luque-Fernandez et al. (2018)):

- Outcome Y: (systolic) blood pressure (continuous)
- Treatment T: sodium intake (1 if above 3.5 mg and 0 if below)
- Covariates X: age and amount of protein excreted in urine
- Simulation: so we know the "true" ATE is 1.05

Estimation of ATE

True ATE: $\mathbb{E}[Y(1) - Y(0)] = 1.05$

Identification: $\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X\left[\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]\right]$

Estimation: $\frac{1}{n}\sum_x \left[\underbrace{\mathbb{E}[Y \mid T = 1, x]} - \underbrace{\mathbb{E}[Y \mid T = 0, x]}\right]$

Model (linear regression)

Estimate: 0.85 $\quad \frac{|0.85 - 1.05|}{1.05} \times 100\% = 19\%$

Naive: $\mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0]$

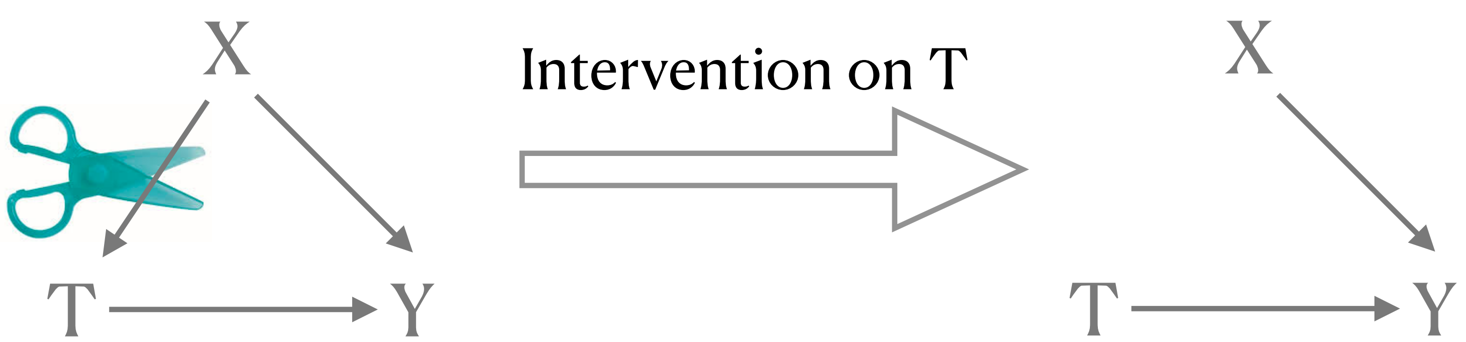
Naive estimate: 5.33 $\quad \frac{|5.33 - 1.05|}{1.05} \times 100\% = 407\%$

(Adapted from Brady Neal, 2020)

How to Estimate Causal Effect With Confounders?

1) Randomization

$$E[Y(1) - Y(0)] = E[Y \mid T = 1] - E[Y \mid T = 0]$$



2) Statistical adjustment

$$ATE = E_X[E[Y \mid T = 1, X] - E[Y \mid T = 0, X]]$$

Covariates Adjustments

$$ATE = E_X[E[Y \mid T = 1, X] - E[Y \mid T = 0, X]]$$

- Regression adjustments
- Matching
  - Mahalanobis distance matching
  - Propensity Score matching
- Inverse propensity score reweighting
- Doubly robust method

Covariates Adjustments

$$ATE = E_X[E[Y|T=1,X] - E[Y|T=0,X]]$$

- **Regression adjustments**
- Matching
  - Mahalanobis distance matching
  - Propensity Score matching
- Inverse propensity score reweighting
- Doubly robust method

Regression Adjustments

- Regression adjustments under ignorability / unconfoundedness

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i$$

- We can express the ATE in terms of conditional response,

$$ATE = E[Y_i(1) - Y_i(0)] = E[Y_i(1)] - E[Y_i(0)]$$

$$= E[E[Y_i(1) \mid X_i] - E[Y_i(0) \mid X_i]]$$

$$= E[E[Y_i(1) \mid T_i = 1, X_i] - E[Y_i(0) \mid T_i = 0, X_i]]$$

$$= E[E[Y_i \mid T_i = 1, X_i] - E[Y_i \mid T_i = 0, X_i]]$$

$$= E[\mu_{(1)}(X_i)] - E[\mu_{(0)}(X_i)]$$

where $\mu_{(t)}(x) = E[Y_i \mid T_i = t, X_i = x]$

Regression Adjustments

- Given ignorability, we have $\tau = E[\mu_{(1)}(X_i)] - E[\mu_{(0)}(X_i)]$,

  with $\mu_{(t)}(x) = E[Y_i \mid X_i = x, T_i = t]$

  - Fit $\hat{\mu}_t(x)$ via linear regression

  - Fit $\hat{\mu}_t(x)$ via non-parametric approach

- One may use the following estimation strategy

  1. Learn $\hat{\mu}_0(x)$ by predicting Y from X on controls
  2. Learn $\hat{\mu}_1(x)$ by predicting Y from X on treated units
  3. Estimate $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$

$\hat{\tau}$ is consistent if $\hat{\mu}_t(x)$ is consistent for $\mu_t(x)$…

Covariates Adjustments

$$ATE = E_X[E[Y|T=1,X] - E[Y|T=0,X]]$$

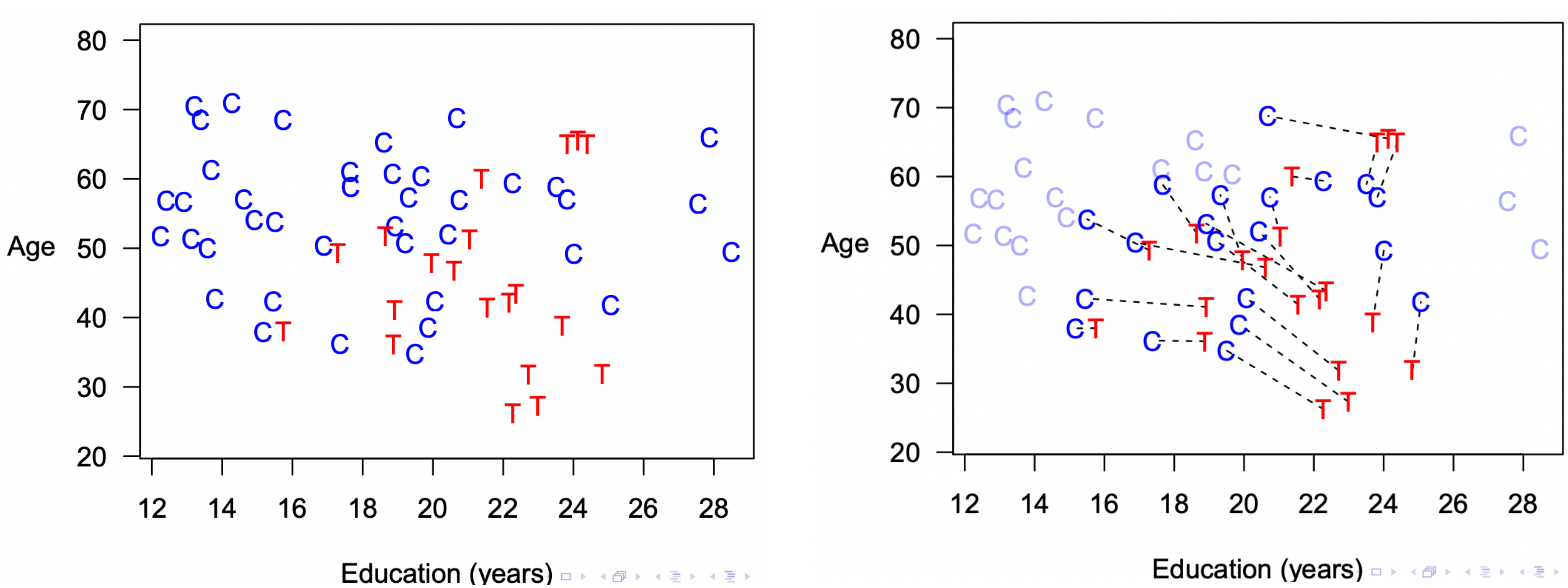
- Regression adjustments
- **Matching**
    - Mahalanobis distance matching
    - Propensity Score matching
- Inverse propensity score reweighting
- Doubly robust method

