Regression

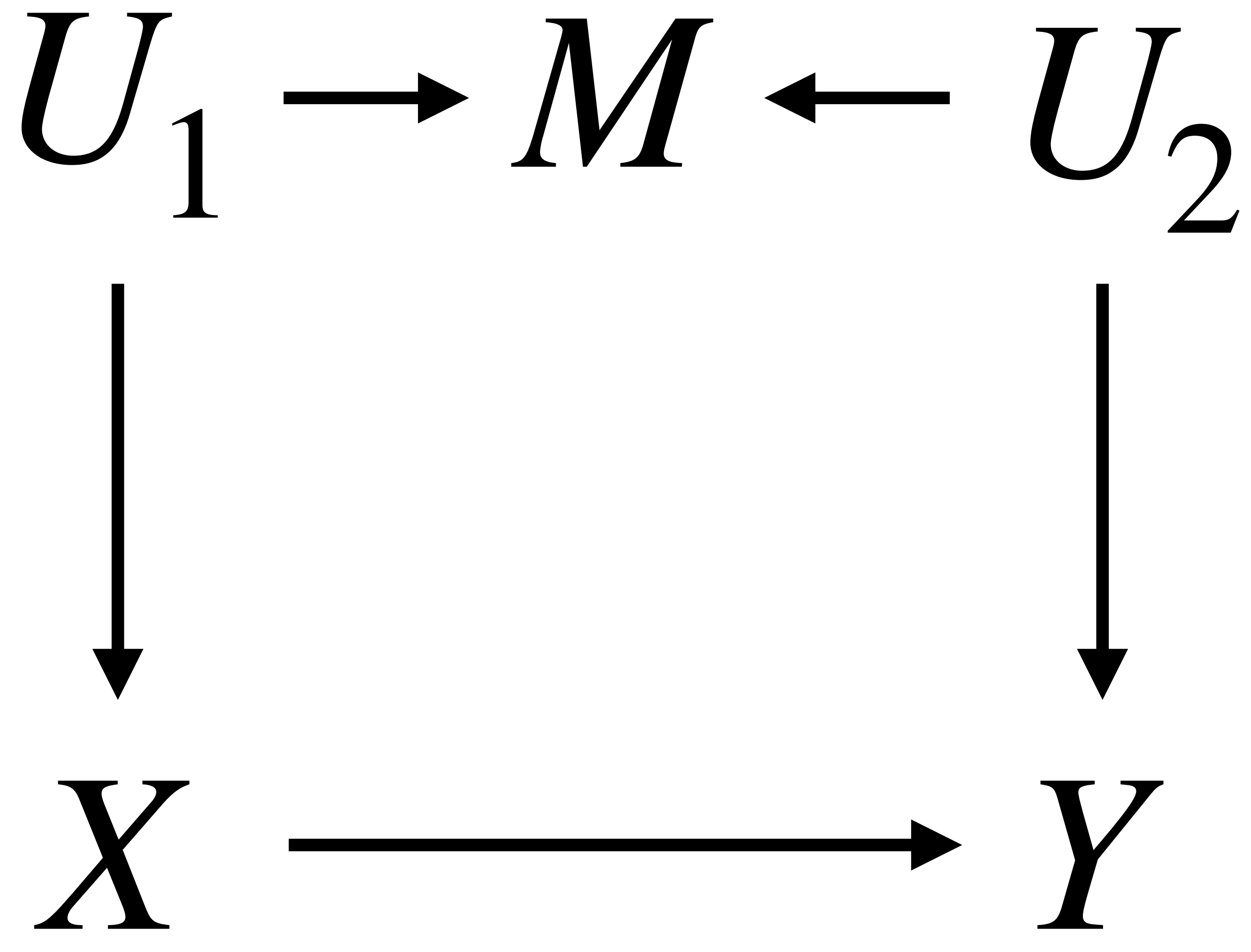
Manoel Horta Ribeiro
manoel@cs.princeton.edu

humans & machines lab

COS 598D / Spring 2026

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!



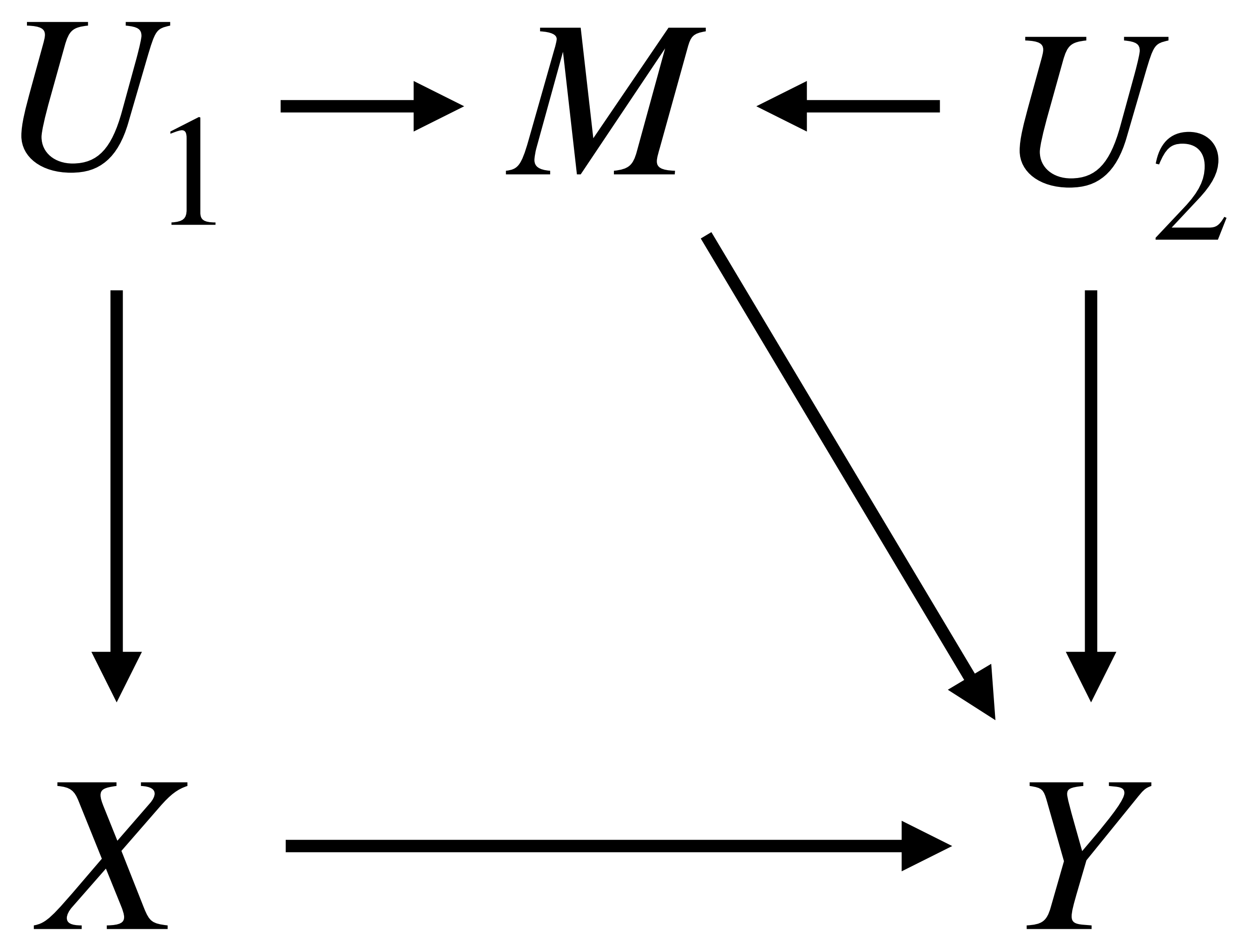
“Damned if you do, damned if you don’t.”

- Suppose that mothers’ diabetes influences their children through mechanisms other than genes (e.g., epigenetics).

- Controlling for M opens the path
  $X \to U_1 \to M \leftarrow U_2 \leftarrow Y$

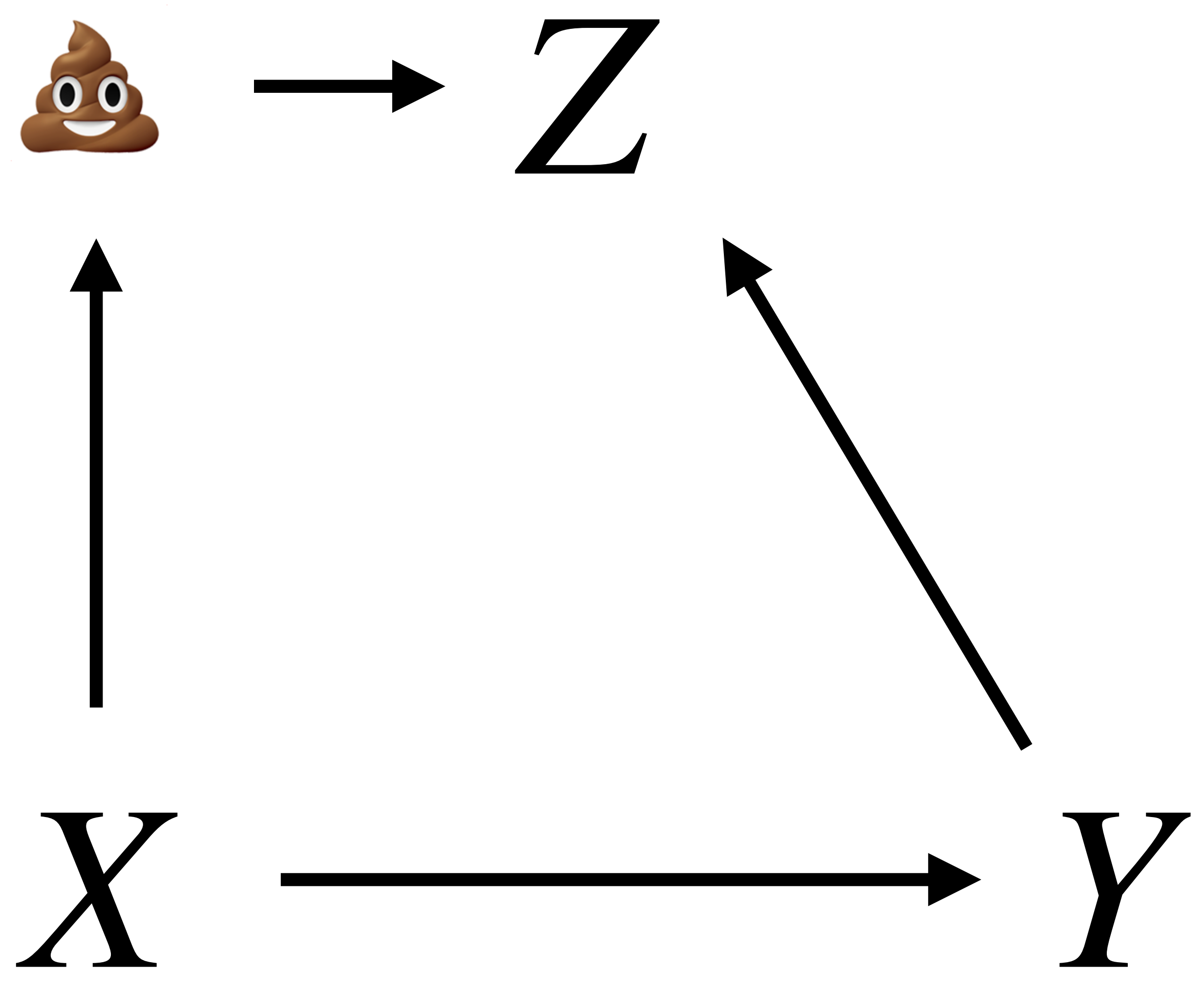
- Not doing so leaves open the path:
  $X \to U_1 \to M \to Y$

- We are toast!



