

Regression #1

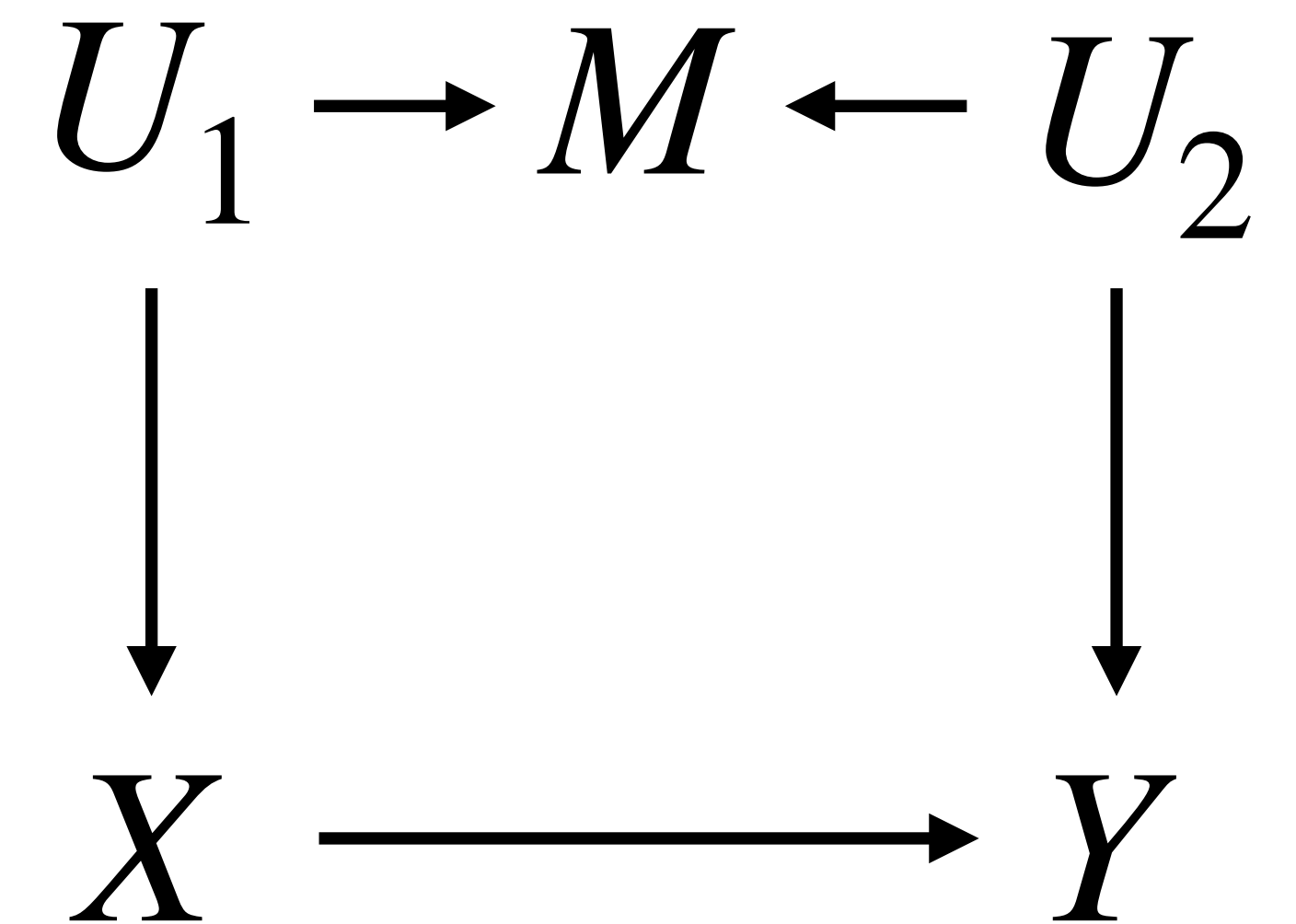
Manoel Horta Ribeiro
manoel@cs.princeton.edu

Recap: Conditional Exchangeability

- Let Y and T be binary random variables specifying treatments and outcomes, and X be a confounder.
 - E.g., T is the prompt style, Y is whether the question is solved, and X is the question difficulty, again, “in the wild.”
- **Conditional Exchangeability** entails that:
$$Y^a \perp T \text{ for all } | X \quad a \in \{0,1\}$$
 - I.e., within difficulty levels, treatments are if as random!

Recap: DAGs

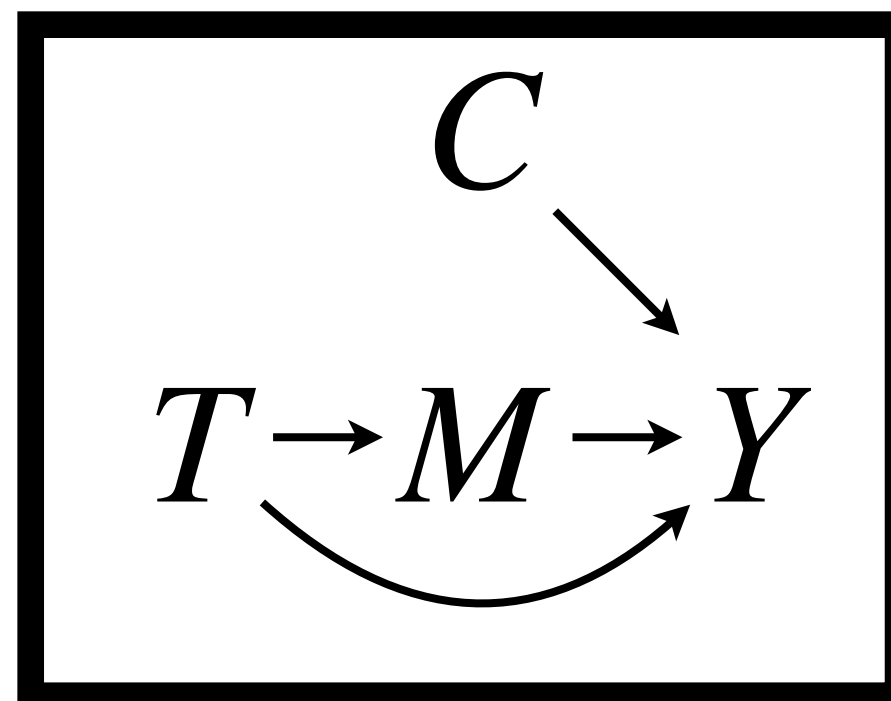
- A graphical way to represent random variables
- $X \rightarrow Y$: X *may cause* Y
- Why is it useful? Let's you know whether causal effects can be identified!



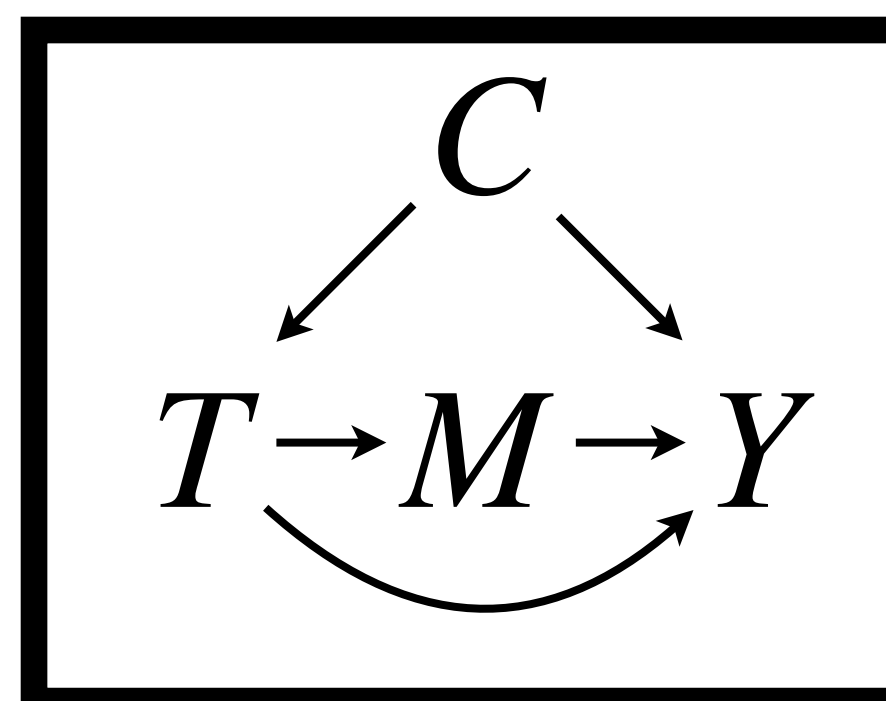
Recap: DAGs

- Y : whether the model solves the problem
- C : the difficulty of the question.
- T : indicate the prompt (CoT/standard)
- M : whether the model used intermediate tokens