Matching I: Mahalanobis Distance Matching

- **Mahalanobis distance matching**: match the feature of each treated unit to the nearest control unit, with the distance

$$D(X_i, X_j) = \sqrt{((X_i - X_j)^T S^{-1}(X_i - X_j))}$$

  - Control units: pruned if unused
  - Prune matches if distance > threshold



Propensity Score

- The propensity score measures the probability of being treated conditionally on X_i , i.e.,

$$e(x) = P(T_i = 1 \mid X_i = x)$$

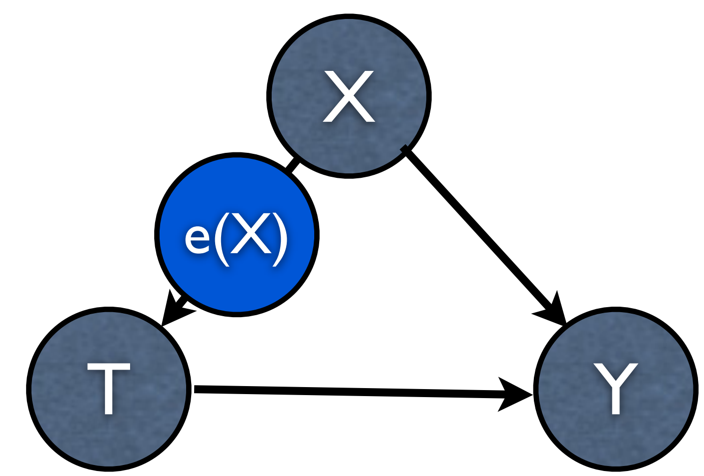
- In a randomized trial, the propensity score is constant

$$e(x) = e_0 \in (0,1)$$

  - At least qualitatively, the variability of the propensity score gives a measure of how far we are from a randomized trial

Matching 2: Propensity Score Matching

- One way is to match covariates *X, but it is hard especially for high-dimensional X*
- Propensity Score
  - Let $e(X) = P(T=1 \mid X)$; $T \perp\!\!\!\perp X \mid e(X)$
  - Then $e(X)$ and X are (confounding)-equivalent
  - $\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i \Longrightarrow \{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid e(X_i)$
    - Unconfoundness given X implies unconfoundness given $e(X)$
    - X may be high-dimensional, while $e(X)$ is one-dimension



Matching 2: Propensity Score Matching

- Propensity Score
  - Let $e(X) = P(T{=}1 \mid X)$; $T \perp\!\!\!\perp X \mid e(X)$ — *The probability of T=1, given X*
  - Then $e(X)$ and X are (confounding)-equivalent:

$$\sum_x P(Y \mid t, x)P(x) = \sum_x \sum_e P(Y \mid t, x)P(e)P(x \mid e)$$

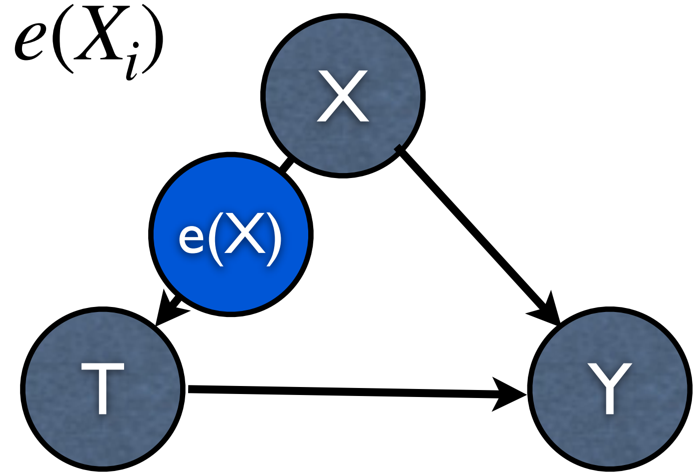
$$= \sum_x \sum_e P(Y \mid t, x, e)P(e)P(x \mid t, e) = \sum_x \sum_e P(Y, x \mid t, e)P(e)$$

$$= \sum_e P(Y \mid t, e)P(e)$$

Matching 2: Propensity Score Matching

- Unconfoundness given X implies unconfoundness given $e(X)$

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid X_i \;\Longrightarrow\; \{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \mid e(X_i)$$



- If directly matching on X, overlap decreases with he dimensionality of the adjustment set

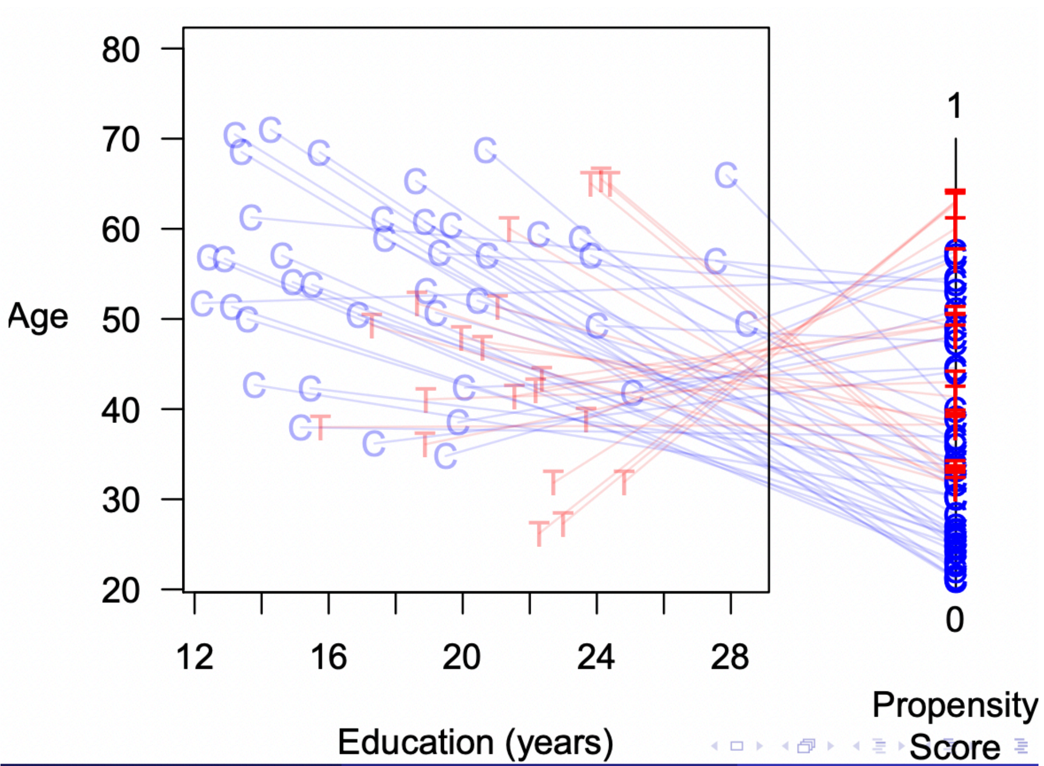
- The propensity score magically reduces the dimensionality of the adjustment set to 1!

- However, we do not have access to the propensity score.
  - The best we can do is to model it, e.g., with logistic regression, shifting the high-dimensionality problem to the modeling of $e(X)$