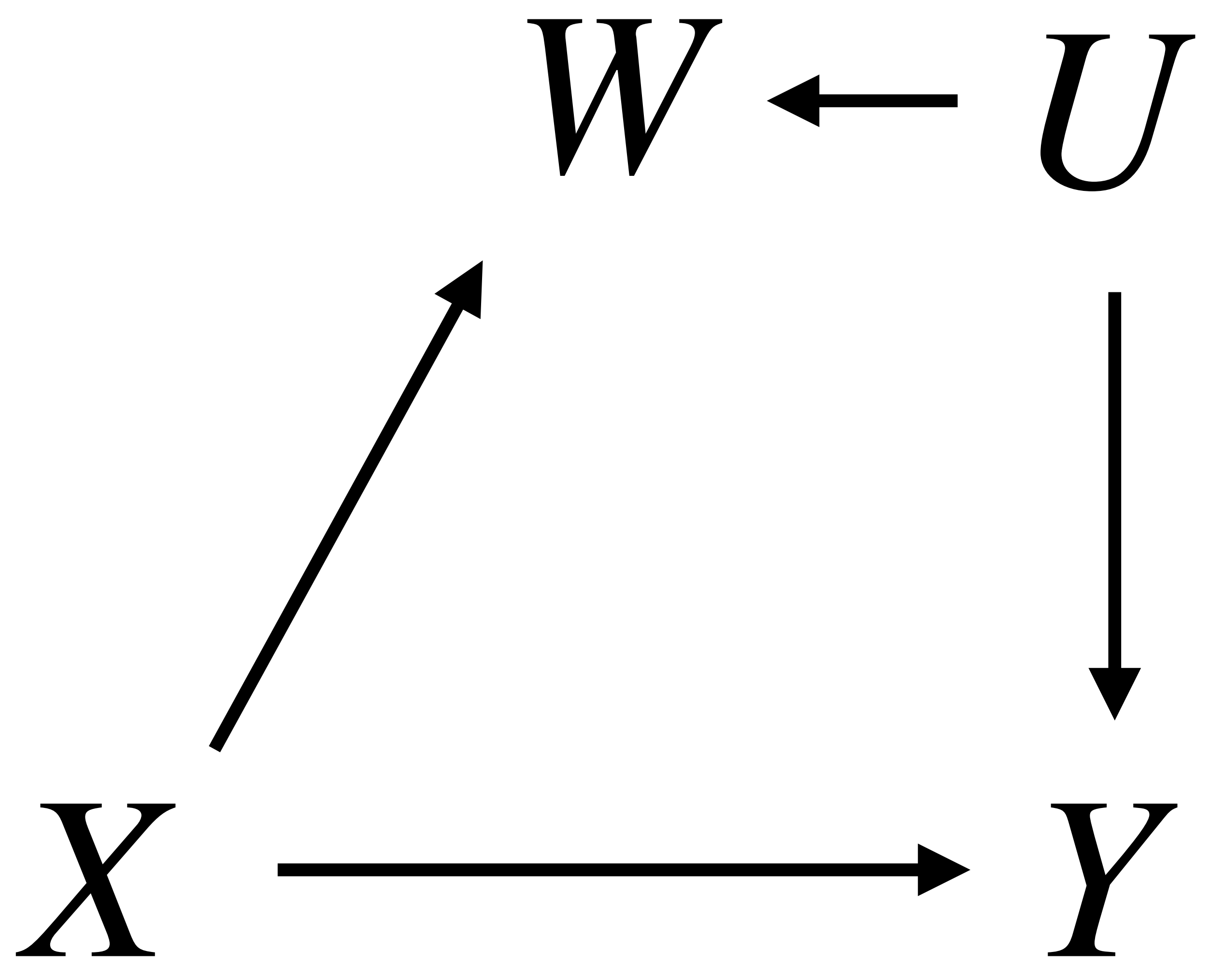
Case-control Bias

- 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 $X \rightarrow$ 💩 $\rightarrow Z \leftarrow Y$



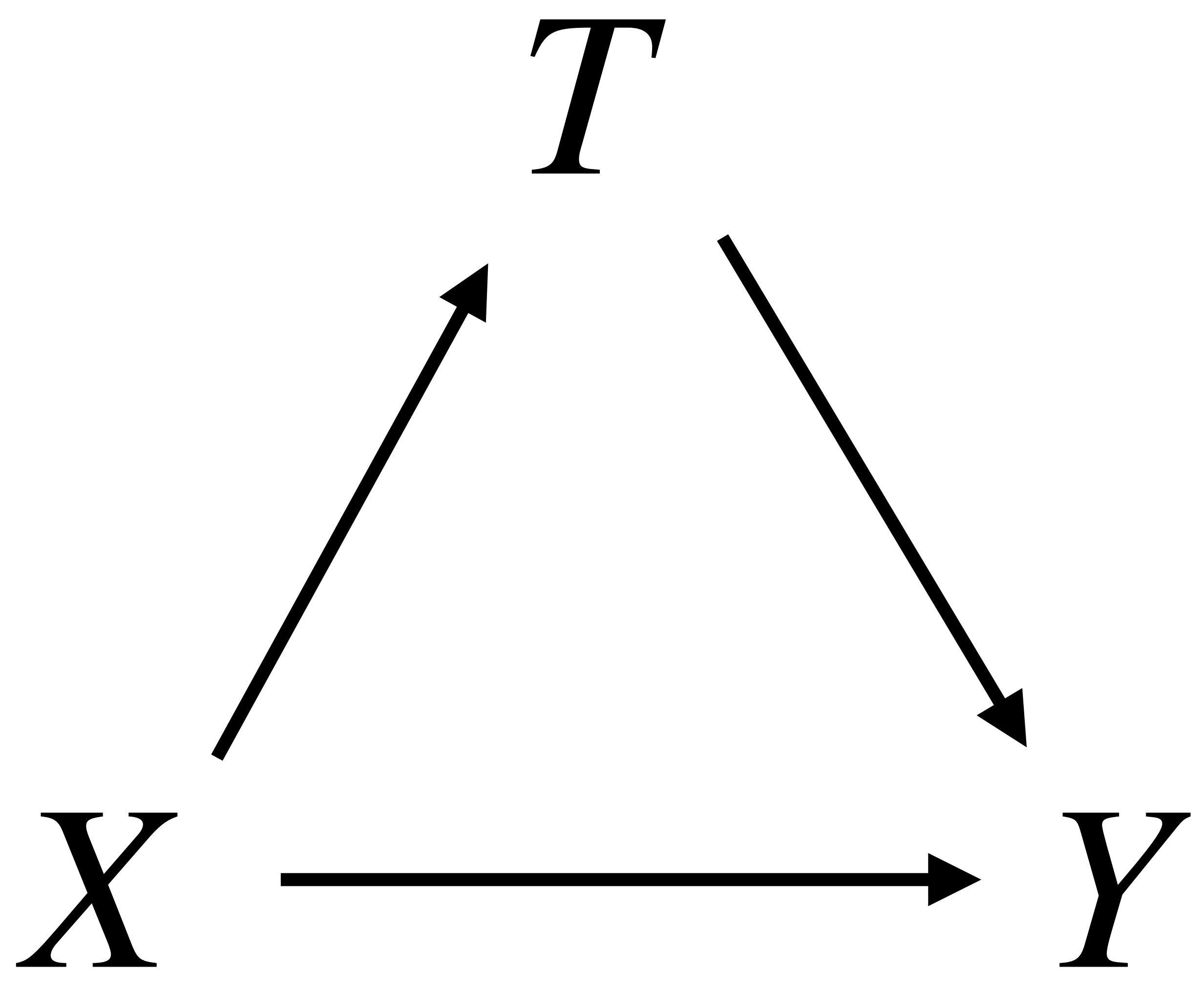
Selection Bias

- 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!



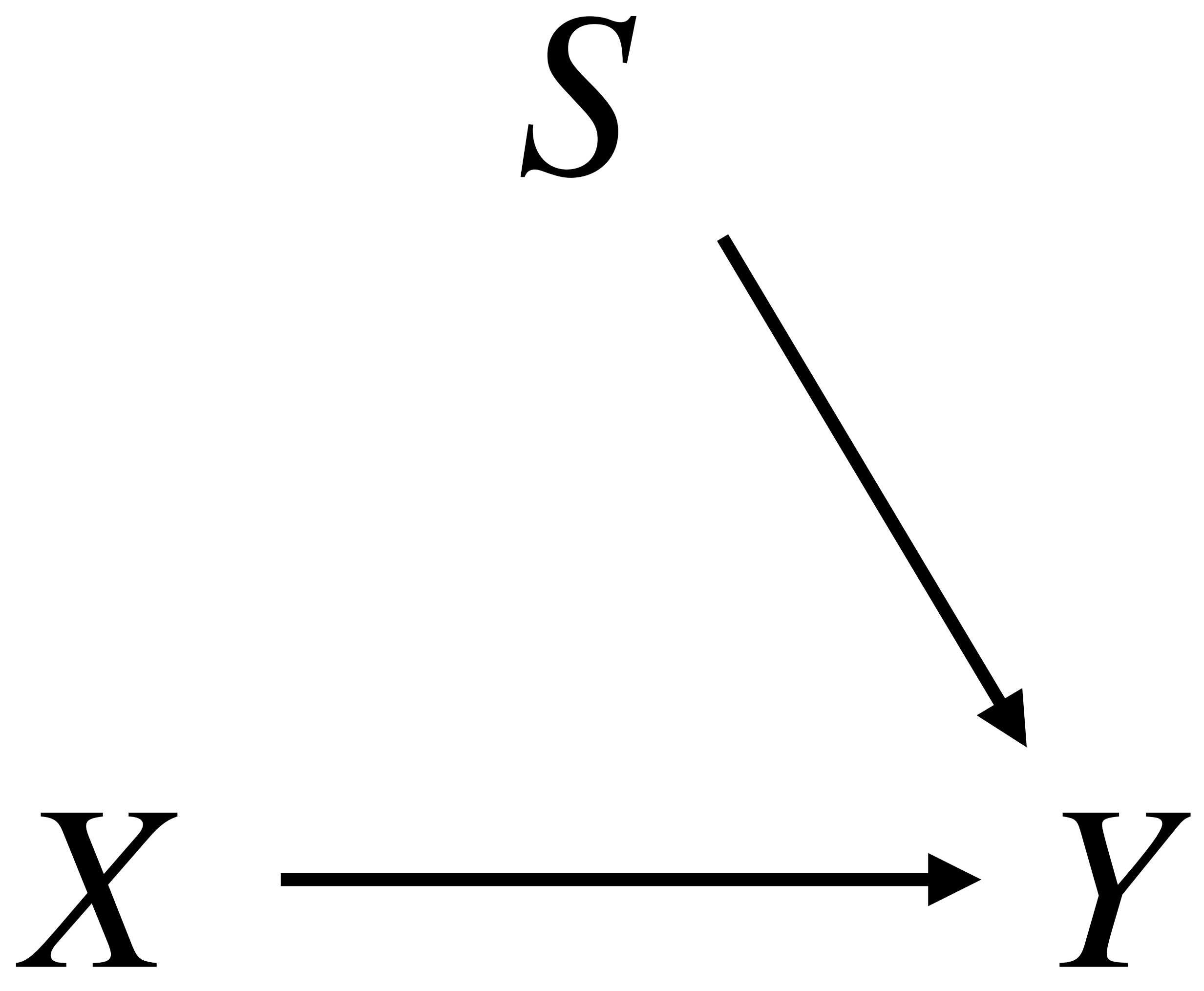
Overcontrolling

- 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 $X \to T \to Y$
  - In other ways $X \to Y$
- We block a legitimate path of the causal effect!