Experiment



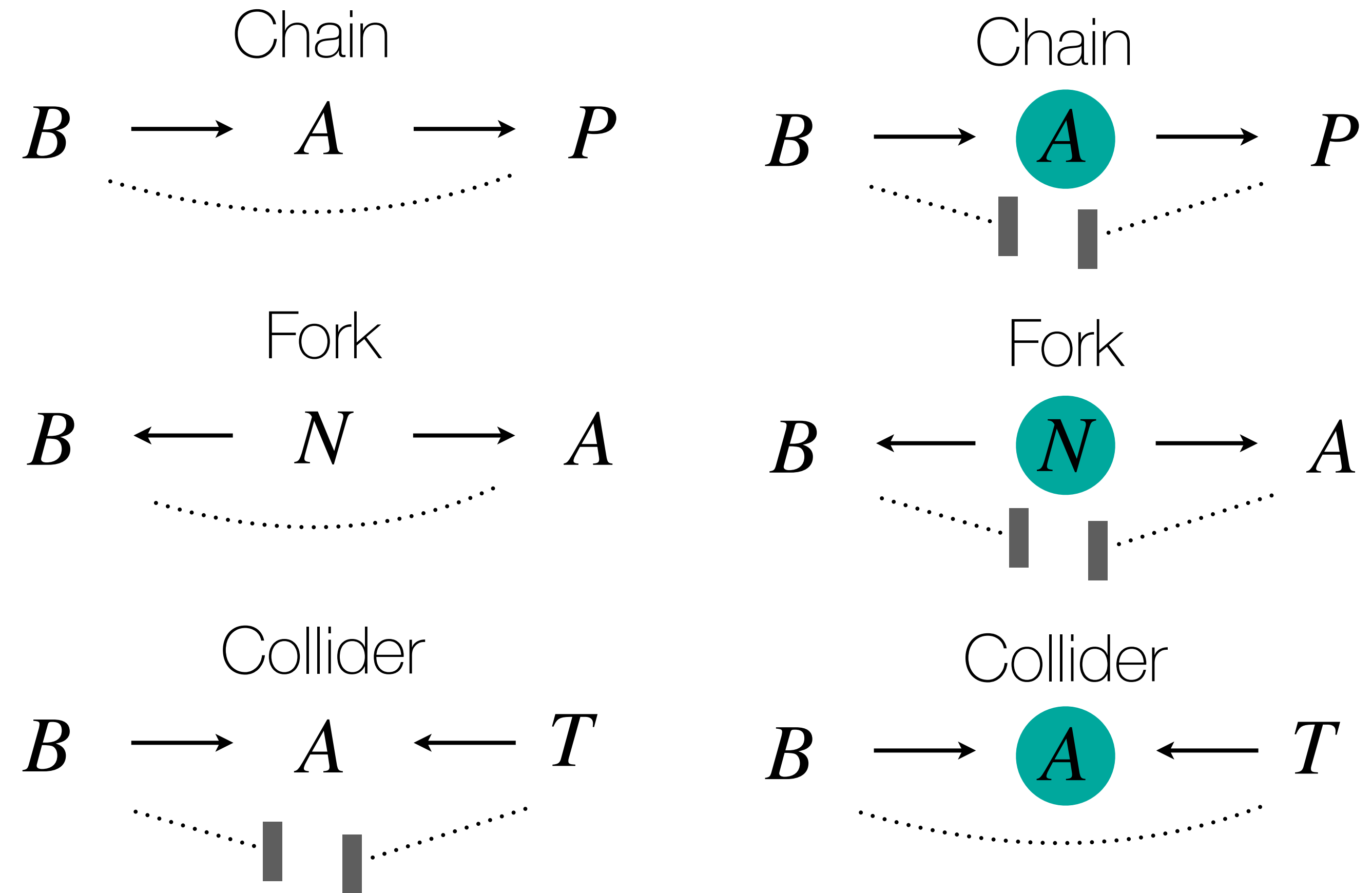
Observational



$T \rightarrow Y$ direct causal effect

$T \rightarrow M \rightarrow Y$ indirect causal effect

Recap: DAG building blocks



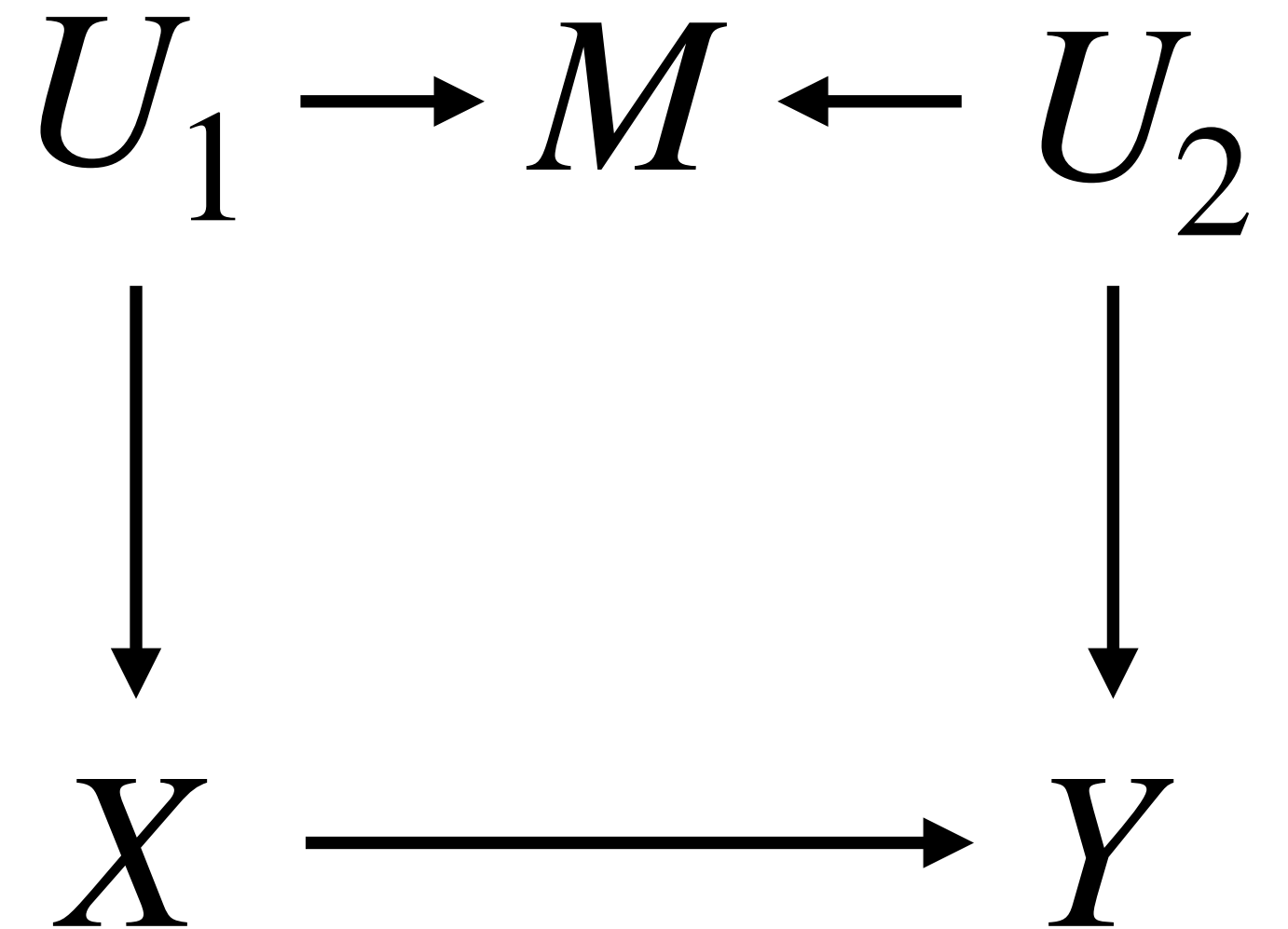
Recap: Backdoor criteria

- A graph-based criterion that allows researchers to identify a set of variables $\mathbf{Z}$ that, when controlled for, allow them to estimate the causal effect of a variable X on another variable Y

1. No node in $\mathbf{Z}$ must be a descendant of X .
2. $\mathbf{Z}$ blocks every path from X to Y that starts with an arrow into X .

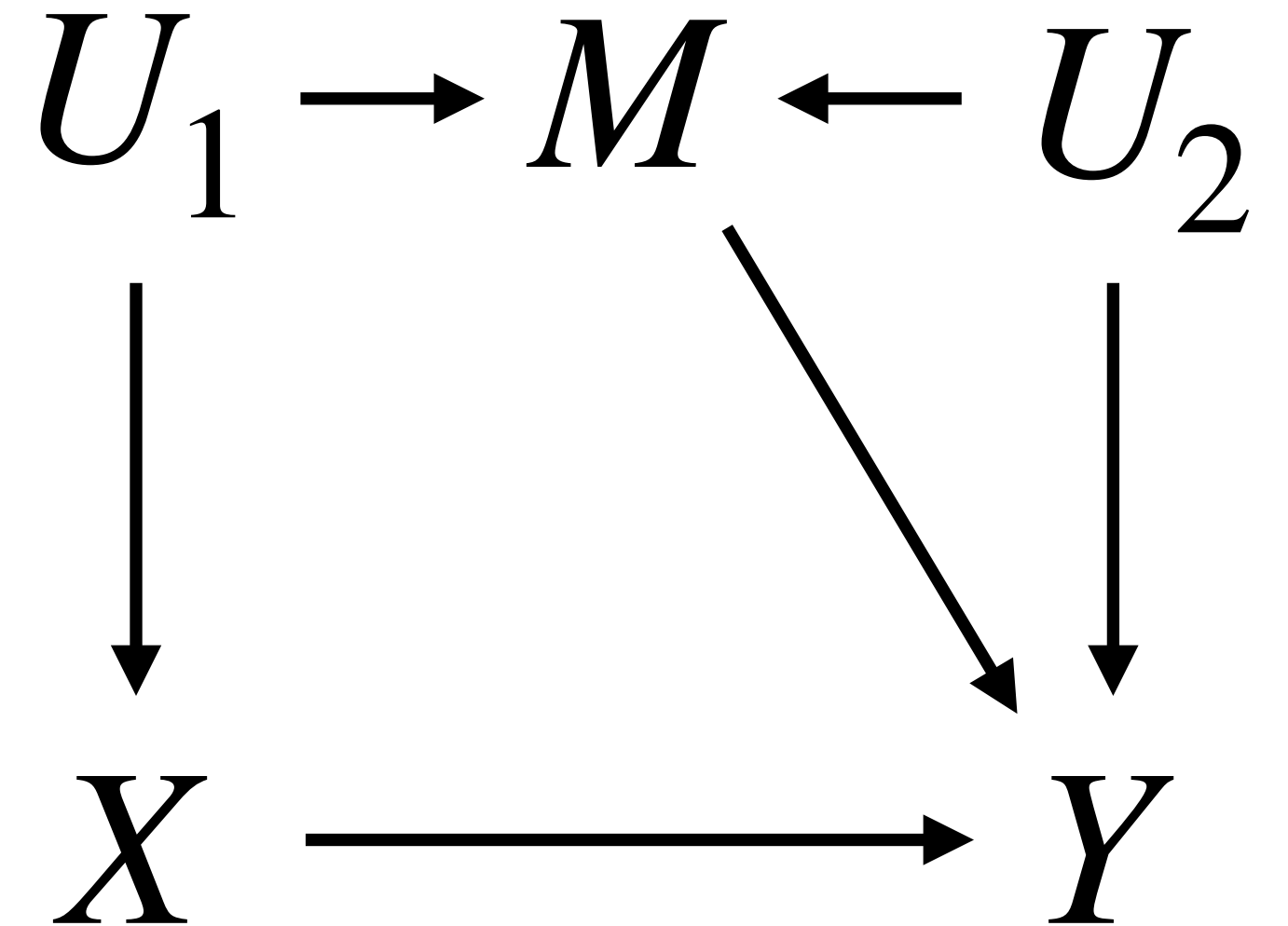
M -bias

- Suppose we care about the effect of education X on diabetes Y .
- We also observe the mother's history of diabetes M .
- Assume two non-observed variables:
 - U_1 : Income during childhood
 - U_2 : Genetic risk of diabetes
- Controlling for M opens a backdoor path!