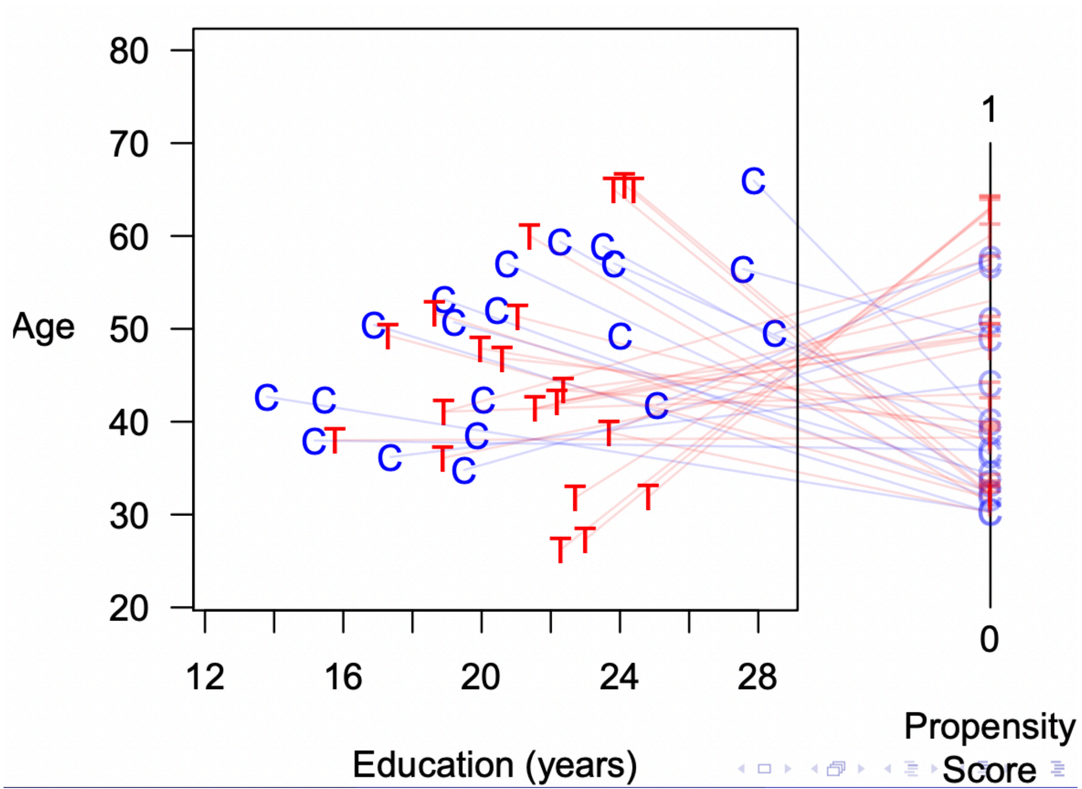
Matching 2: Propensity Score Matching

General procedures of propensity score matching:

1. Estimate propensity scores *c*(*X*) = *P*(*T*=*1* | *X*), e.g. with logistic regression
2. Match each treated to the nearest untreated on propensity score
   - Nearest neighbor matching
   - Optimal full matching …



Estimate propensity scores



Matching

Matching 2: Propensity Score Matching

Questions:

- What is the intuition behind why we can condition on $e(X)$, instead of X?
- What is attractive about conditioning on $e(X)$, instead of X?
- Why does this not really solve the positivity issue when X is high-dimensional?

Identification of Causal Effects & Counterfactual Inference: Outline

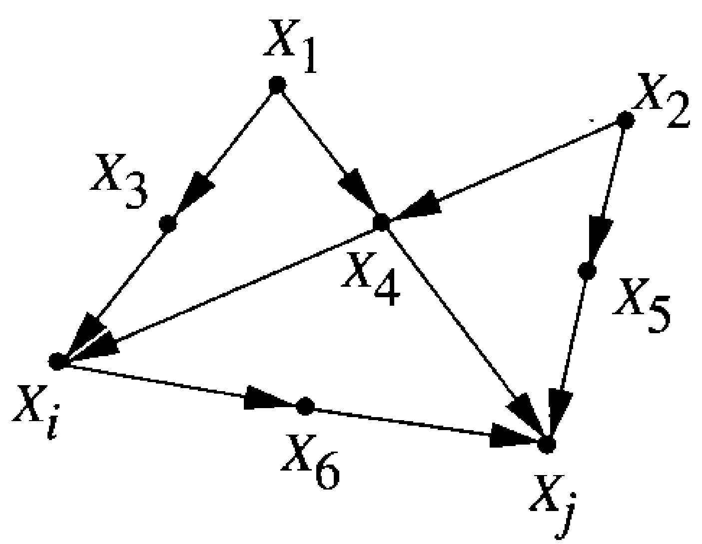
- Problem definition
- Potential outcome framework
  - Propensity score
- Backdoor criterion and front door criterion
- Counterfactual inference

Graphical Criterion: Back-Door Criterion

Definition 3.3.1 (Back-Door)
A set of variables Z satisfies the back-door *criterion relative to an ordered pair of variables* (X_i, X_j) *in a DAG G if:*

(i) *no node in* Z *is a descendant of* X_i; *and*

(ii) Z *blocks every path between* X_i *and* X_j *that contains an arrow into* X_i.



- What if $Z = \{X_3, X_4\}$?
  $Z = \{X_4, X_5\}$?
  $Z = \{X_4\}$?
- What if there is a confounder?

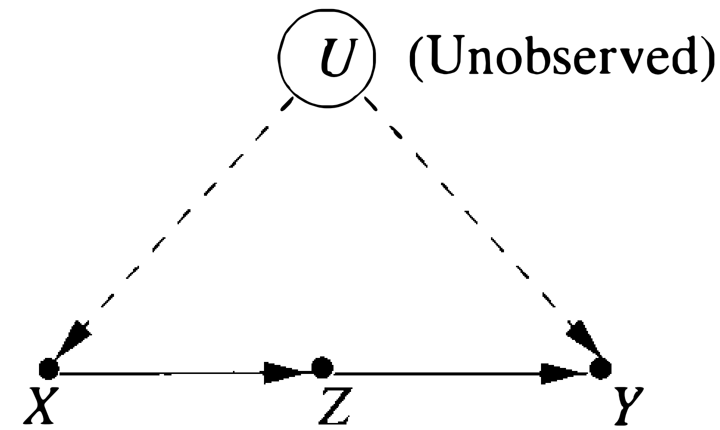
Theorem 3.3.2 (Back-Door Adjustment)
If a set of variables Z satisfies the back-door criterion relative to (X, Y), *then the causal effect of X on Y is identifiable and is given by the formula*

$$\boxed{P(y \mid \hat{x})} = \sum_z P(y \mid x, z) P(z).$$

Or P(Y=y | do(X=x))

*

Front-Door Criterion



Definition 3.3.3 (Front-Door)

A set of variables Z is said to satisfy the front-door *criterion relative to an ordered pair of variables (X, Y) if:*

(i) *Z intercepts all directed paths from X to Y;*

(ii) *there is no back-door path from X to Z; and*

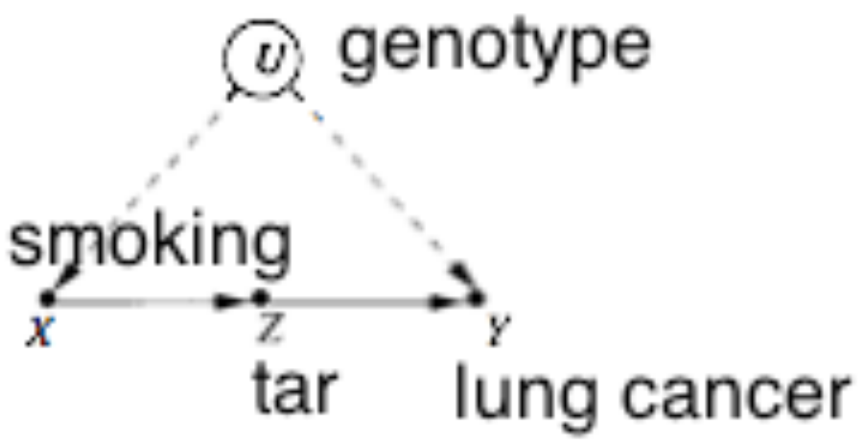
(iii) *all back-door paths from Z to Y are blocked by X.*

Theorem 3.3.4 (Front-Door Adjustment)

If Z satisfies the front-door criterion relative to (X, Y) and if $P(x, z) > 0$, then the causal effect of X on Y is identifiable and is given by the formula

$$P(y \mid \hat{x}) = \sum_{z} P(z \mid x) \sum_{x'} P(y \mid x', z) P(x'). \tag{3.29}$$