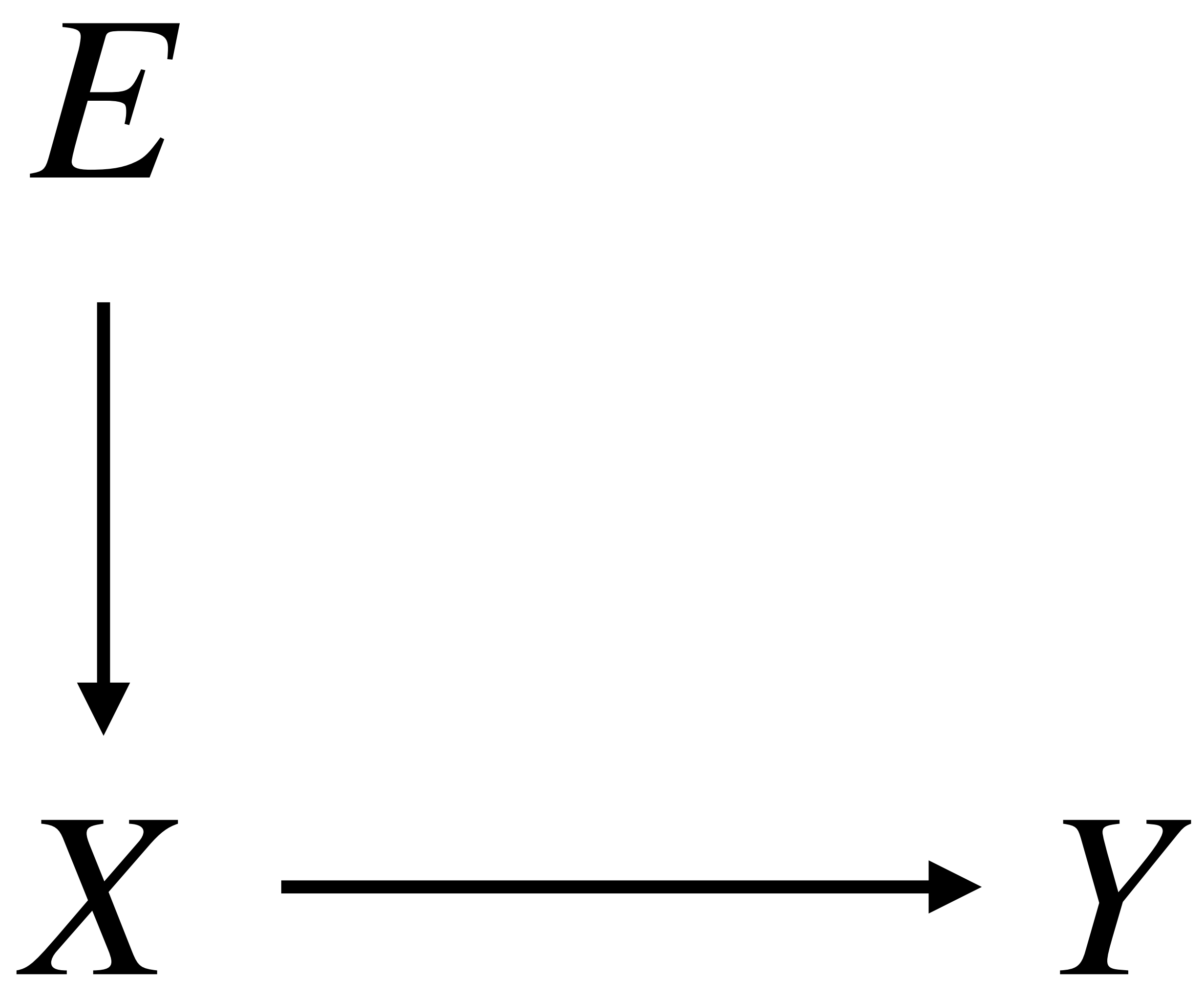
Reducing Variance

- 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

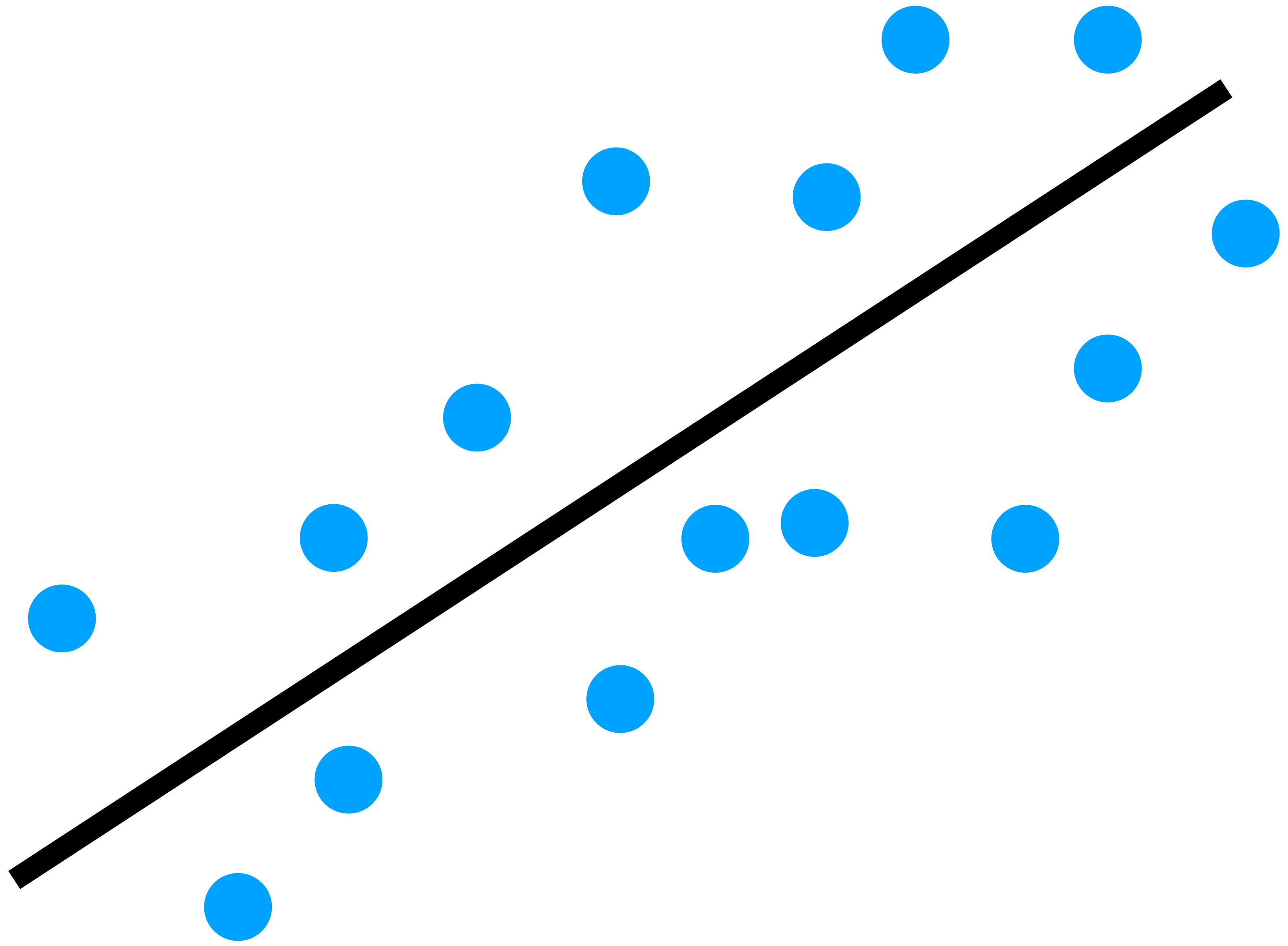
- We care about the effect of attending office hours X on the final exam Y.
- Students are randomly enrolled to receive reminders E, which strongly predict attendance!
- Controlling for E makes you compare attendees *vs.* non-attendees within reminder groups!
  - $E = 1$: almost everyone attended;
  - $E = 0$: 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

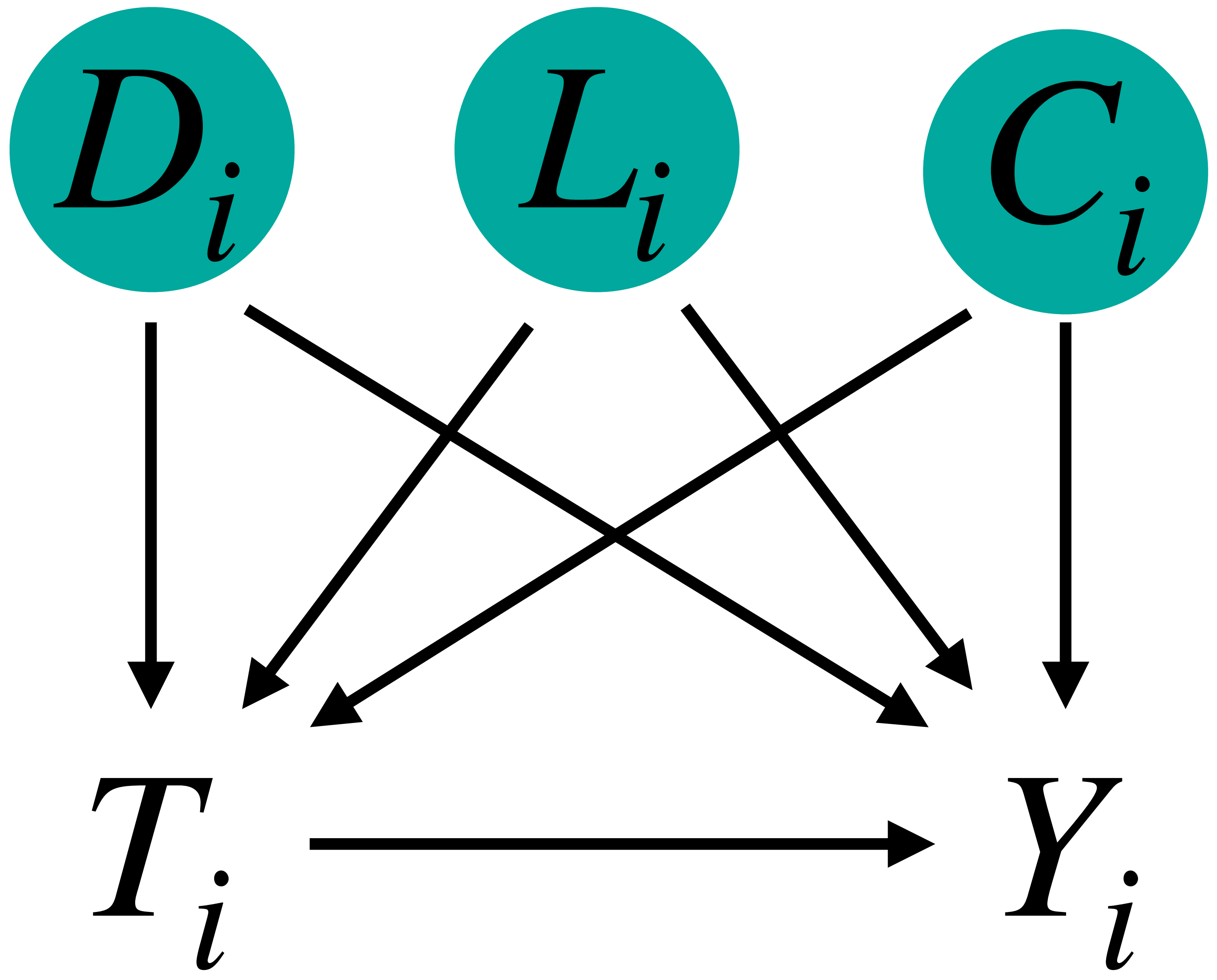
- Linear regression is typically described in CS as a *prediction* algorithm.
- But in other disciplines, it often used a powerful analytical tool to:
  - Describe data;
  - Estimate causal effects;