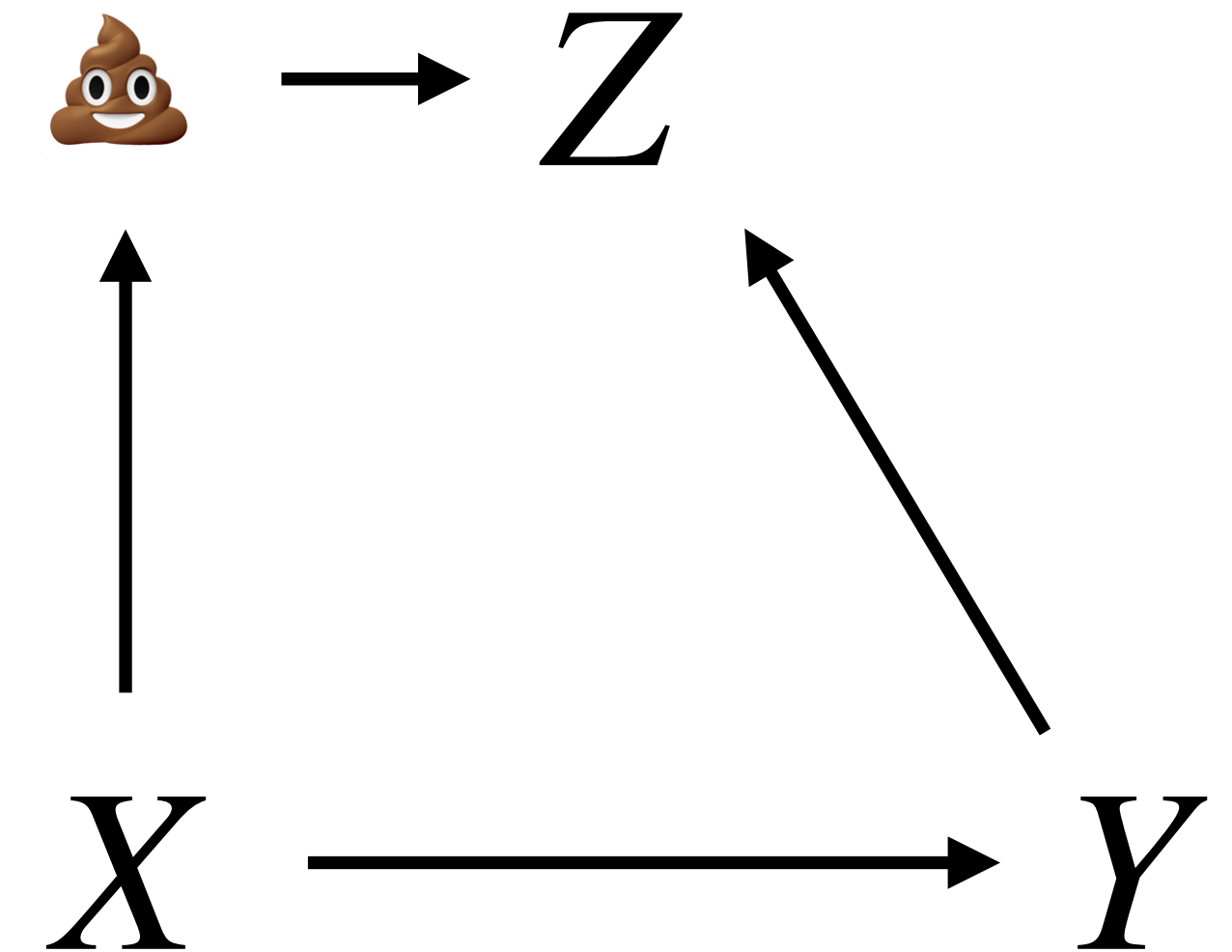
“Damned if you do, damned if you don’t.”

- Suppose that mothers’ diabetes influences their children through mechanisms other than genes (e.g., epigenetics).
- Controlling for M opens the path
 $X \rightarrow U_1 \rightarrow M \leftarrow U_2 \leftarrow Y$
- Not doing so leaves open the path:
 $X \rightarrow U_1 \rightarrow M \rightarrow Y$
- We are toast!



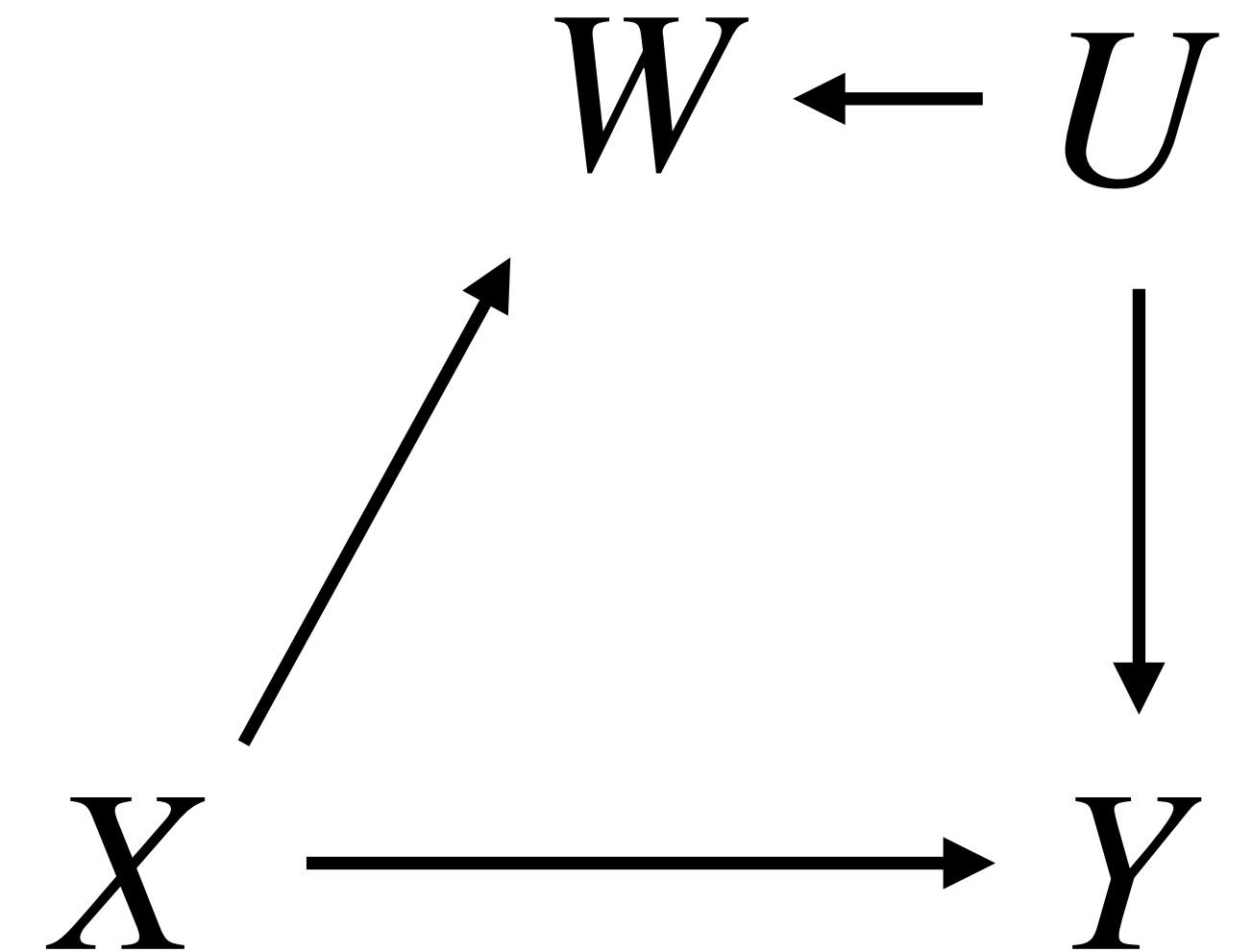
Case-control Bias

- Suppose we care about the effect of coffee drinking X on heart conditions Y .
- But we only consider in our sample people who have been admitted to a hospital Z .
- Imagine coffee does not cause heart problems.
- But it does cause intestinal problems (💩) that lead to hospitalization!
- Controlling for Z opens the path
 $X \rightarrow \text{💩} \rightarrow Z \leftarrow Y$