* Example: Smoking & Genotype Theory



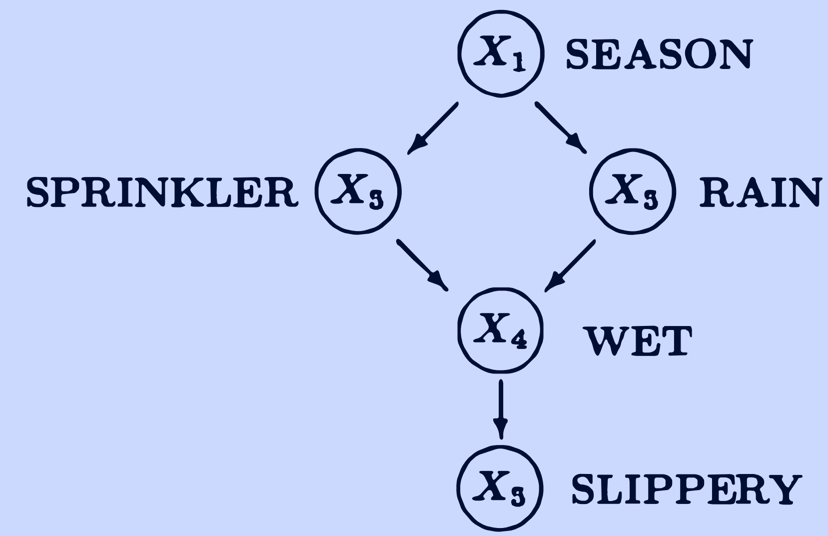
	Group Type	$P(x, z)$ Group Size (% of Population)	$P(Y = 1 \mid x, z)$ % of Cancer Cases in Group
$X = 0,\ Z = 0$	Nonsmokers, No tar	47.5	10
$X = 1,\ Z = 0$	Smokers, No tar	2.5	90
$X = 0,\ Z = 1$	Nonsmokers, Tar	2.5	5
$X = 1,\ Z = 1$	Smokers, Tar	47.5	85

$$\begin{aligned} P(Y = 1 \mid do(X = 1)) &= .05(.10 \times .50 + .90 \times .50) \\ &\quad + .95(.05 \times .50 + .85 \times .50) \\ &= .05 \times .50 + .95 \times .45 = .4525, \end{aligned}$$

$$\begin{aligned} P(Y = 1 \mid do(X = 0)) &= .95(.10 \times .50 + .90 \times .50) \\ &\quad + .05(.05 \times .50 + .85 \times .50) \\ &= .95 \times .50 + .05 \times .45 = .4975. \end{aligned}$$

Remember Structural Causal Models?

- For simplicity, suppose we have X and Y:
  - SEM: $X = E_X;\ Y = f(X, E_Y)$
  - A particular experimental unit (e.g., a patient) u has its values for exogenous variables E_X and E_Y, say, e_x and e_y
  - Do intervention on X: $X = x;\ Y = f(x, E_Y)$
  - Potential outcome $Y(x,u)$ or $Y_x(u)$
  - $Y(x)$: counterfactual variable



$$PA_i \longrightarrow X_i$$

$$\begin{aligned}
X_1 &= E_1,\\
X_2 &= f_2(X_1, E_2),\\
X_3 &= f_3(X_1, E_3),\\
X_4 &= f_2(X_3, X_2, E_4),\\
X_5 &= f_5(X_4, E_5)
\end{aligned}$$

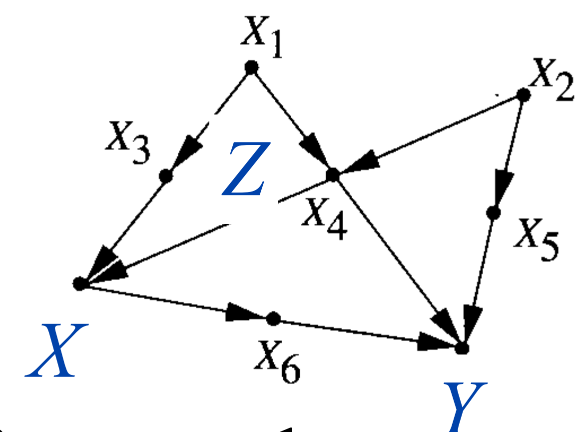
* Relation to Ignorability (Potential Outcome Framework)

Definition 3.3.1 (Back-Door)

A set of variables Z satisfies the back-door *criterion relative to an ordered pair of variables* (X_i, X_j) *in a DAG G if:*

(i) *no node in Z is a descendant of* X_i; *and*

(ii) *Z blocks every path between* X_i *and* X_j *that contains an arrow into* X_i.



- (Conditional) ignorability assumption in the potential outcome framework:

$$Y(x) \perp\!\!\!\perp X \mid Z.$$

$Y(x,u)$: the value attained by Y in unit u under intervention do(x);
$Y(x)$: counterfactual variable (u is treated as a variable)

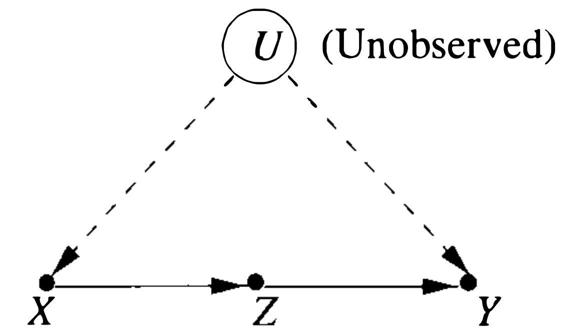
Definition 3.3.3 (Front-Door)

A set of variables Z is said to satisfy the front-door *crite[rion relative to an ordered pair] of variables* (X, Y) *if:*

(i) *Z intercepts all directed paths from X to Y*;

(ii) *there is no back-door path from X to Z*; *and*

(iii) *all back-door paths from Z to Y are blocked by X.*



- $Y(z,x) = Y(z)$; $\{Y(z), X\} \perp\!\!\!\perp Z(x)$.

* A Unification of the Graphical Criteria

- (Pear & Tian, 2002) A sufficient condition for identifying the causal effect $P(y \mid do(x))$ is that there exists no bi-directed path (i.e., a path composed entirely of bi-directed arcs) between X and any of its children.

- Necessary & sufficient conditions also exist (e.g., Shpitser and Pearl, 2008)…

- Examples:

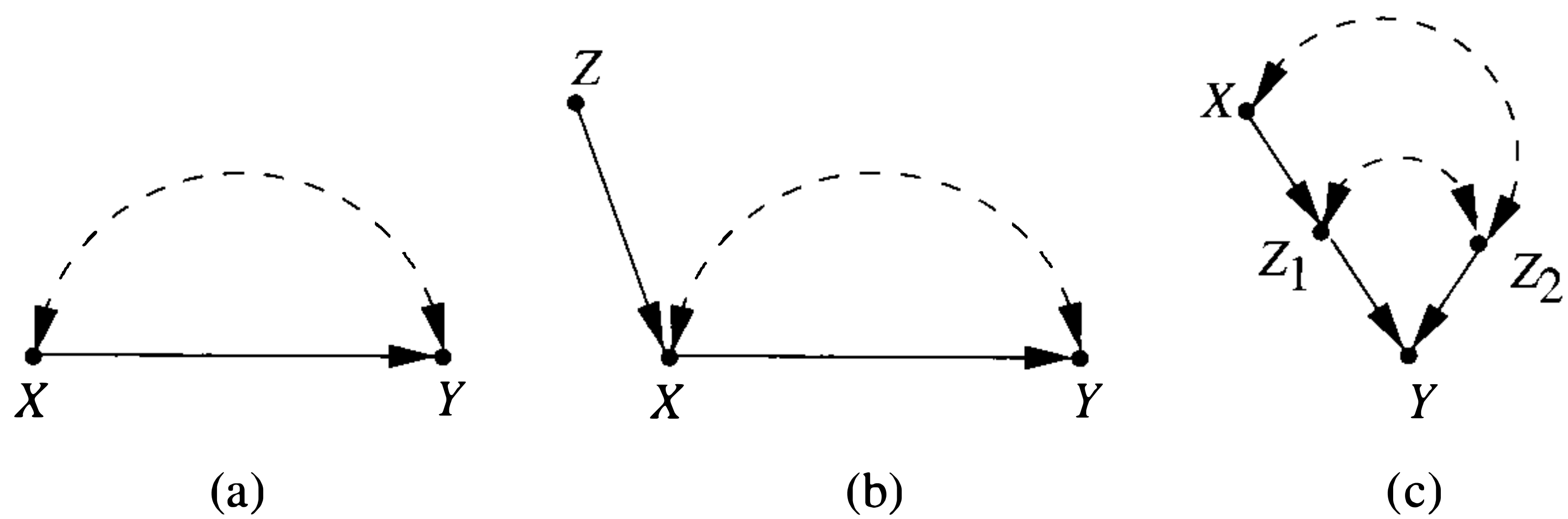


Figure 3.7 (a) A bow pattern: a confounding arc embracing a causal link $X \rightarrow Y$, thus preventing the identification of $P(y \mid \hat{x})$ even in the presence of an instrumental variable Z, as in (b). (c) A bowless graph that still prohibits the identification of $P(y \mid \hat{x})$.