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"
- $D_i \in \{\text{math, biology}, \ldots\}$ —Domain



What do we need to control for?

Calculating the Treatment Effect

- Conditional exchangeability saves us:

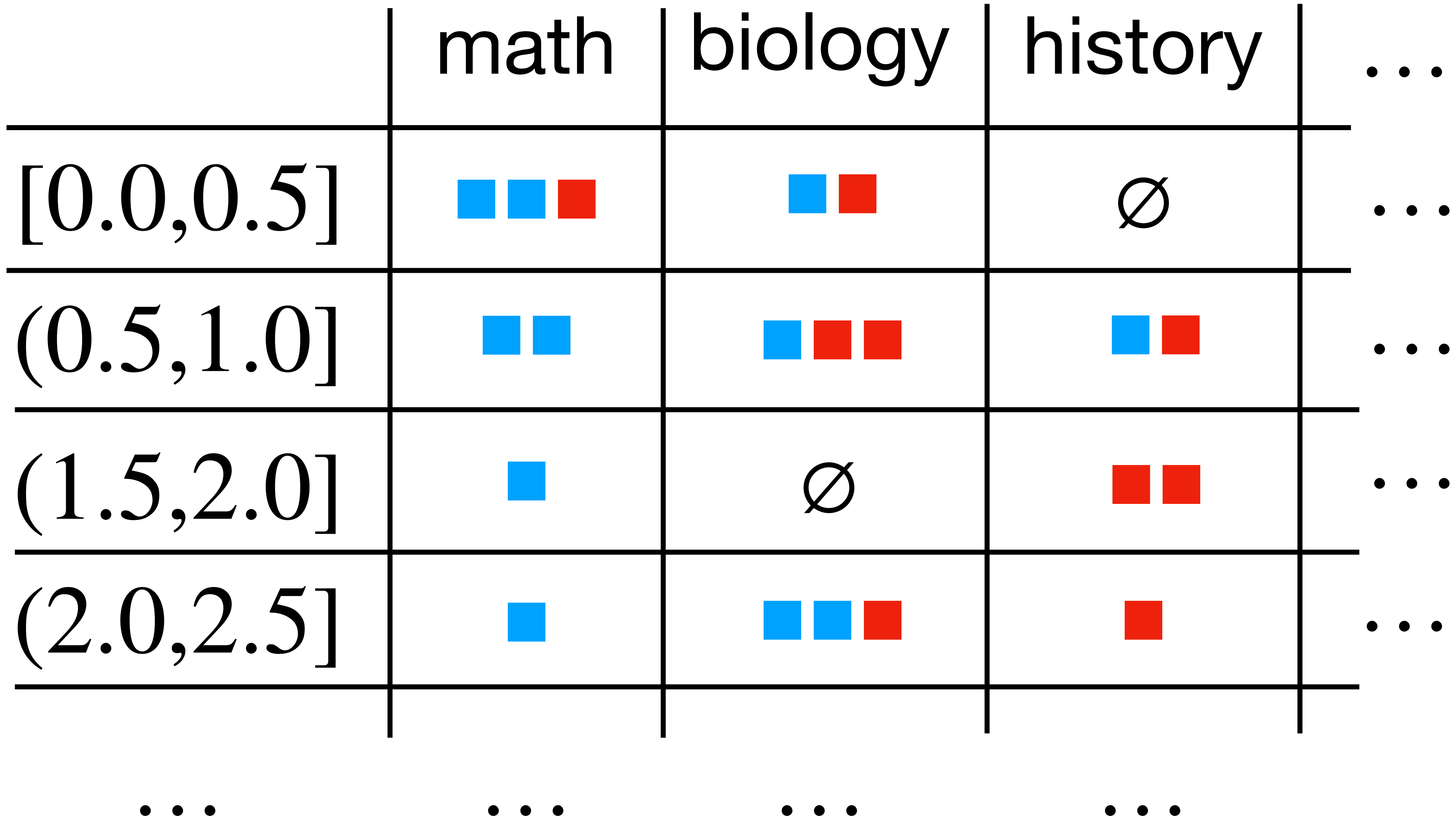
$$\tau = E_{D,L,C}\Big[E\big[Y \mid D, L, C, T = 1\big]\Big] - E_{D,L,C}\Big[E\big[Y \mid D, L, C, T = 0\big]\Big]$$

- 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	$\dots$
$[0.0, 0.5]$				$\dots$
$(0.5, 1.0]$				$\dots$
$(1.5, 2.0]$				$\dots$
$(2.0, 2.5]$				$\dots$
$\dots$	$\dots$	$\dots$	$\dots$	

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!



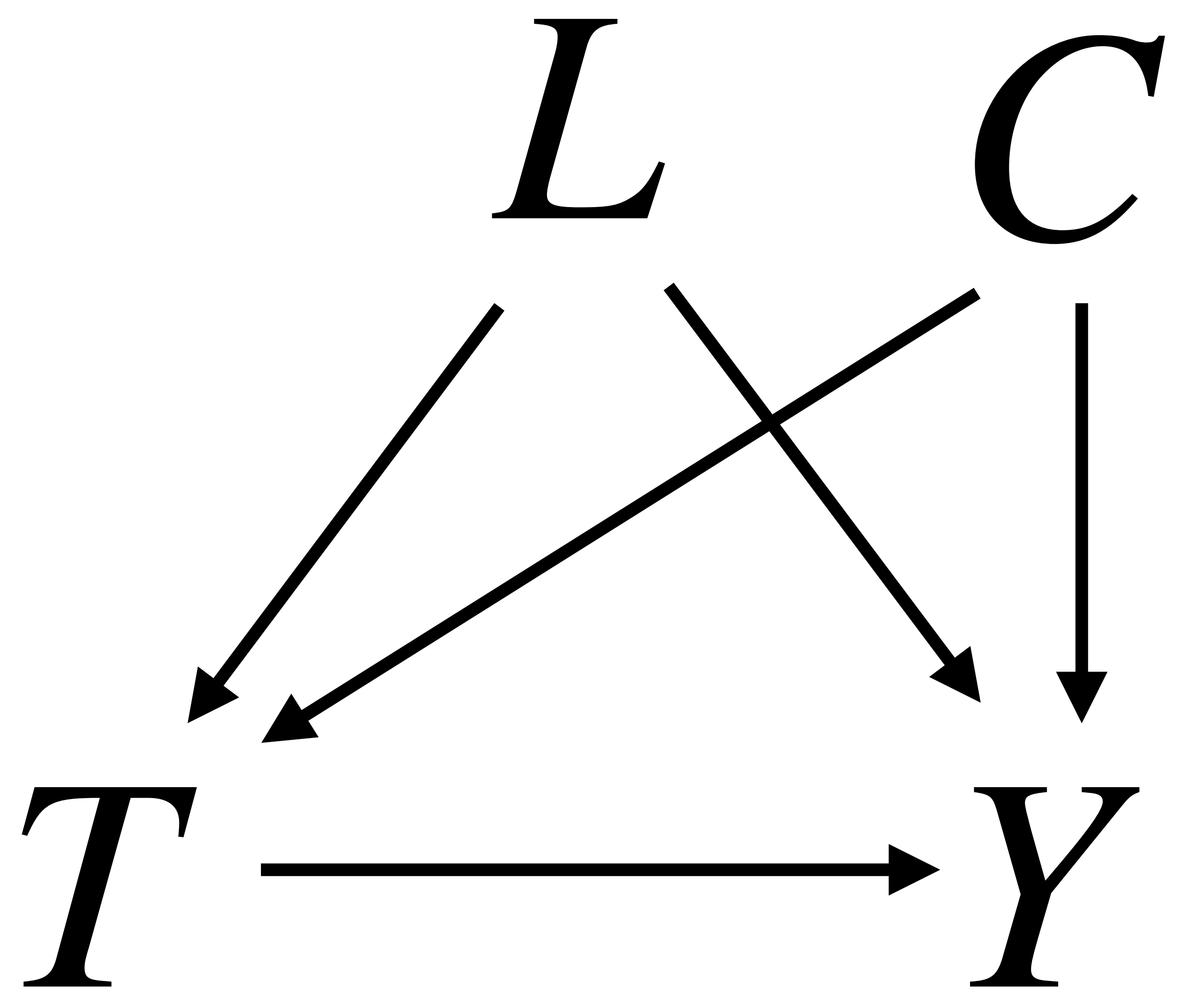
Idea: Assume a Model

- *E*stimate $E\left[Y \mid C, L, T\right]$ parametrically.

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

- Using a model fills the gaps by extrapolation!