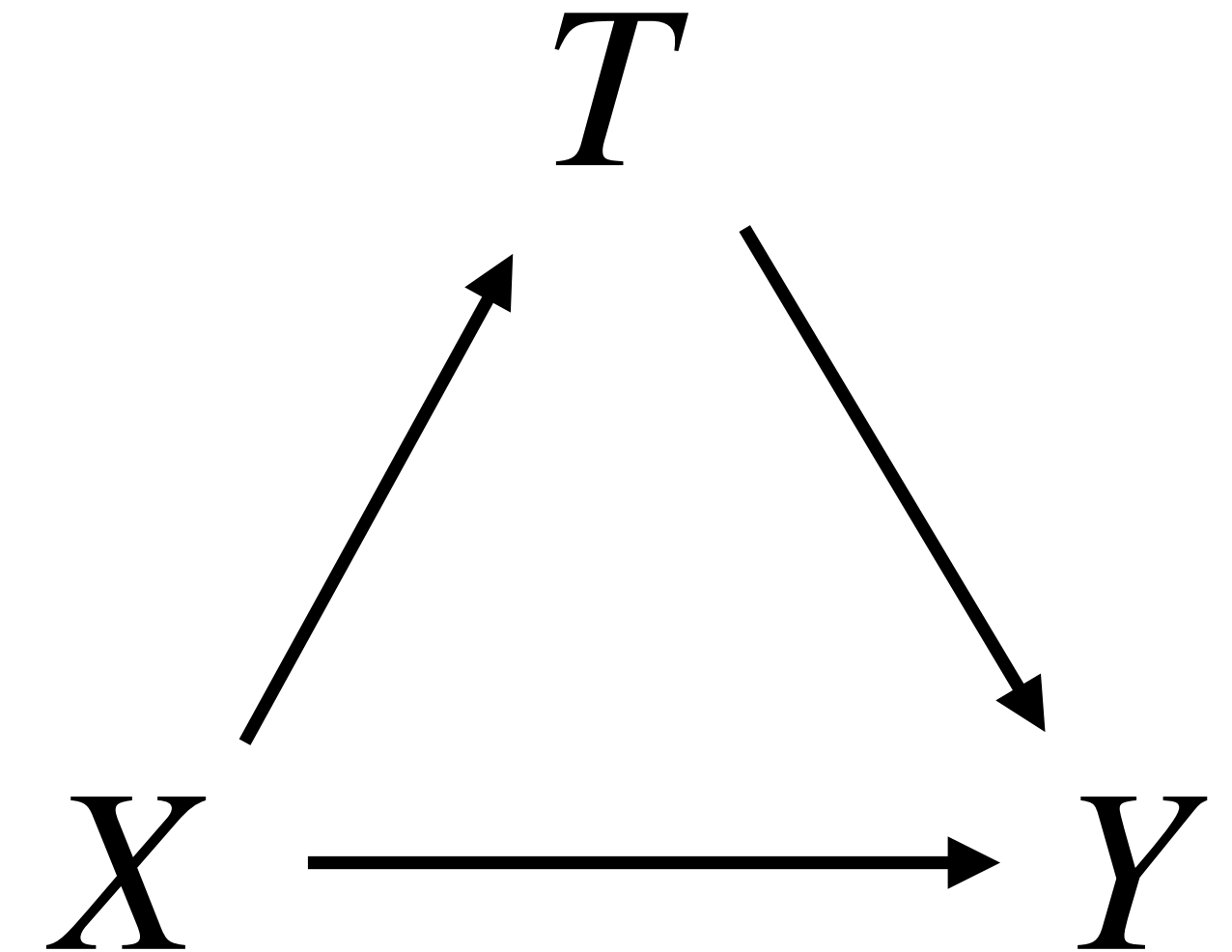
Selection Bias

- Suppose we care about the effect of smoking X on infant mortality Y before 4 weeks.
- We decided to control for weight at birth W
- But there are serious genetic conditions that cause both low weight at birth and infant mortality!



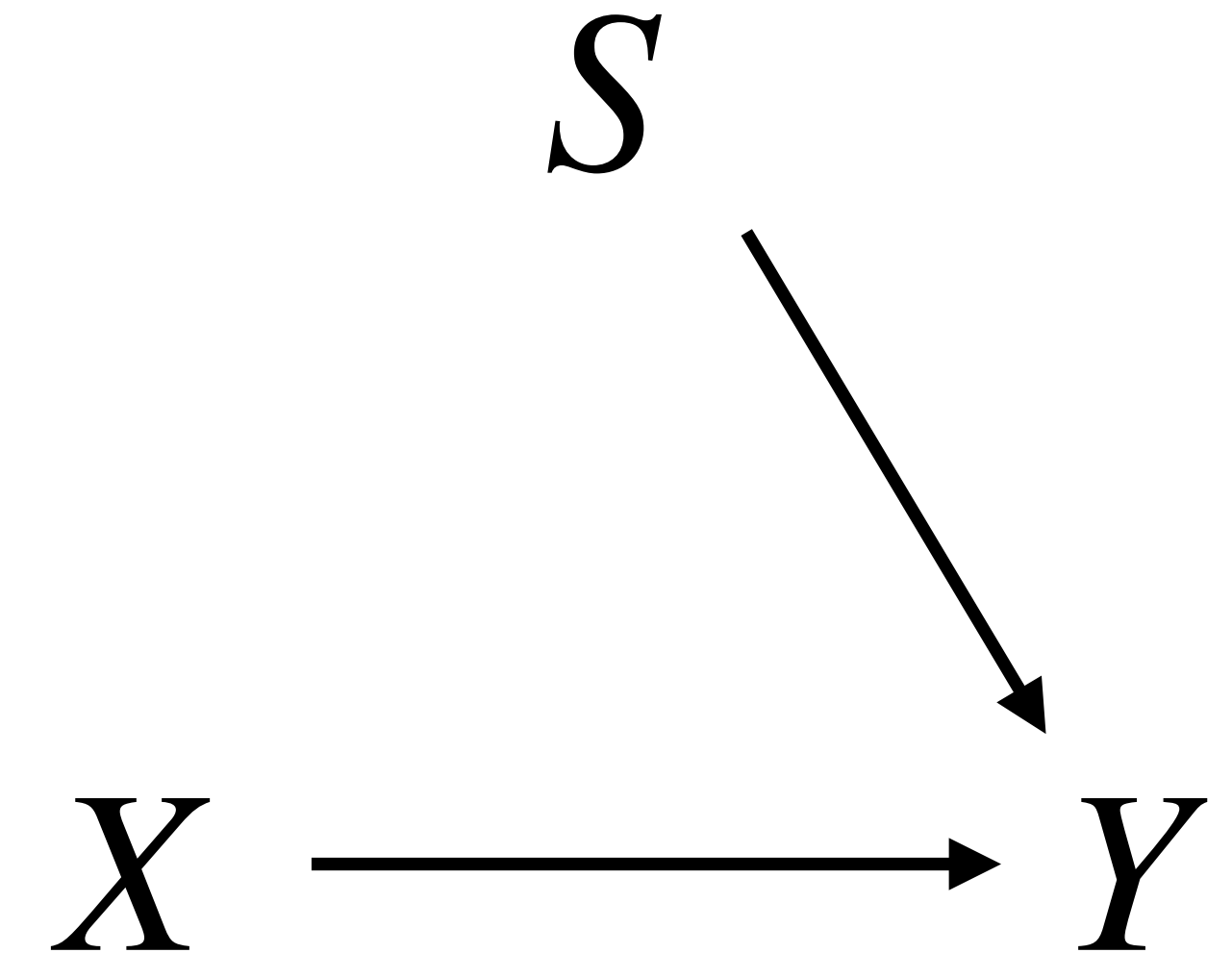
Overcontrolling

- Suppose we care about the effect of smoking X on mortality Y .
- We control for tar in the lungs T
- Smoking can kill you:
 - Through tar $X \rightarrow T \rightarrow Y$
 - In other ways $X \rightarrow Y$
- We block a legitimate path of the causal effect!



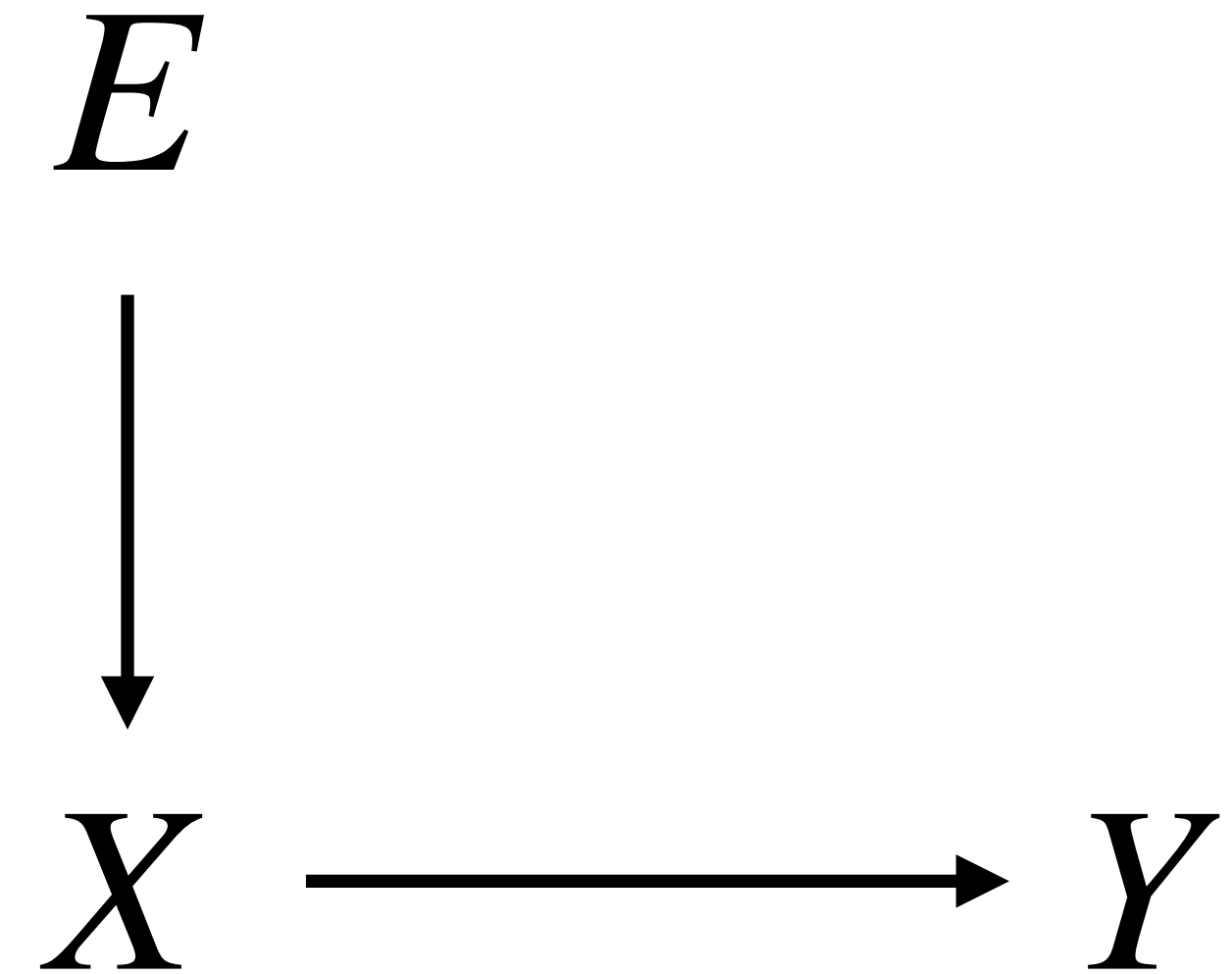
Reducing Variance

- Suppose we care about the effect of a drug X on mortality Y from a disease.
- Mortality from that disease varies widely by biological sex S .
- When we control for biological sex (S), this prevents spurious variation in the treatment and control group