*

A Unification: Examples

- Examples:

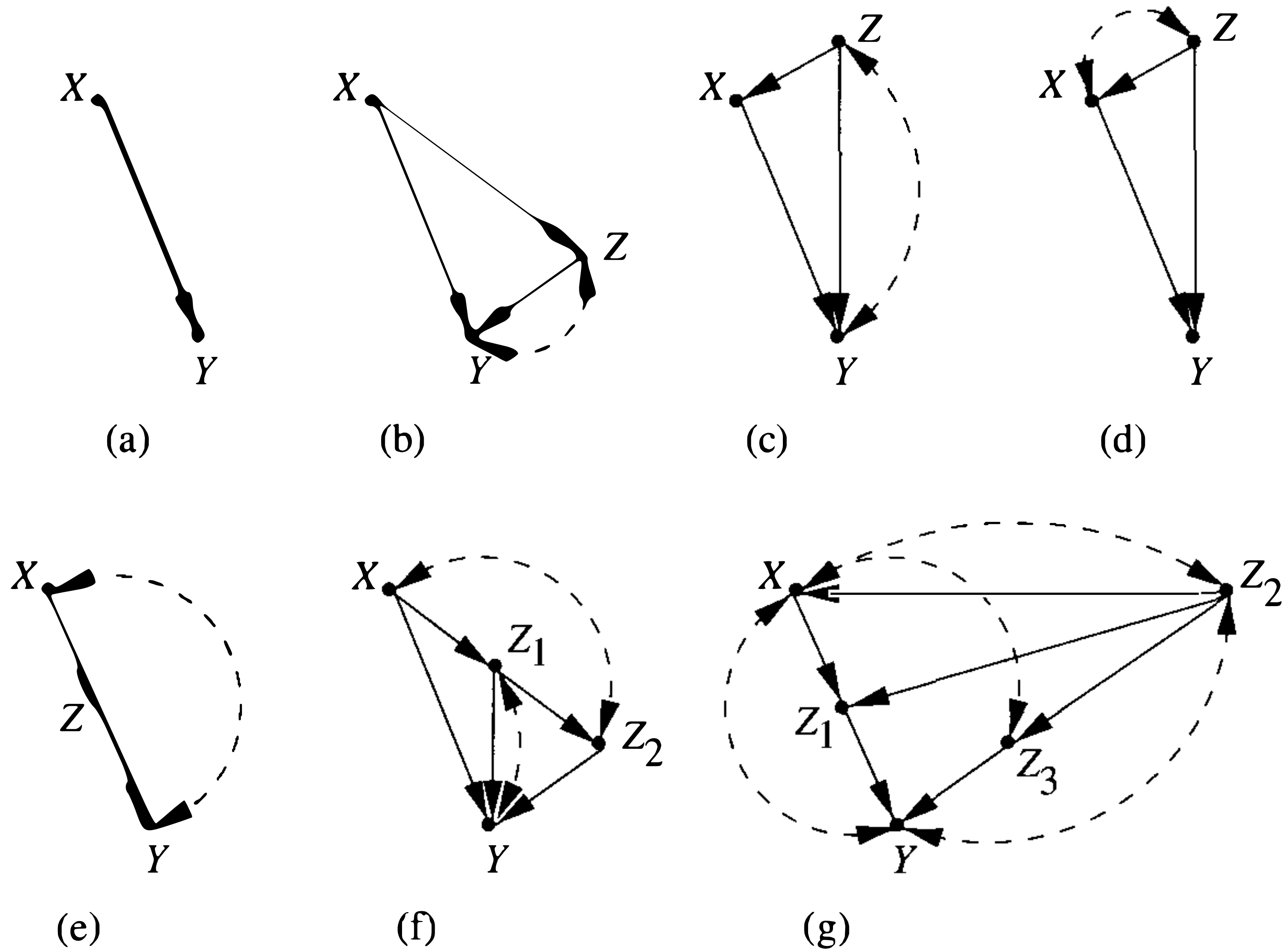


Figure 3.8 Typical models in which the effect of X on Y is identifiable. Dashed arcs represent confounding paths, and Z represents observed covariates.

*

A Unification: Examples

- Examples:

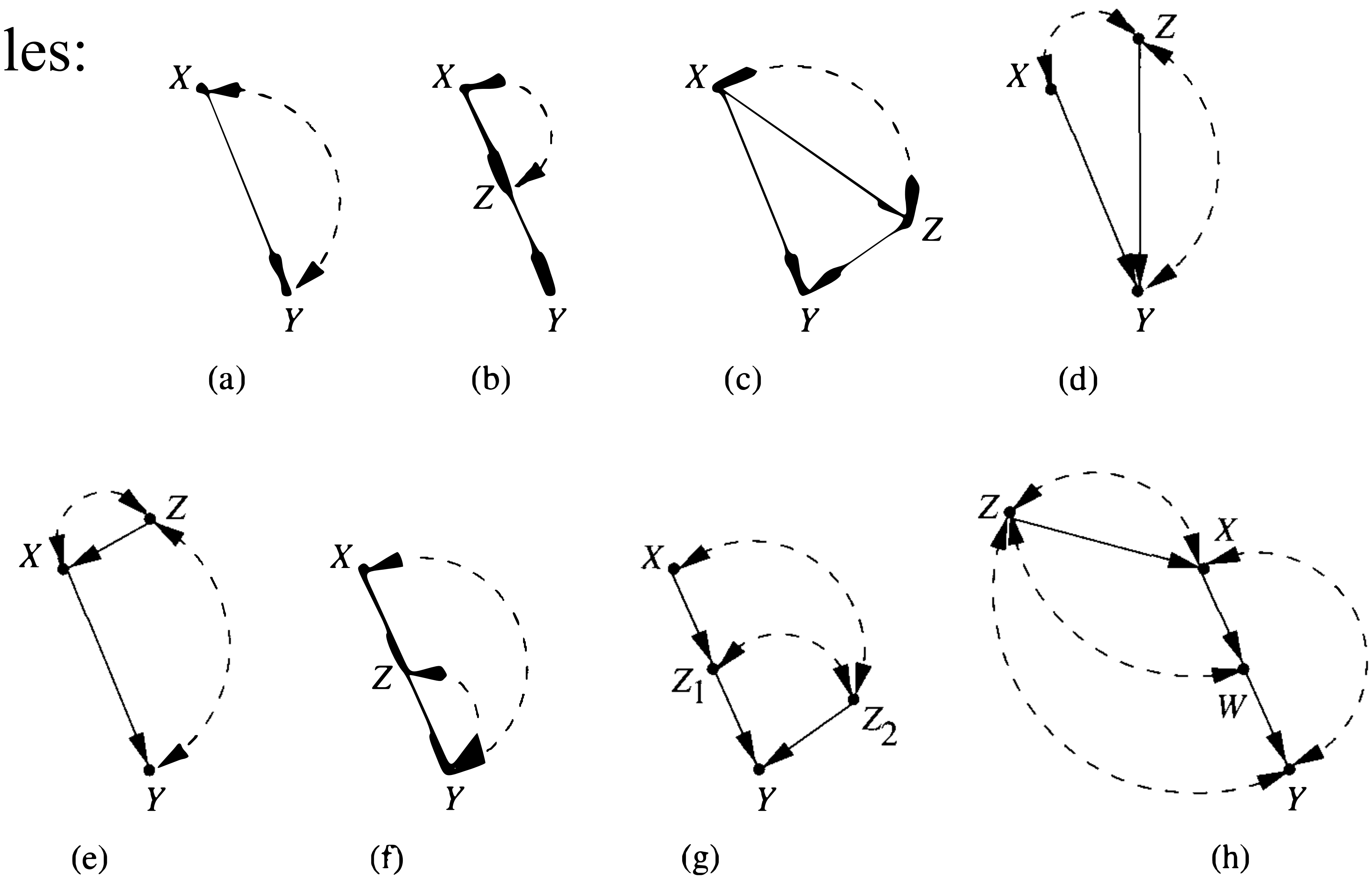
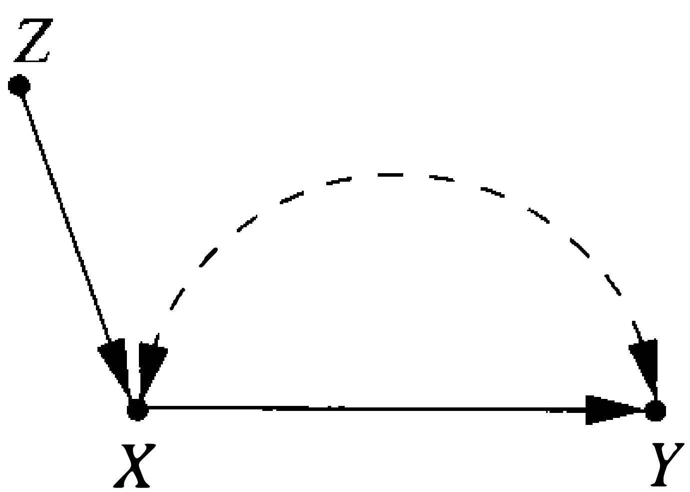
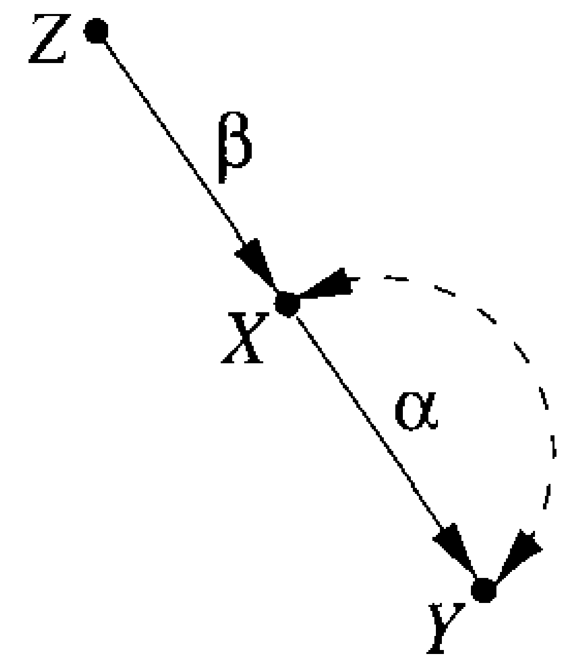