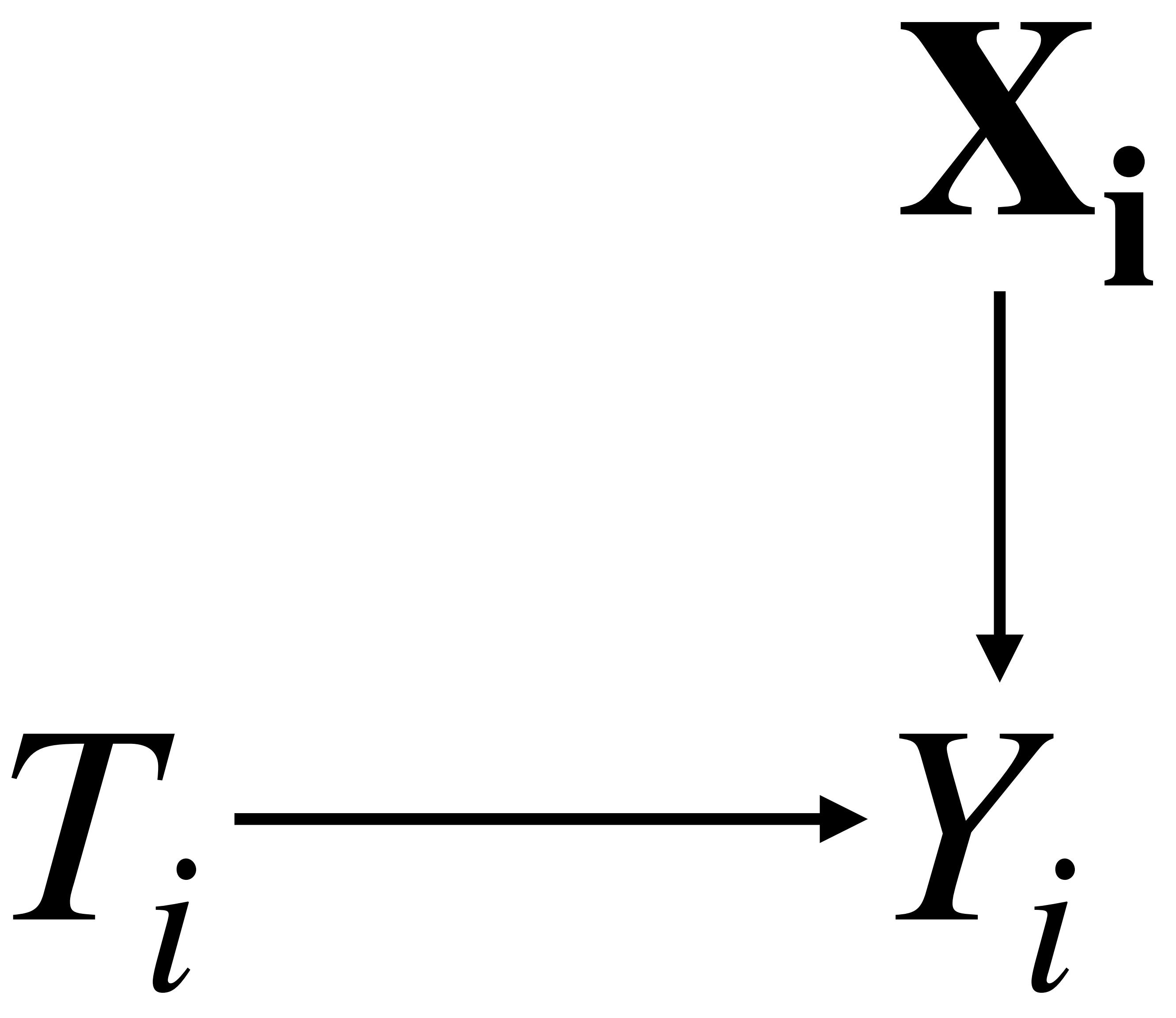
- Correctness depends on the modeling assumptions!
  - E.g., linearity, additivity.



Backtracking: Experimental Case

- Assume it is an experiment, and we don't have access to confounders $\mathbf{X}$.
- We observe $\langle Y_i, T_i \rangle$ $i = 1, \ldots, n$ 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

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

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

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

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

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

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

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

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

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

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

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

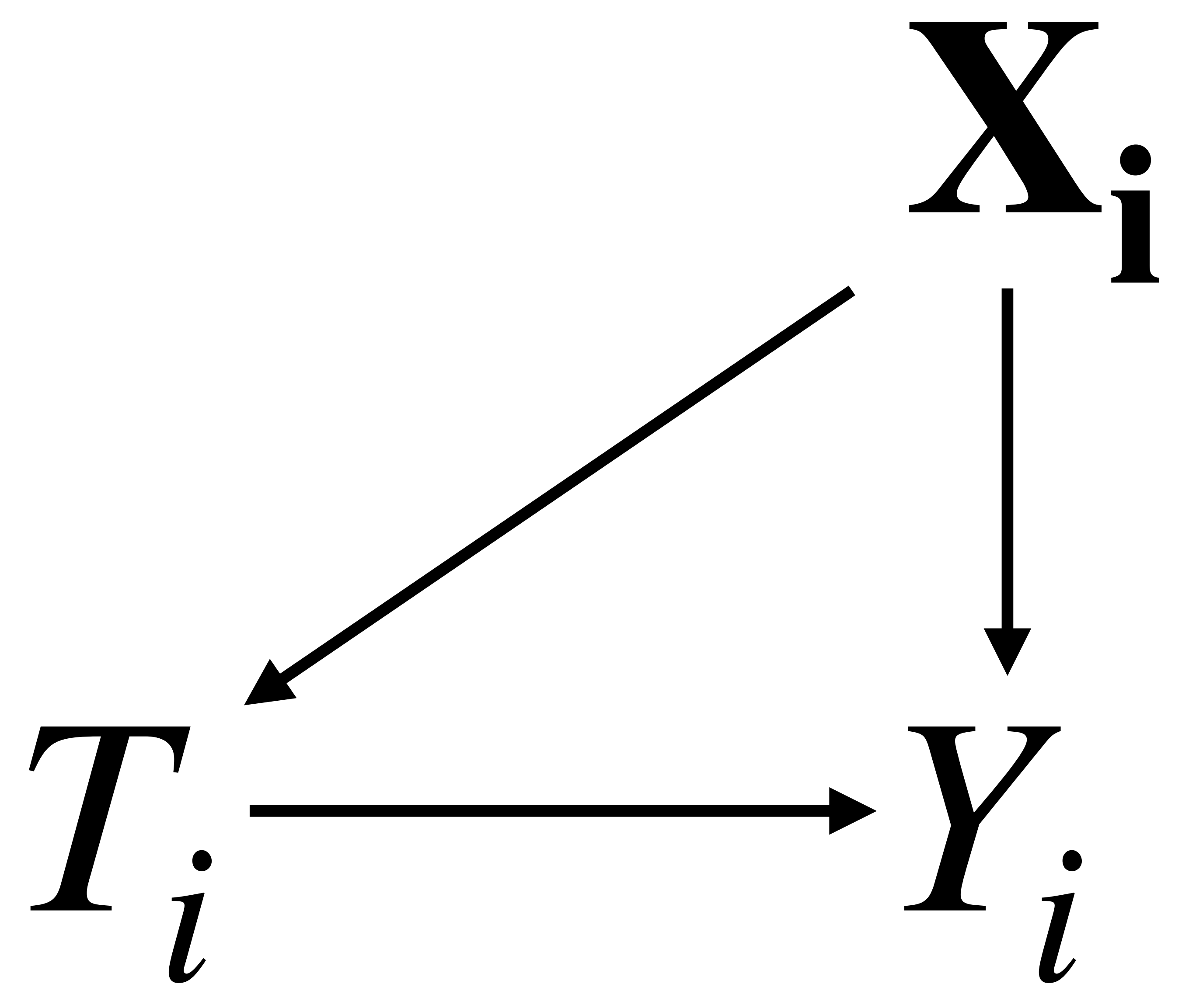
What if we have controls?

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

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

- We assume the model:

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



Ordinary Least Squares (Multivariate)

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

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

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

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

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

$$\hat{\boldsymbol{\beta}} \;=\; (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top Y$$

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:

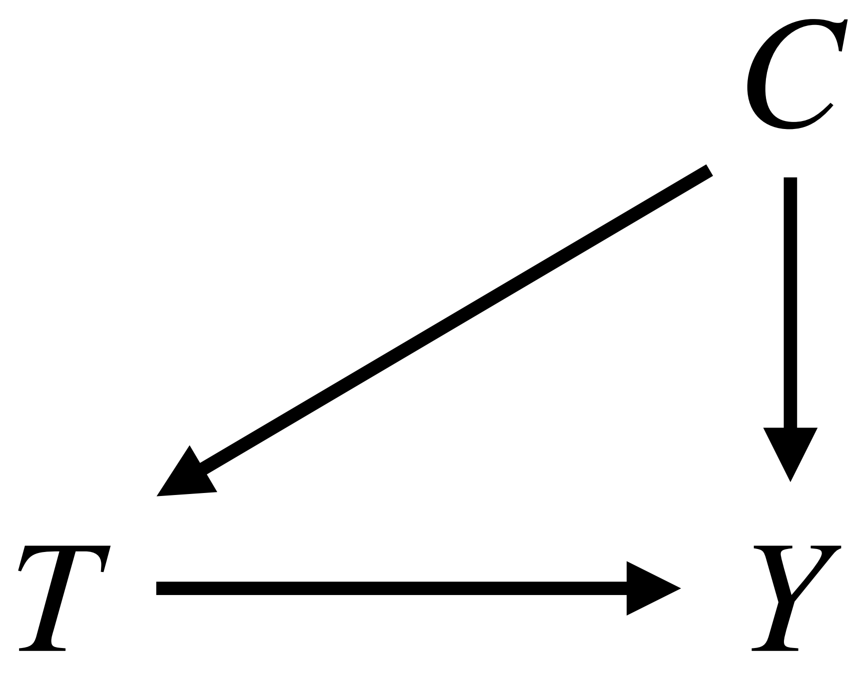
$$T_i = \alpha_0 + \alpha_1 C_i + u_i, \quad Y_i = \gamma_0 + \gamma_1 C_i + v_i$$

- Then, computing:

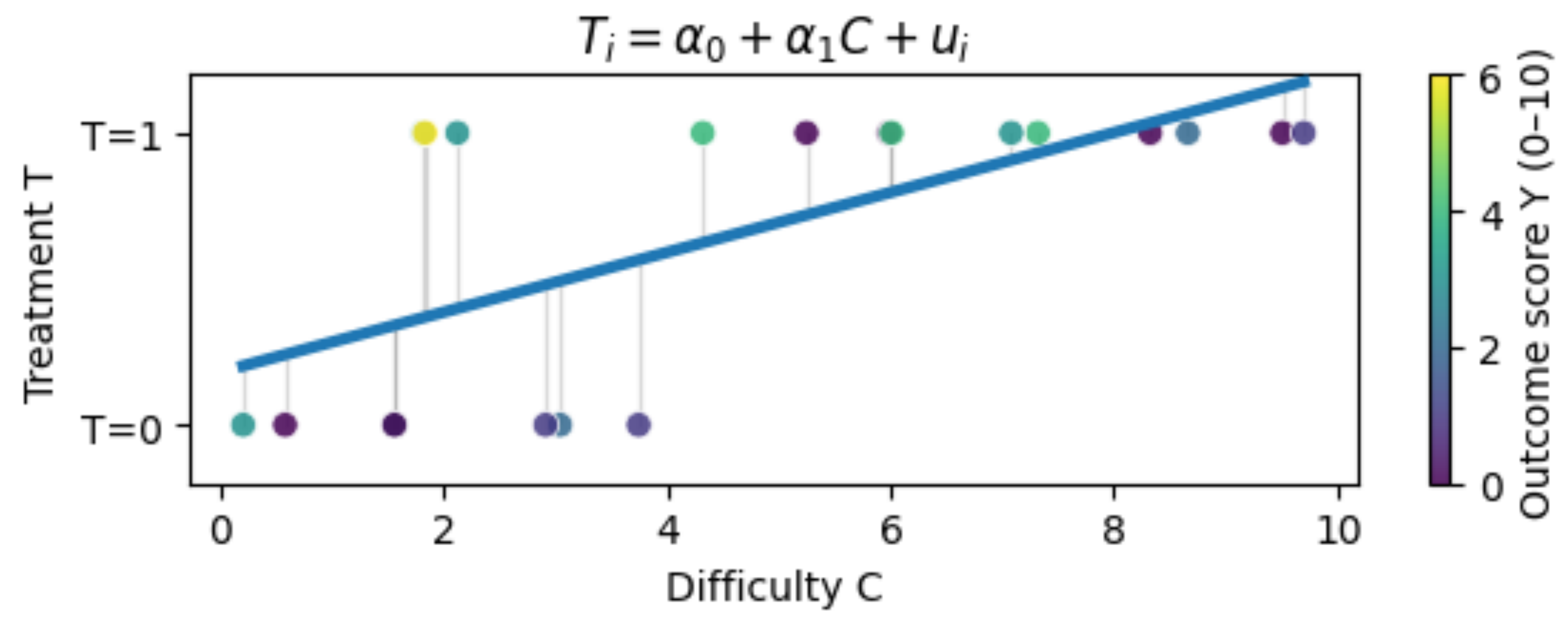
$$\tilde{T}_i = T_i - \alpha_0 + \alpha_1 C_i, \quad \tilde{Y}_i = Y_i - \gamma_0 + \gamma_1 C_i$$