Increasing Variance

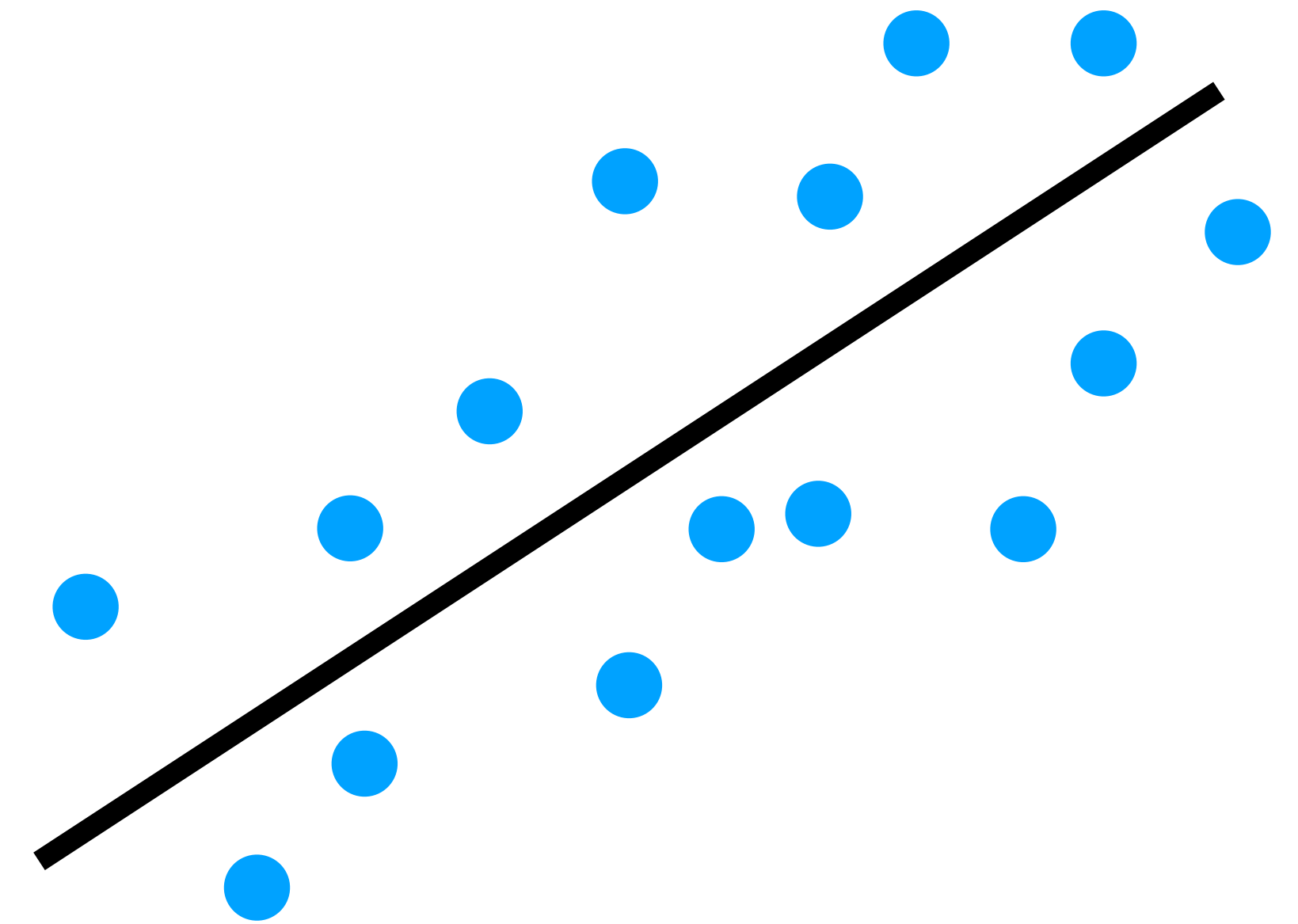
- We care about the effect of attending office hours X on the final exam Y .
- Students are randomly enrolled to receive reminders E , which strongly predict attendance!
- Controlling for E makes you compare attendees vs. non-attendees within reminder groups!
 - $E = 1$: almost everyone attended;
 - $E = 0$: almost no one attended;
- Bad for precision: treatment varies little!



$$Z = \frac{\hat{\tau} - 0}{SE(\hat{\tau})} = \frac{\bar{Y}_1 - \bar{Y}_0}{SE(\bar{Y}_1 - \bar{Y}_0)}$$

Linear Regression

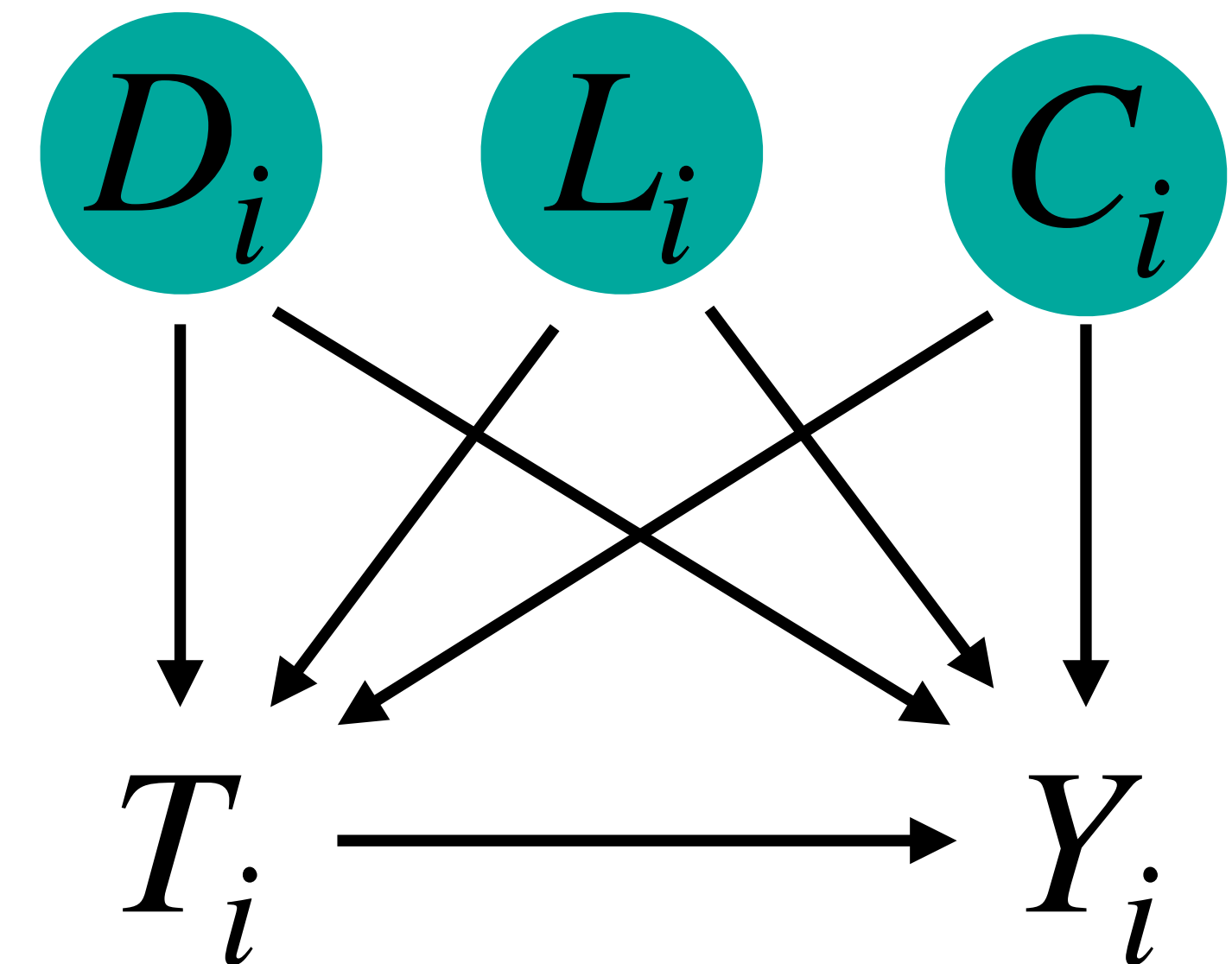
- Linear regression is typically described in CS as a *prediction* algorithm.
- But in other disciplines, it often used a powerful analytical tool to:
 - Describe data;
 - Estimate causal effects;



Problem Setup

Does CoT prompting helps any question?

- $T_i \in \{0,1\}$ — prompt-type
- $Y_i \in [0,10]$ — Correctness score
- $C_i \in [0,10]$ — Question difficulty
- $L_i \in [0,10]$ — Question “clarity”
- $D_i \in \{\text{math, biology, ...}\}$ — Domain



What do we need to control for?