Figure 3.9 Typical models in which $P(y \mid \hat{x})$ is not identifiable.

*

Nonparametric vs. Parametric





- What if the causal relations are linear?

$\beta = r_{XZ}$ (regression coefficient of regressing X on Z)

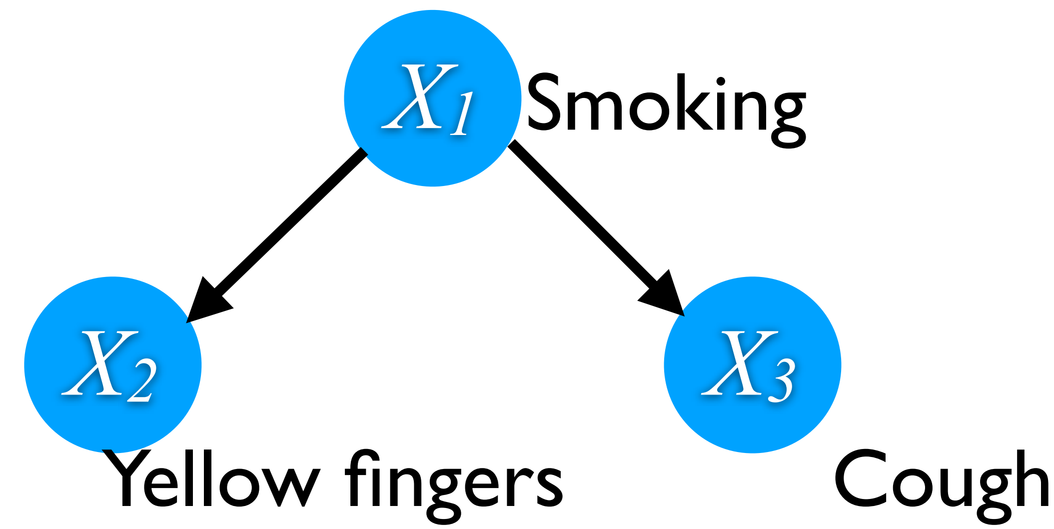
$\alpha\beta = r_{YZ}$

so $\alpha = r_{YZ}/r_{XZ}$.

Identification of Causal Effects & Counterfactual Inference: Outline

- Problem definition
- Potential outcome framework
    - Propensity score
- Backdoor criterion and front door criterion
- Counterfactual inference

Three Types of Problems in Current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
…	…	…

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

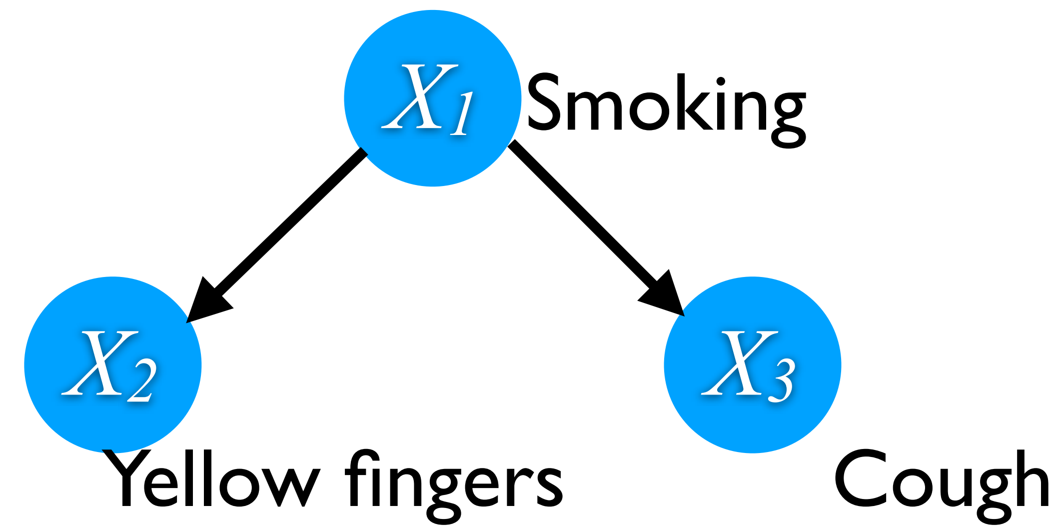
- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

$$P(X3 \mid do\,(X2=1))$$

- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X3_{X2=1} \mid X2 = 0, X3 = 1)$$

Three Types of Problems in Current AI



- Three questions:

X_1	X_2	X_3
1	0	0
0	0	1
0	1	1
1	1	1
0	0	0
0	1	0
1	1	1
1	1	1
0	0	0
1	0	0
…	…	…

- **Prediction**: Would the person cough if we *find* he/she has yellow fingers?

$$P(X3 \mid X2=1)$$

- **Intervention**: Would the person cough if we *make sure* that he/she has yellow fingers?