- And then, regressing

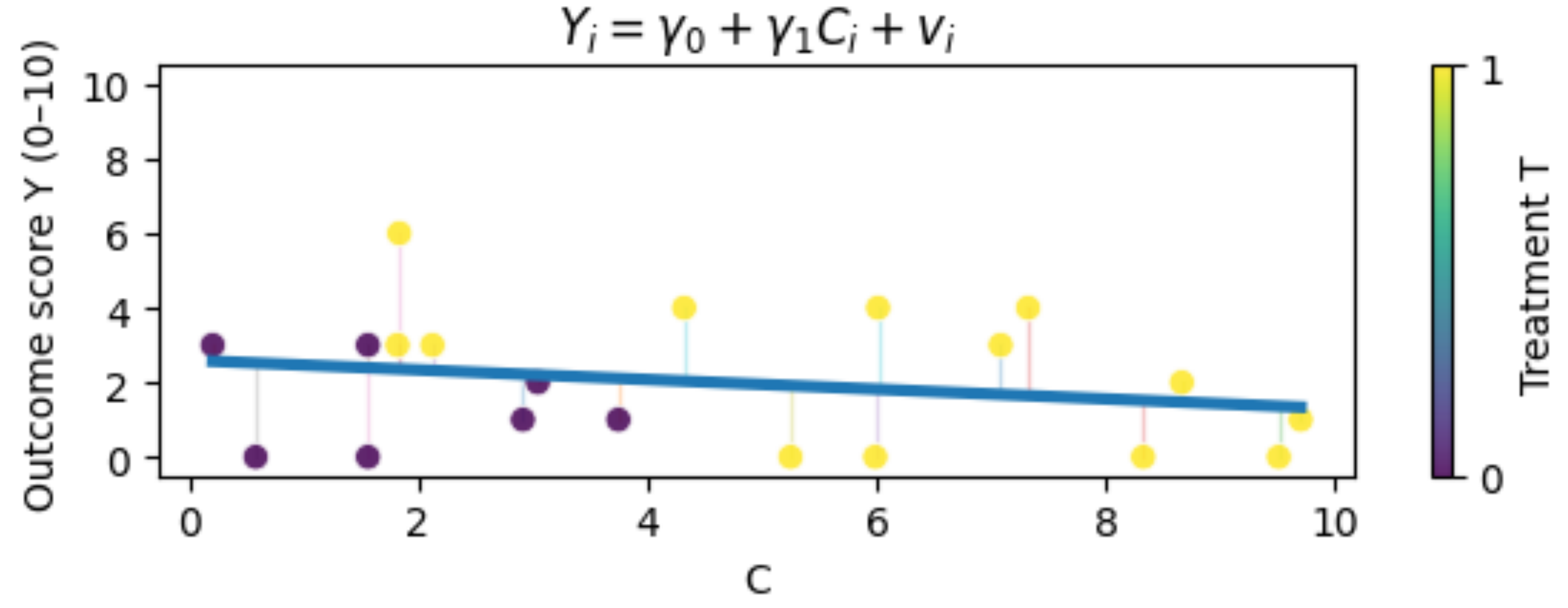
$$\tilde{Y}_i = \theta_0 + \theta_1 \tilde{T}_i + \eta_i$$



The Frisch-Waugh-Lovell Theorem



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



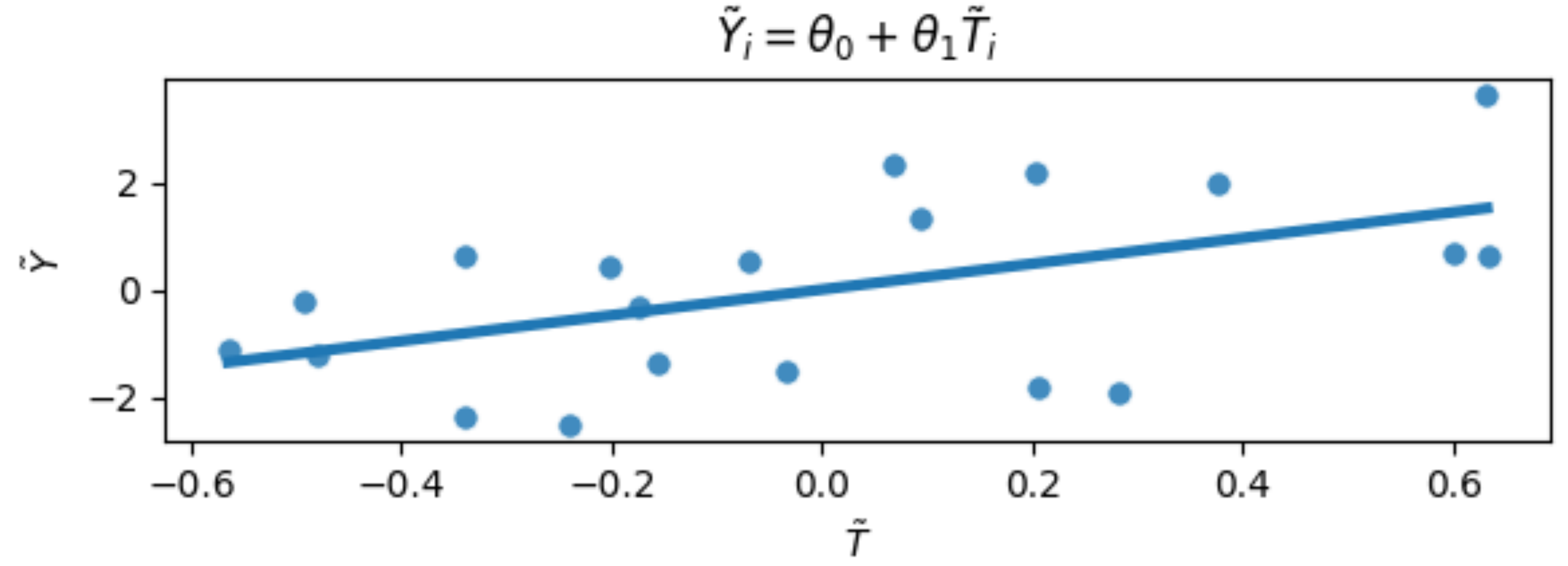
We calculate the average outcome for each treated level.

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

$\tilde{T}_i$: 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, $Y_i = \boldsymbol{\beta}^T \mathbf{Z}_i + \epsilon_i = \beta_0 + \beta_1 T_i + \beta_2 X_1 + \beta_3 X_2 + \ldots + \epsilon_i$

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

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

Let's appreciate how cool this is

- $Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_{1,i} + \beta_3 X_{2,i} + \ldots + \epsilon_i$
- 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 $T_i = \alpha_0 + \alpha_1 X_{1,i} + \alpha_1 X_{2,i} + \ldots + u_i$.
    - 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!

Statistical Tests in a Regression

- A regression estimates *coefficients*. $Y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 C_i + u_i$
- And it turns out that each of these coefficients is normally distributed! (Trust me! I won't prove this).
- We can do Z tests on these coefficients!

$$H_0 : \ \alpha_1 = 0, \quad H_1 : \ \alpha_1 \neq 0, Z \ = \ \frac{\hat{\alpha}_1 - 0}{SE(\hat{\alpha}_1)}$$

- We can do a Z-test using only the part of T_i that C_i can't explain!

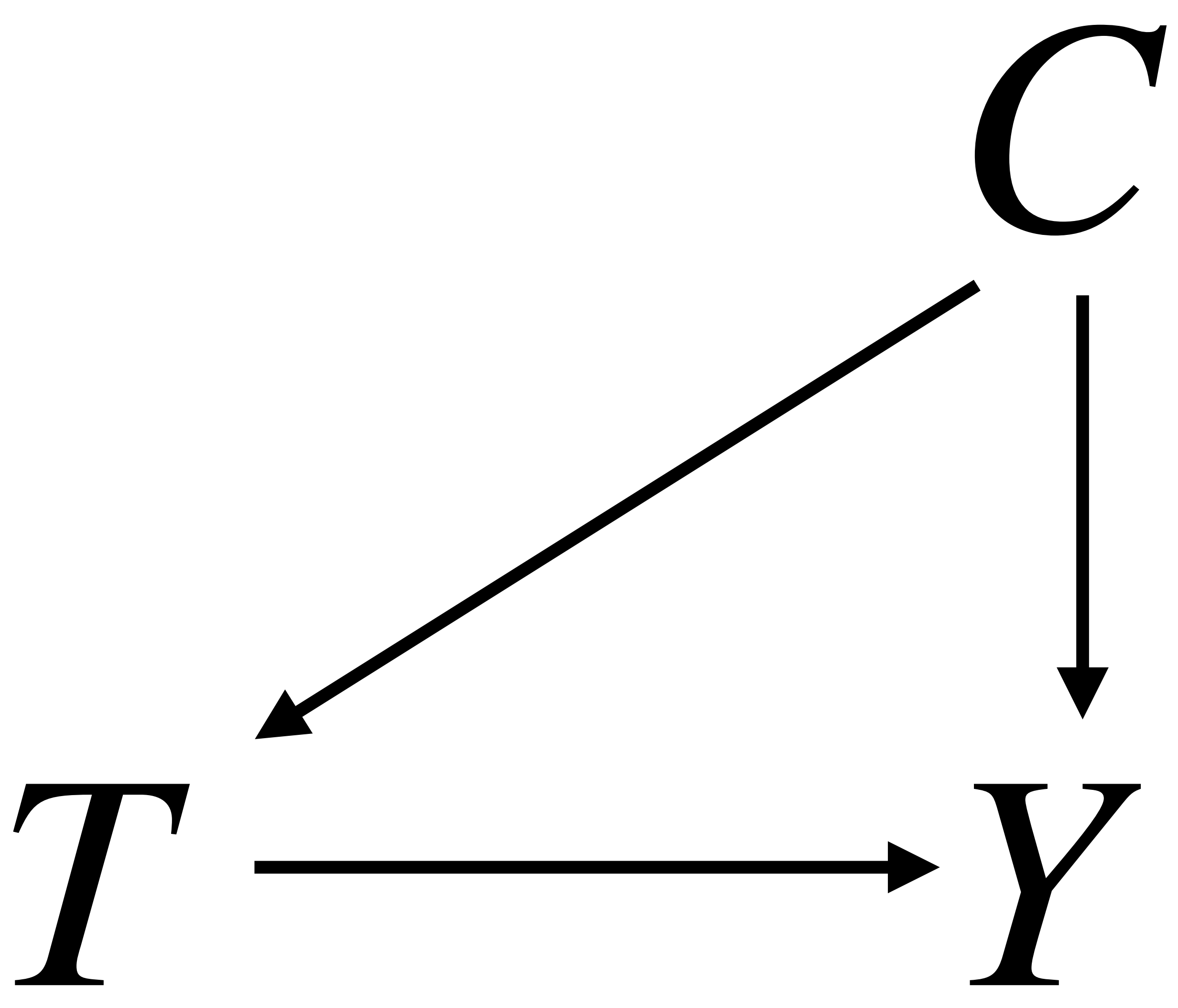
What if our model is wrong?