Calculating the Treatment Effect

- Conditional exchangeability saves us:

$$\tau = E_{D,L,C} \left[E[Y \mid D, L, C, T = 1] \right] - E_{D,L,C} \left[E[Y \mid D, L, C, T = 0] \right]$$

- How do we even handle a “continuous” variable? Maybe we can bin it?
- What if there are 20 different domains?
 $20 \times 20 = 400$ “cells”

	math	biology	history	...
[0.0,0.5]				...
(0.5,1.0]				...
(1.5,2.0]				...
(2.0,2.5]				...
...	...	...	...	...

Calculating the Treatment Effect

- For each cell, we could calculate the average outcome on the treated vs. control units
- As the number of confounding variables increases, you may find many cells with zero treated or untreated units!

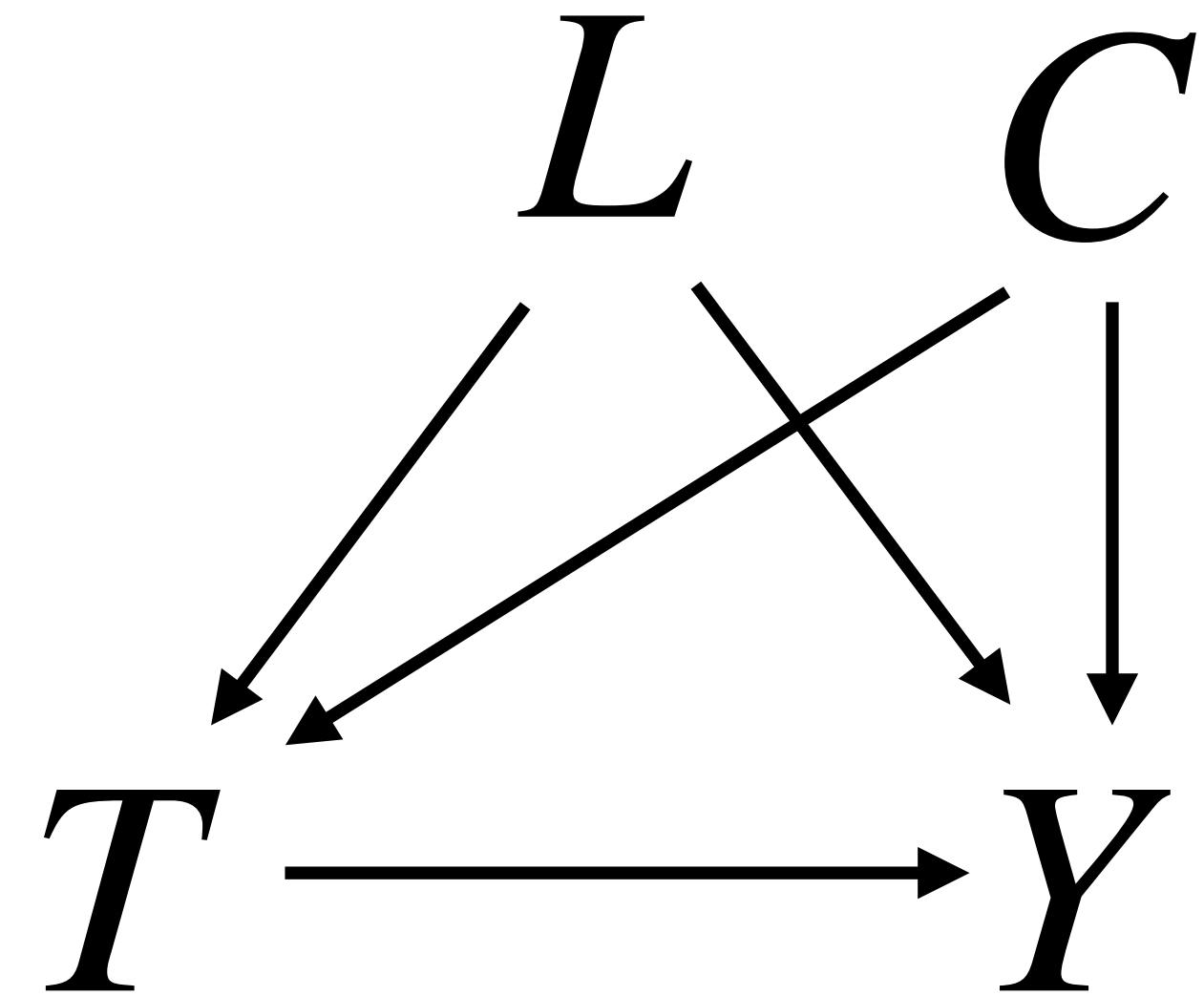
	math	biology	history	...
[0.0,0.5]			∅	...
(0.5,1.0]				...
(1.5,2.0]		∅		...
(2.0,2.5]				...
...	...	...	...	

Idea: Assume a Model

- Estimate $E[Y \mid C, L, T]$ parametrically.

$$Y = \beta_0 + \beta_1 T + \beta_2 C + \beta_3 L + \epsilon$$

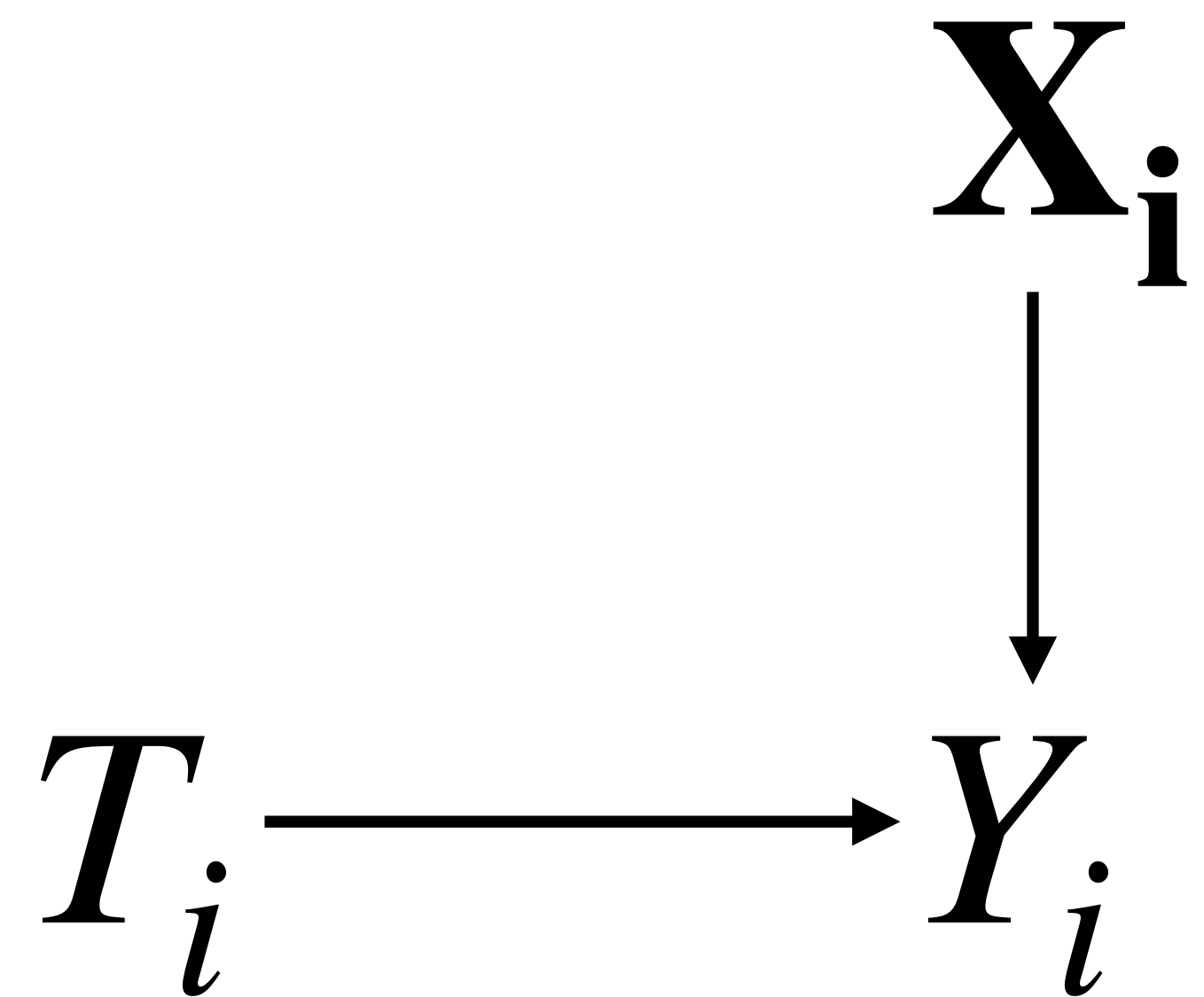
- Using a model fills the gaps by extrapolation!
- Correctness depends on the modeling assumptions!
 - E.g., linearity, additivity.



Backtracking: Experimental Case

- Assume it is an experiment, and we don't have access to confounders $\mathbf{X}$.
- We observe $\langle Y_i, T_i \rangle$ $i = 1, \dots, n$ where $T_i \in \{0, 1\}$ and $Y_i \in \{0, 1\}$
- We assume the model:

$$Y_i = \beta_0 + \beta_1 T_i + \epsilon_i$$



Ordinary Least Squares

- Choose $\hat{\beta}_0, \hat{\beta}_1$ to minimize the sum of squared residuals.