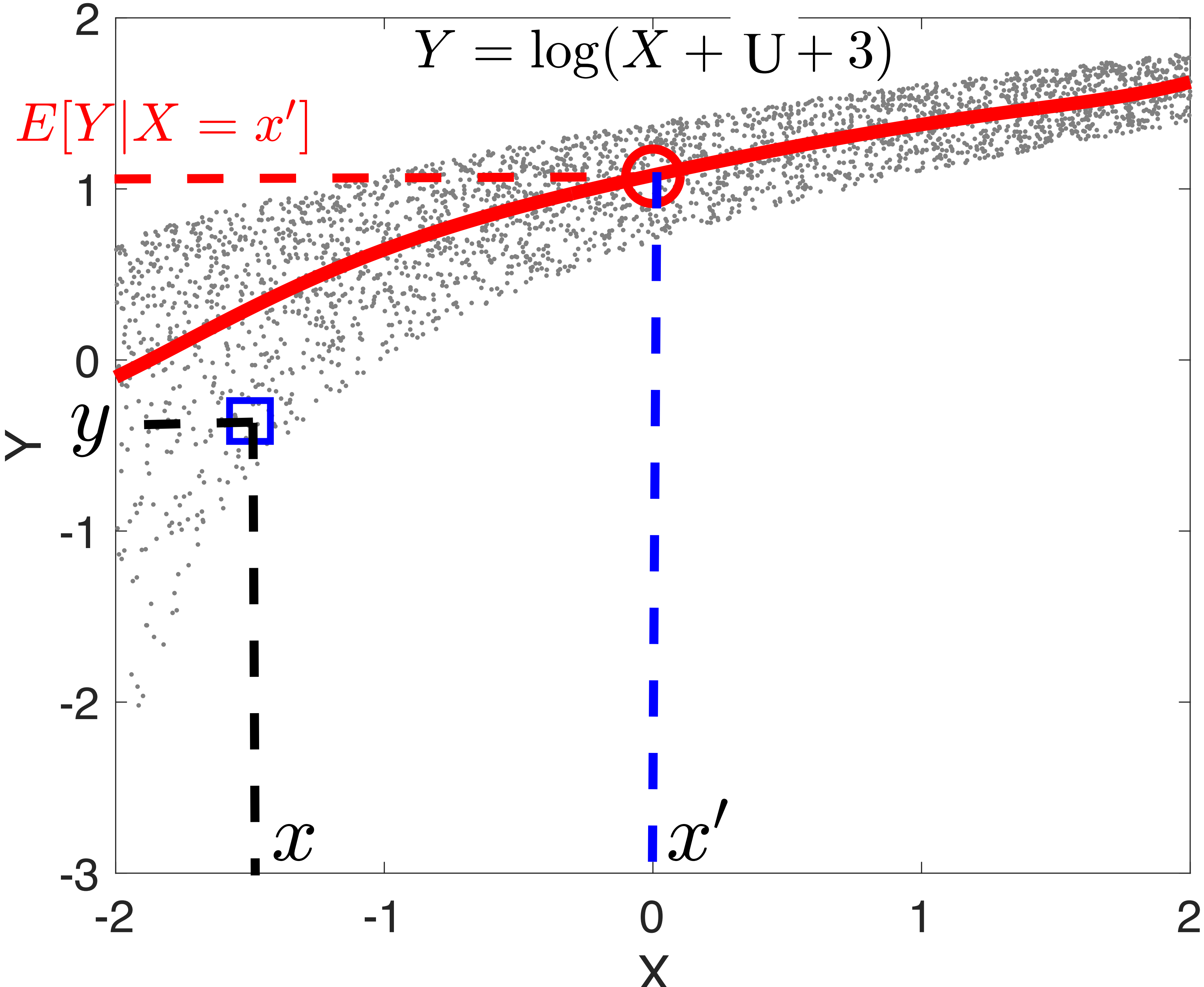
$$P(X3 \mid do\ (X2=1))$$

- **Counterfactual**: Would George cough *had* he had yellow fingers, *given that he does not have yellow fingers and coughs*?

$$P(X3_{X2=1} \mid X2 = 0, X3 = 1)$$

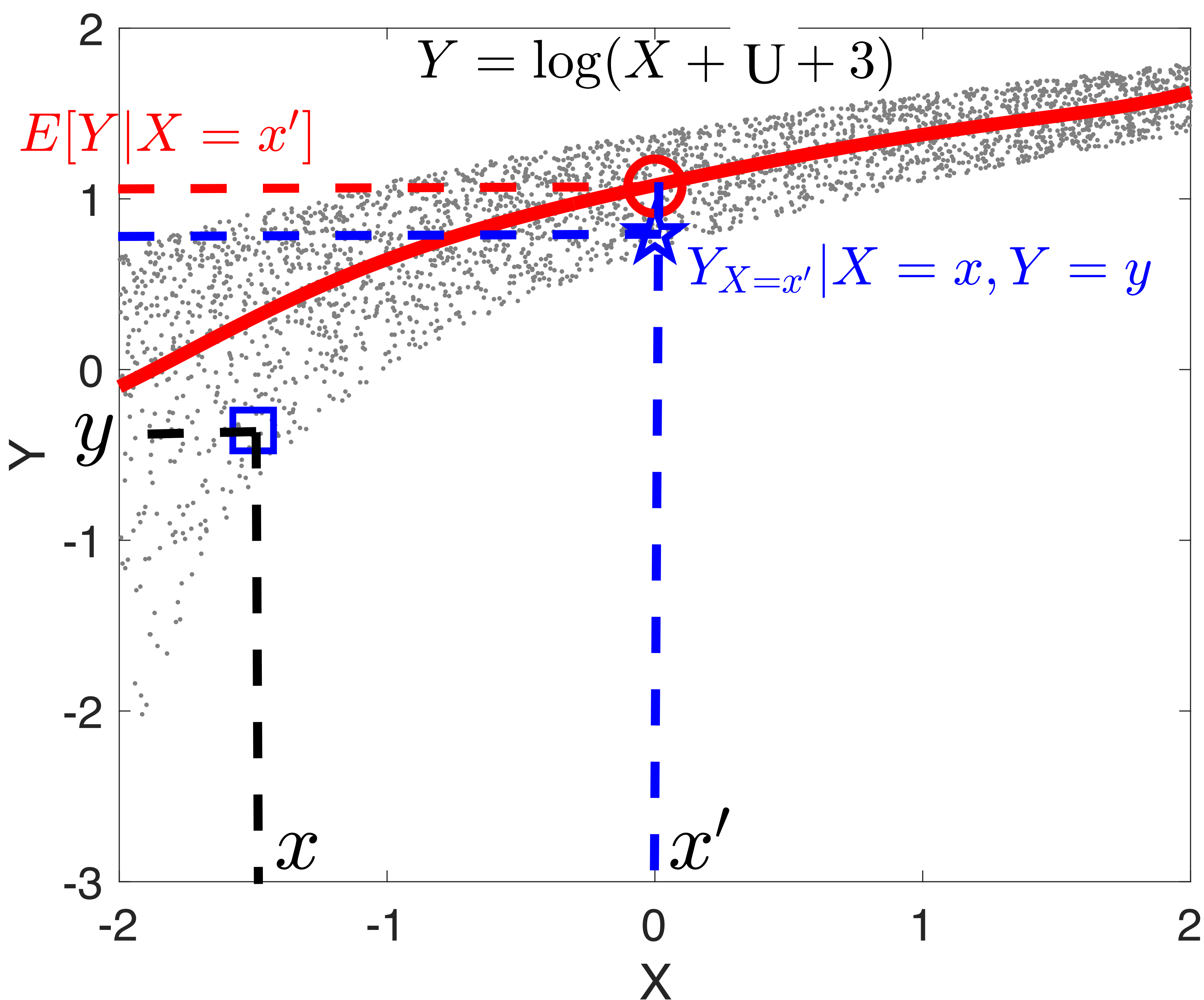
Counterfactual Inference vs. Prediction

- Suppose X→Y (*attendance* → *grade*) with Y = log(X + U + 3). For an individual with (x,y), what would Y be if X had been x'?



Counterfactual Inference vs. Prediction

- Suppose X→Y with Y = log(X + U + 3). For an individual with (*x*,*y*), what would Y be if X had been *x'* ?

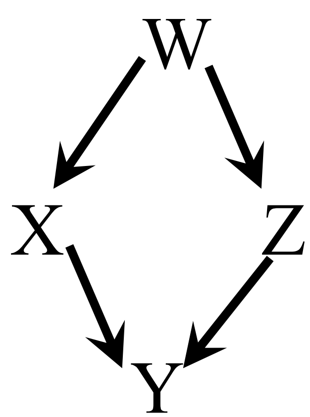


Standard Counterfactual Questions

- We talk about a particular situation (or unit) $U = u$, in which $X = x$ and $Y = y$
- What value would Y be had X been x' in situation u? I.e., we want to know $Y_{X=x'}(u)$, the value of Y in situation u if we do($X=x'$)
- u is not directly observable, so $P(Y_{X=x'} \mid X = x, Y = y)$ instead

For identification of causal effects, U is randomized. It is fixed for counterfactual inference.