- What if we ignore C? We consider:

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

- When in reality, we have:

$$Y_i = \alpha_0 + \alpha_1 T_i + \alpha_2 C_i + u_i$$



- We can use our formula…

$$\hat{\beta}_1 = \frac{\text{Cov}(Y_i, T_i)}{\text{Var}(T_i)} = \frac{\text{Cov}(\alpha_0 + \alpha_1 T_i + \alpha_2 C_i + u_i, T_i)}{\text{Var}(T_i)} = \alpha_1 + \alpha_2 \underbrace{\frac{\text{Cov}(C_i, T_i)}{\text{Var}(T_i)}}_{\text{Omitted Variable Bias}}$$

What if our model is wrong?

$$\hat{\beta}_1 = \frac{\text{Cov}(Y_i, T_i)}{\text{Var}(T_i)} = \frac{\text{Cov}(\alpha_0 + \alpha_1 T_i + \alpha_2 C_i + u_i, T_i)}{\text{Var}(T_i)} = \alpha_1 + \alpha_2 \frac{\text{Cov}(C_i, T_i)}{\text{Var}(T_i)}$$

"Short equals long
plus the effect of omitted
times the regression of omitted on included"
— Joshua Angrist

- Causal inference with *non-random* or observational data should be taken with a grain of salt! What if there's unobserved variable bias?

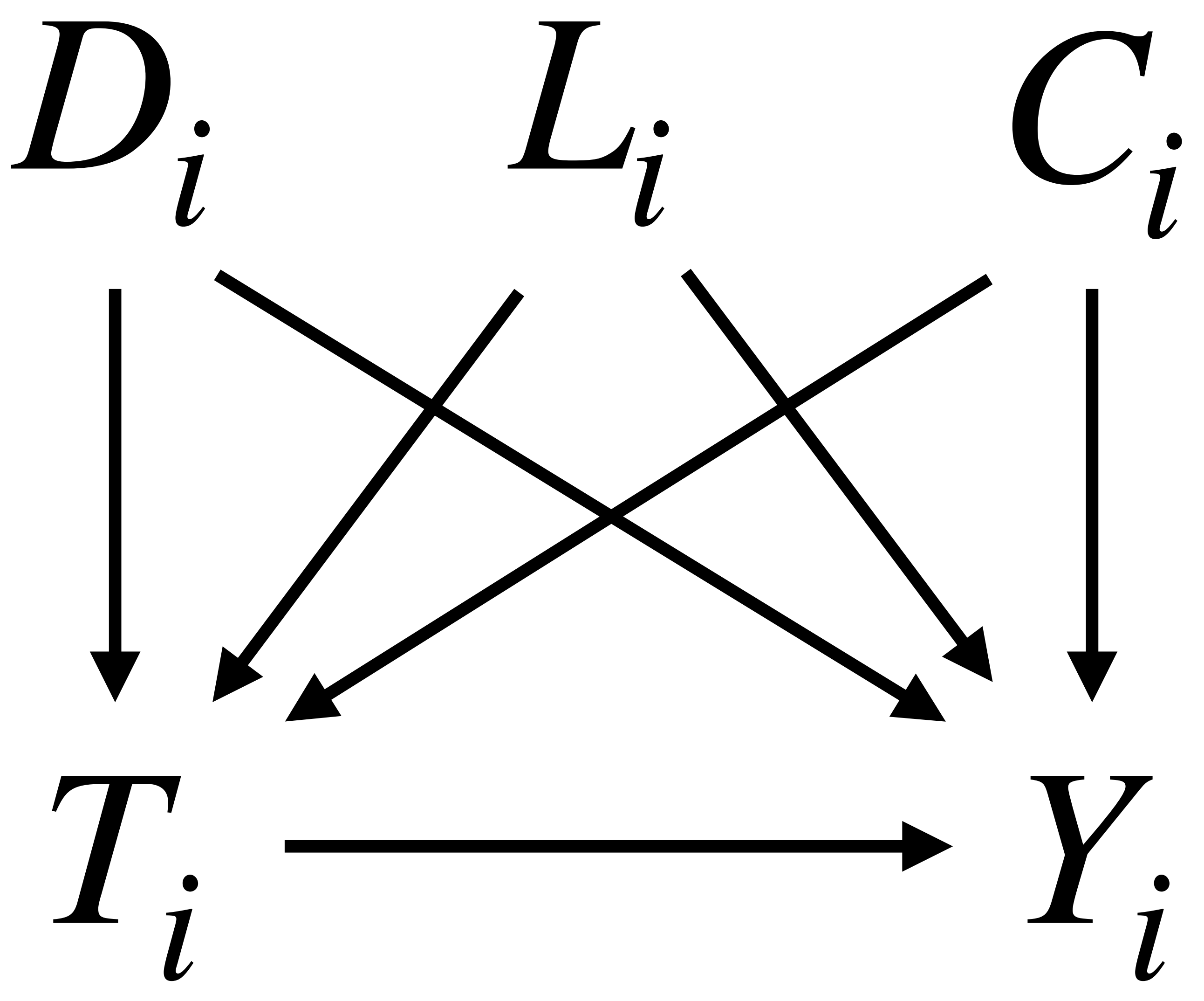
Dummy Variables

So far, we considered two types of variables:

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

But how do we deal with something like:

- $D_i \in \{\text{math, biology}, \ldots\}$ —Domain



Naïve Idea

- Create a different variable for each level of the categorical variable

- Instead of $D_i \in \{\text{math, biology, } \dots\}$, we can instead have:
  - $D_i^{\text{math}} \in \{0,1\}$
  - $D_i^{\text{biology}} \in \{0,1\}$
  - …

D_i	D_i^{math}	D_i^{bio}	D_i^{hist}	$\cdots$
math	1	0	0	$\cdots$
biology	0	1	0	$\cdots$
math	1	0	0	$\cdots$
history	0	0	1	$\cdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$	

Ordinary Least Squares (Multivariate)

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

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

- Given conve

A square matrix is not invertible if one of its columns can be written as a linear combination of the other columns.

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

$$D_i^{\text{math}} = 1 - D_i^{\text{bio}} - D_i^{\text{hist}} - \dots$$

Less Naïve Idea: Dummy Encoding

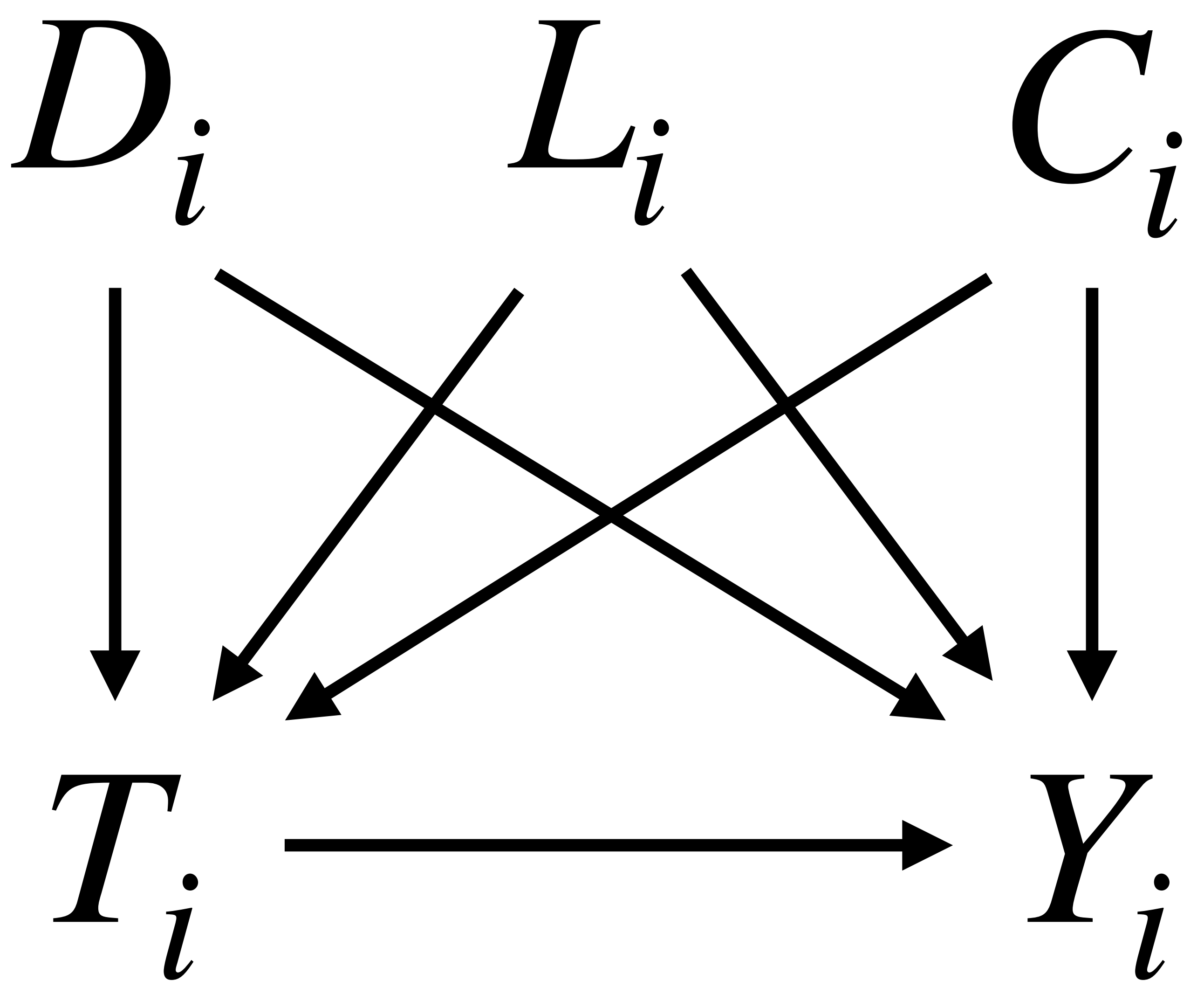
- Choose one variable as "the reference."
- Create a different variable for each *other level* of the categorical variable.
- Note that $D_i^{\text{hist}} \sim \text{Bernoulli}(p^{\text{hist}})$