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 T_i)^2$$

- Given convexity, we find the global minima with first-order conditions:

$$\frac{\partial}{\partial \beta_0} \sum_i (Y_i - \beta_0 - \beta_1 T_i)^2 = -2 \sum_i (Y_i - \beta_0 - \beta_1 T_i) = 0 \Rightarrow \sum_i \hat{\epsilon}_i = 0$$

$$\frac{\partial}{\partial \beta_1} \sum_i (Y_i - \beta_0 - \beta_1 T_i)^2 = -2 \sum_i T_i (Y_i - \beta_0 - \beta_1 T_i) = 0 \Rightarrow \sum_i T_i \hat{\epsilon}_i = 0$$

Ordinary Least Squares

$$\sum_i T_i \hat{\epsilon}_i = 0$$

$$\sum_i \hat{\epsilon}_i = 0$$

$$\sum_{i:T_i=1} T_i (Y_i - \beta_0 - \beta_1 T_i) = 0$$

$$\sum_i (Y_i - \beta_0 - \beta_1 T_i) = 0$$

$$\frac{1}{n_1} \sum_{i:T_i=1} (Y_i - \beta_0 - \beta_1) = 0$$

$$\frac{1}{n_1} \sum_{i:T_i=1} (Y_i - \beta_0 - \beta_1) + \frac{1}{n_0} \sum_{i:T_i=0} (Y_i - \beta_0) = 0$$

$$\frac{1}{n_1} \sum_{i:T_i=1} Y_i = \beta_0 + \beta_1$$

$$\frac{1}{n_0} \sum_{i:T_i=0} (Y_i - \beta_0) = 0$$

$$\bar{Y}_1 = \beta_0 + \beta_1$$

$$\bar{Y}_0 = \beta_0$$

$$\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0 = \text{ATE}$$

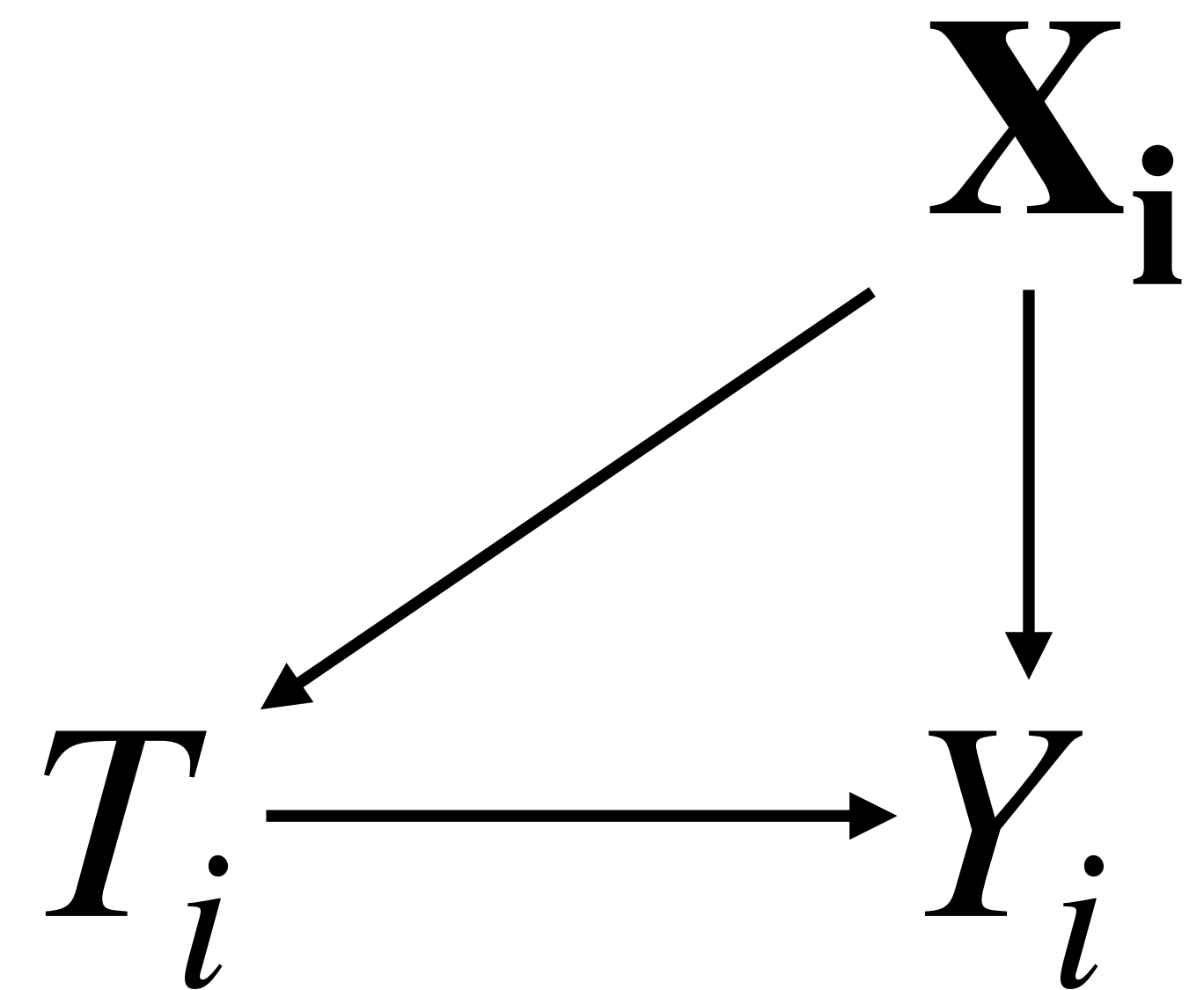
What if we have controls?

- Assume it is not an experiment; we have access to confounders $\mathbf{X}$.
- We observe $\langle Y_i, T_i, \mathbf{X}_i \rangle$ $i = 1, \dots, n$ where $T_i/Y_i \in \{0, 1\}$ and $\mathbf{X}_i \in \mathbb{R}^p$
- Let's design a new matrix

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1} & T & X_1 & \dots & X_p \end{bmatrix}$$

- We assume the model:

$$Y_i = \boldsymbol{\beta}^T \mathbf{Z}_i + \epsilon_i$$



Ordinary Least Squares (Multivariate)

- Choose $\hat{\beta}$ to minimize the sum of squared residuals.

$$\hat{\beta} = \arg \min \sum_{i=1}^n (Y_i - \beta^T \mathbf{Z}_i)$$

- Given convexity, we find the global minima with FOCs:

$$\mathbf{Z}^T(\mathbf{Y} - \mathbf{Z}\beta) = 0$$

- If $\mathbf{Z}$ has full column rank (so $\mathbf{Z}^T\mathbf{Z}$ is invertible), the solution is:

$$\hat{\beta} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}$$

But what does adding controls do?

- When we run: $Y_i = \beta_0 + \beta_1 T_i + \beta_2 C_i + \epsilon_i$

- This is equivalent to running:

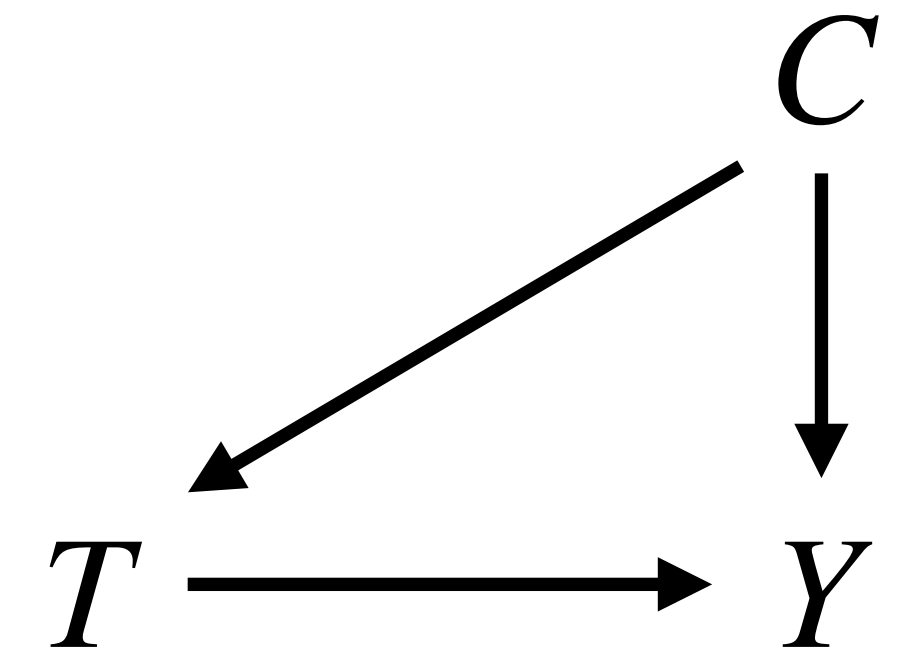
$$T_i = \alpha_0 + \alpha_1 C_i + u_i, \quad Y_i = \gamma_0 + \gamma_1 C_i + v_i$$

- Then, computing:

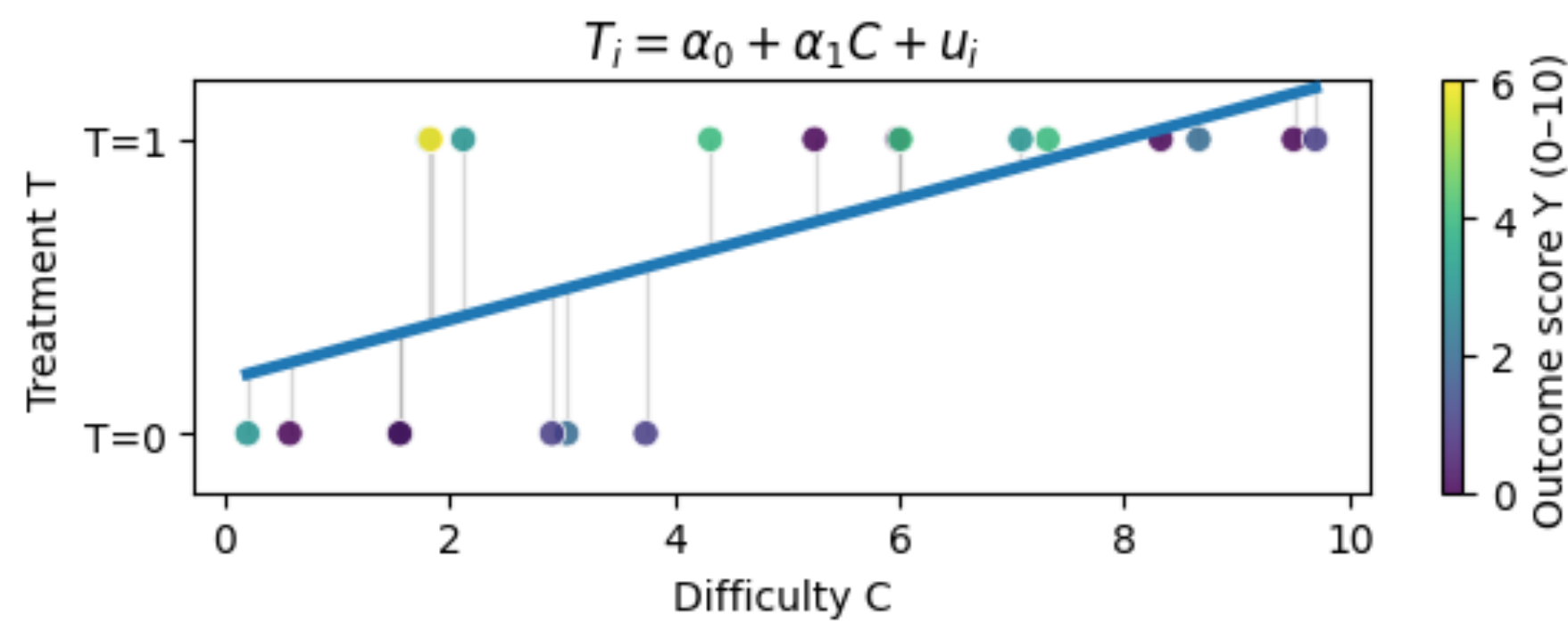
$$\tilde{T}_i = T_i - \alpha_0 - \alpha_1 C_i, \quad \tilde{Y}_i = Y_i - \gamma_0 - \gamma_1 C_i$$

- And then, regressing

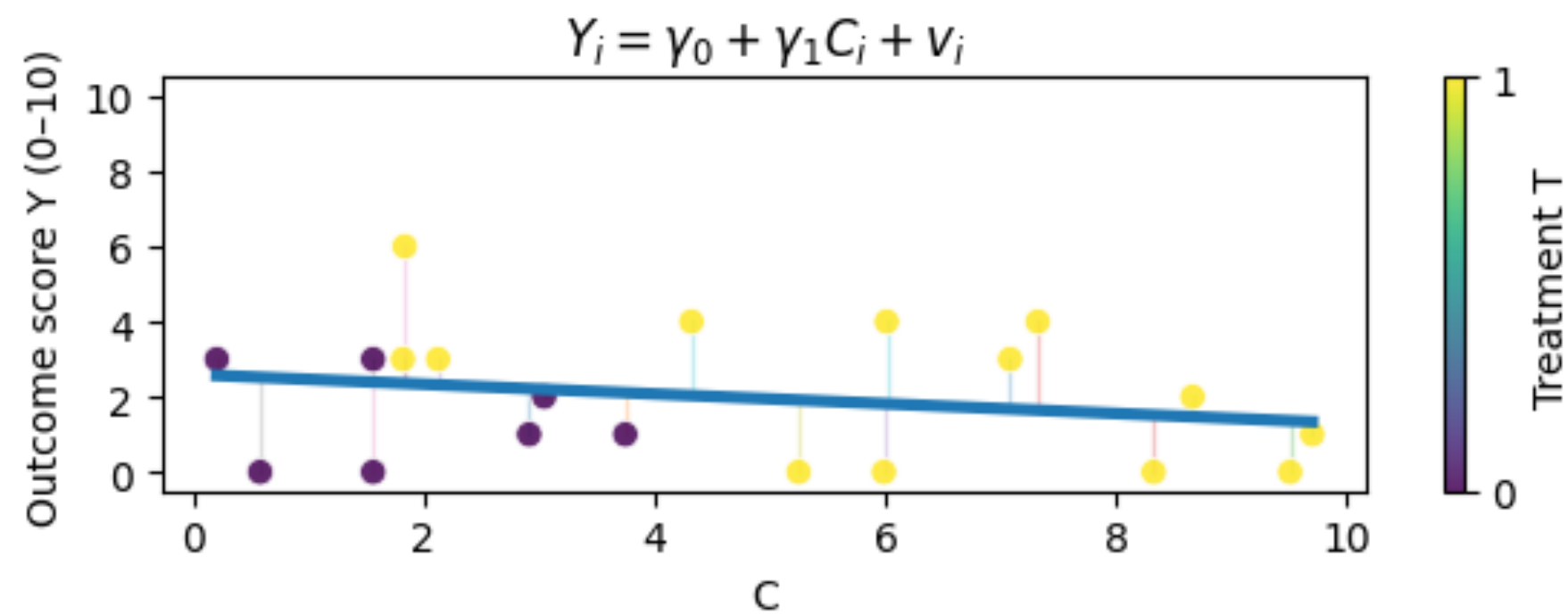
$$\tilde{Y}_i = \theta_0 + \theta_1 \tilde{T}_i + \eta_i$$



The Frisch-Waugh-Lovell Theorem



We calculate the fraction of units treated for each difficulty level!



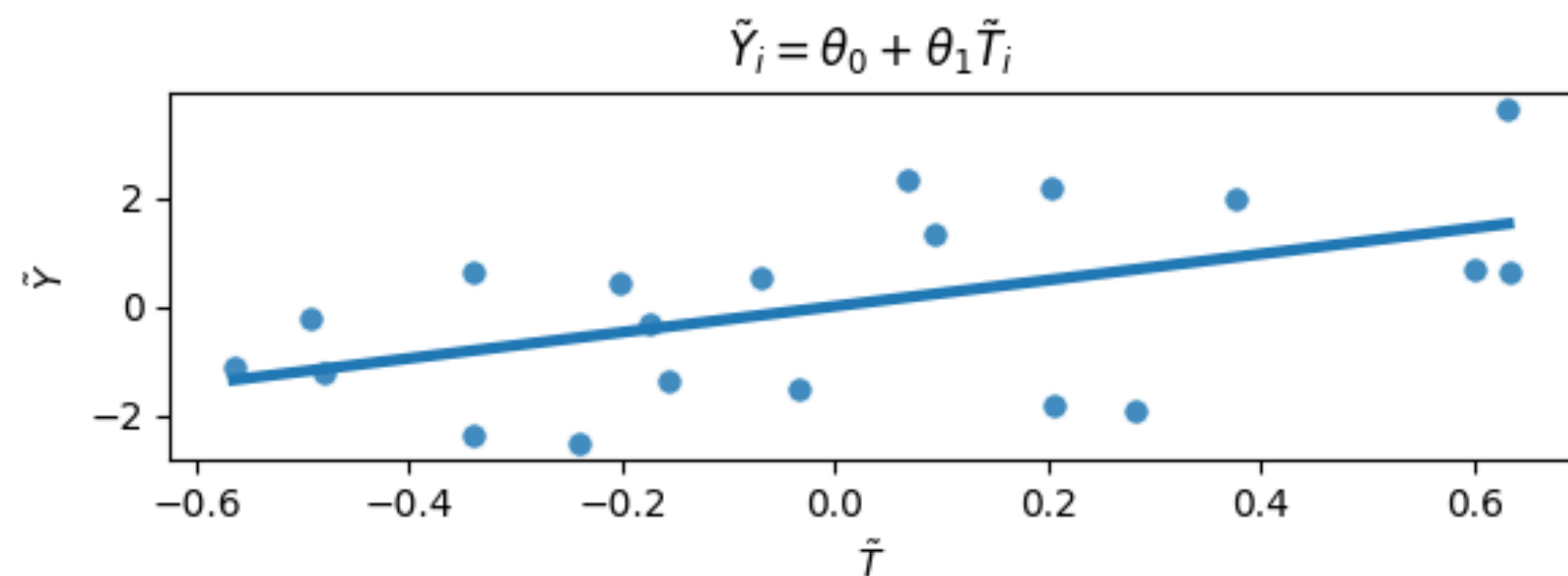
We calculate the average outcome for each treated level.

$$\tilde{T}_i = T_i - \alpha_0 - \alpha_1 C_i$$

$\tilde{T}_i$: we offset 0/1 by the predicted % of treated questions in C_i .

$$\tilde{Y}_i = Y_i - \gamma_0 - \gamma_1 C_i$$

$\tilde{Y}_i$: how much bigger/smaller Y_i is to the predicted average question in C_i .



When treatment happens more than expected, do outcomes also come out better than expected?

Another view

- Simple case, $Y_i = \beta_0 + \beta_1 T_i + \epsilon_i$

$$\hat{\beta}_1 = \bar{Y}_1 - \bar{Y}_0 = \frac{\text{Cov}(Y_i, T_i)}{\text{Var}(T_i)}$$

- Multivariate case, $Y_i = \boldsymbol{\beta}^T \mathbf{Z}_i + \epsilon_i = \beta_0 + \beta_1 T_i + \beta_2 X_1 + \beta_3 X_2 + \dots + \epsilon_i$

$$\hat{\beta}_1 = \frac{\text{Cov}(Y_i, \tilde{T}_i)}{\text{Var}(\tilde{T}_i)}$$

- Where $\tilde{T}_i$ is the residual for a regression of T_i on all other covariates.

Let's appreciate how cool this is

- $Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \dots + \epsilon_i$
- If we can predict T_i with the other variables, it means T_i is *not* random.
- However, it is impossible to (linearly) predict $\tilde{T}$ from the other variables by construction!
 - $\tilde{T}$ is the error term of $T_i = \alpha_0 + \alpha_1 X_{1,i} + \alpha_1 X_{2,i} + \dots + u_i$.
 - By construction, $\tilde{T}_i = u_i$, and $\hat{\alpha}_k = \text{Cov}(u_i, X_{k,i}) / \text{Var}(u_i) = 0$.
 - By the way we find the coefficients of the OLS, we have $\sum_i X_{k,i} \hat{u}_i = 0$
- Controlling makes treatment behave as if it were randomly assigned!



Give us some feedback!



[https://tinyurl.com/
reasoningwithdata](https://tinyurl.com/reasoningwithdata)