Counterfactual Inference



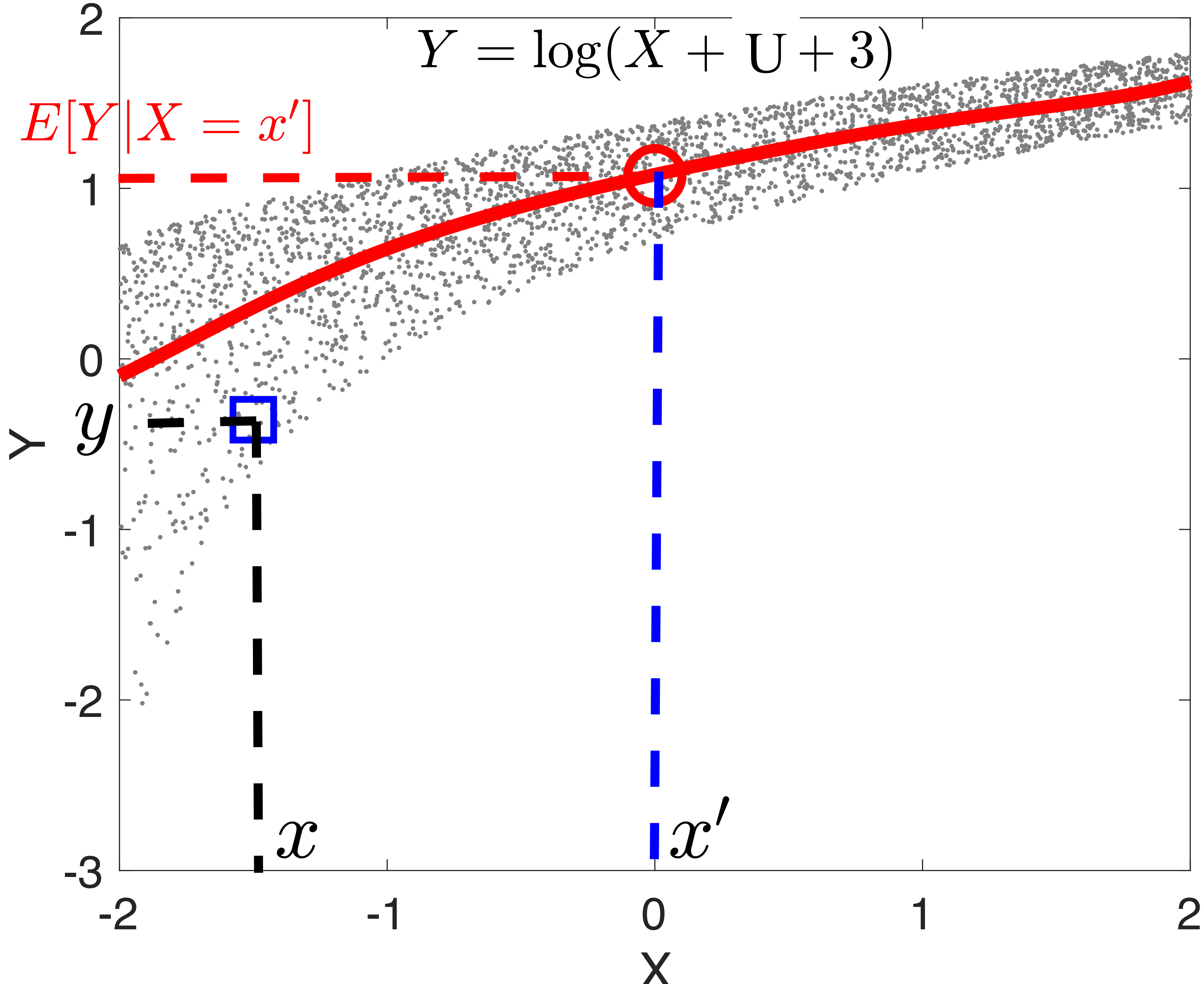
$W = U_W$
$X = f_X(W, U_X)$
$Z = f_Z(W, U_Z)$
$Y = f_Y(X, Z, U_Z)$

$$P(Y_{X=x'} \mid \underbrace{X = x,\ Y = y,\ W = w}_{\textit{evidence}})$$

- Three steps
  - Abduction: find P(U | *evidence*)
  - Action: Replace the equation for X by $X = x'$
  - Prediction: Use the modified model to predict Y

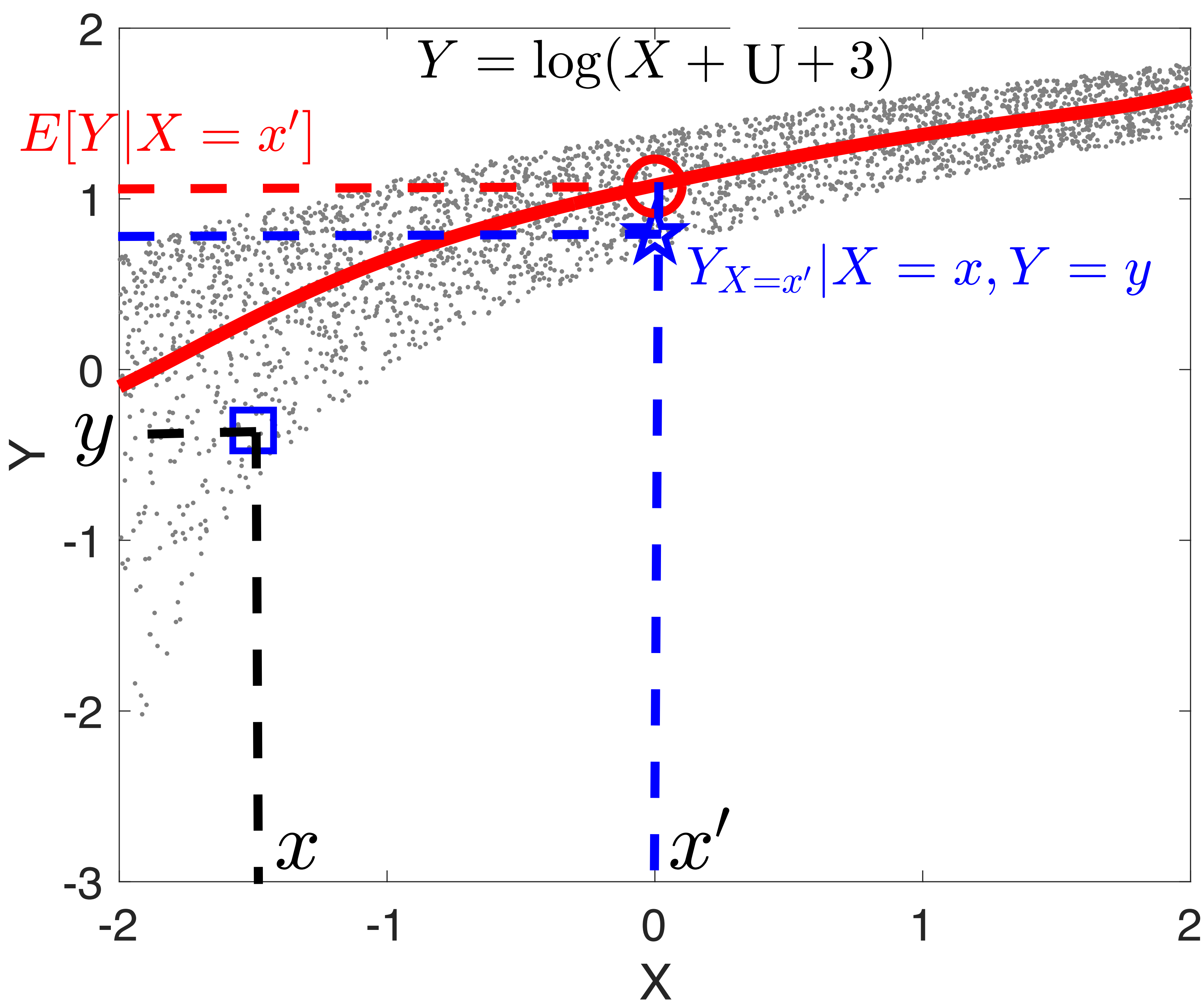
Counterfactual Inference vs. Prediction

- Suppose X→Y (*attendance* → *grade*) with Y = log(X + U + 3). For an individual with (x,y), what would Y be if X had been x' ?



Counterfactual Inference vs. Prediction

- Suppose X→Y with Y = log(X + U + 3). For an individual with (*x*,*y*), what would Y be if X had been *x'* ?



Counterfactual Inference: Discussions

- Discover the hidden features and use them, since you focus on a specific subject
- Do we really need an SCM for counterfactual reasoning?
- Other potential issues?

- Chaochao Lu*, Biwei Huang*, Ke Wang, José Miguel Hernández-Lobato, Kun Zhang, Bernhard Schölkopf, Sample-Efficient Reinforcement Learning via Counterfactual-Based Data Augmentation, NeurIPS Workshop on Offline Reinforcement Learning, 2020

Summary: Causal Effect Identification & Counterfactual Reasoning

- Classical problem
- What is taken as input?
- What does identifiability mean?
- Potential outcomes framework
- Backdoor criterion and unification
- Propensity score
- Difference from counterfactual inference