D_i	D_i^{bio}	D_i^{hist}	$\cdots$
math	0	0	$\cdots$
biology	1	0	$\cdots$
math	0	0	$\cdots$
history	0	1	$\cdots$
$\cdots$	$\cdots$	$\cdots$	

- Consider two categories: math and history

$$Y_i = \beta_0 + \beta_1 D_i^{\text{Hist}} + \epsilon_i$$

- What will the β_0 β_1 capture?
  - p^{math} / p^{hist} = % of questions per category
  - μ^{math} / μ^{hist} = average outcome per category



$$\beta_1 = \frac{\text{Cov}(D_i^{\text{hist}}, Y_i)}{\text{Var}(D_i^{\text{hist}})} = \frac{p^{\text{hist}}\mu^{\text{hist}} - p^{\text{hist}}\mu}{p^{\text{hist}}(1 - p^{\text{hist}})} = \mu^{\text{hist}} - \mu^{\text{math}}$$

$$\beta_0 = E[Y_i] - \beta_1 E[D_i^{\text{hist}}] = \mu - p^{\text{hist}}(\mu^{\text{hist}} - \mu^{\text{math}}) = \mu^{\text{math}}$$

$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]$

$X \sim \text{Bernoulli}(p) \rightarrow E[X] = p$
$\rightarrow Var(X) = p(1 - p)$

Encoding rocks

- This generalizes to any number of levels!

$$Y_i = \beta_0 + \beta_1 D_i^{\text{Hist}} + \beta_2 D_i^{\text{Bio}} \epsilon_i$$

- β_1 captures how much the average outcome from the history question differs from the math one.
- β_2 captures how much the average outcome from the biology question differs from the math one.

D_i	D_i^{bio}	D_i^{hist}	$\cdots$
math	0	0	$\cdots$
biology	1	0	$\cdots$
math	0	0	$\cdots$
history	0	1	$\cdots$
$\cdots$	$\cdots$	$\cdots$	

Sum Encoding

$$Y_i = \beta_0 + \beta_1 I_{1,i} + \beta_2 I_{2,i} + \epsilon_i$$

β_0 captures the grand mean, the mean of the means of the groups.

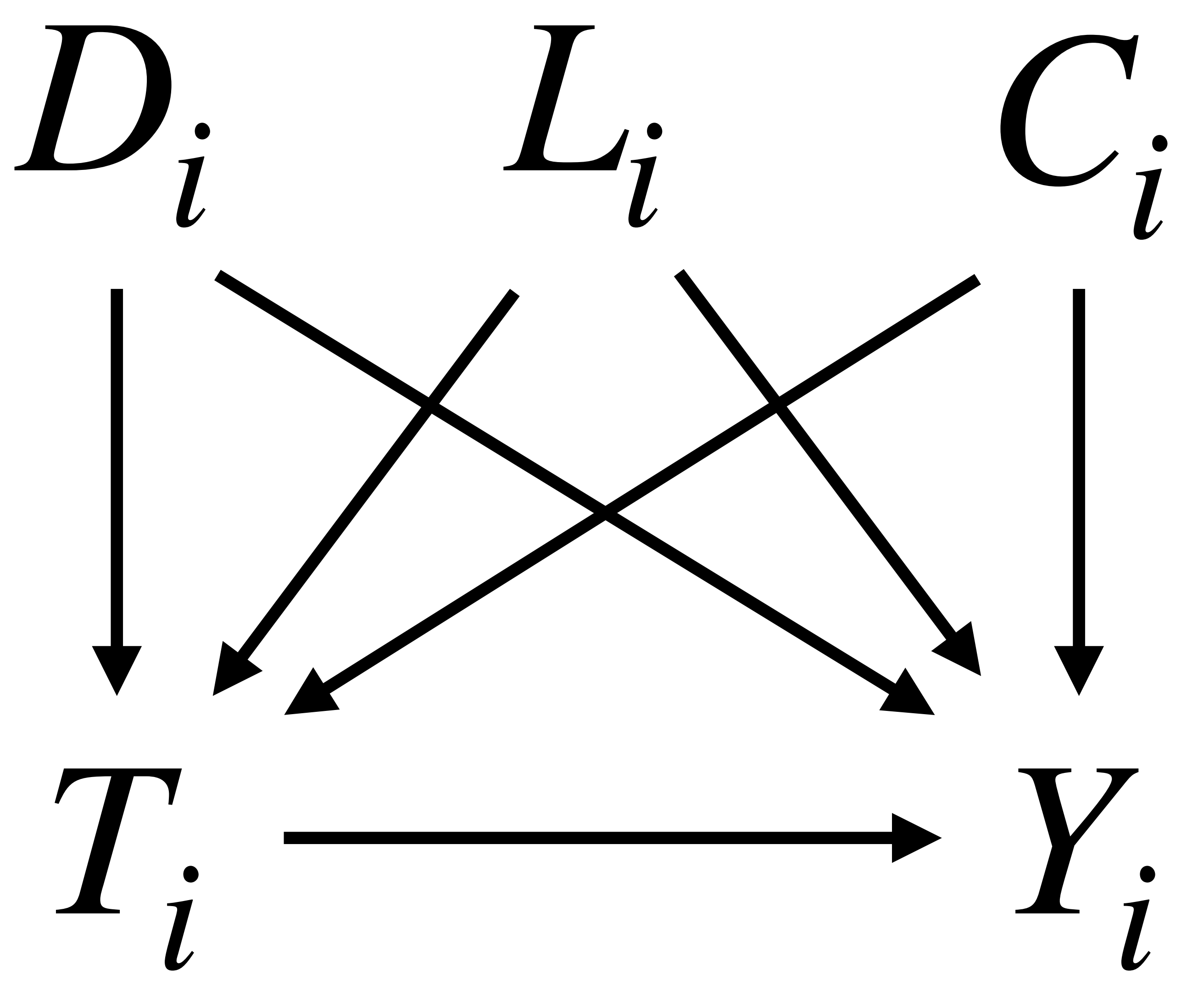
β_1 captures how math questions differ from the grand mean.

β_2 captures how biology questions differ from the grand mean.

$-(\beta_1 + \beta_2)$ captures how history questions differ from the grand mean.

D_i	I_1	I_2
math	1	0
biology	0	1
history	-1	-1

Interpreting Coefficients



$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 C_i + \epsilon_i$$

- $T_i \in \{0,1\}$ — prompt-type
- $Y_i \in [0,10]$ — Correctness score
- $C_i \in [0,10]$ — Question difficulty

$E[Y_i \mid T_i, C_i] = \beta_0 + \beta_1 T_i + \beta_2 C_i$

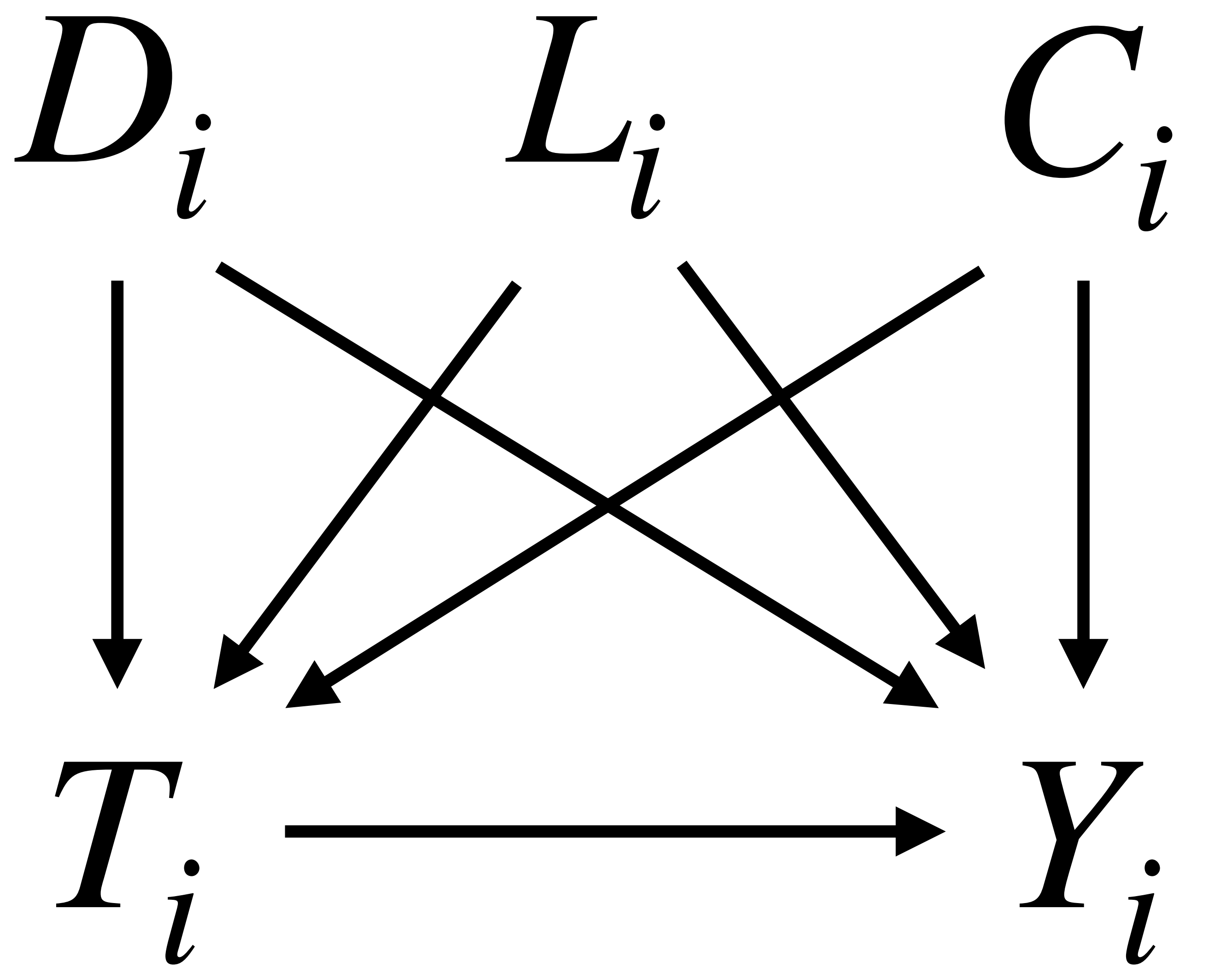
- β_0: Avg Y_i when $C_i = 0$, $T_i = 0$.
- β_1: Difference in Y_i between $T_i = 1$ and $T_i = 0$ holding C_i fixed!
- β_2: Difference in Y_i from increasing C_i by one unit holding T_i fixed!

$E[Y_i \mid T_i = 0, C_i = 0] = \beta_0$

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

$E[Y_i \mid T_i = t, C_i = c + 1] - E[Y_i \mid T_i = t, C_i = c]$

Interpreting Coefficients



$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 C_i + \epsilon_i$$

- $T_i \in \{0,1\}$ – prompt-type
- $Y_i \in \{0,1\}$ – Correctness flag
- $C_i \in [0,10]$ – Question difficulty

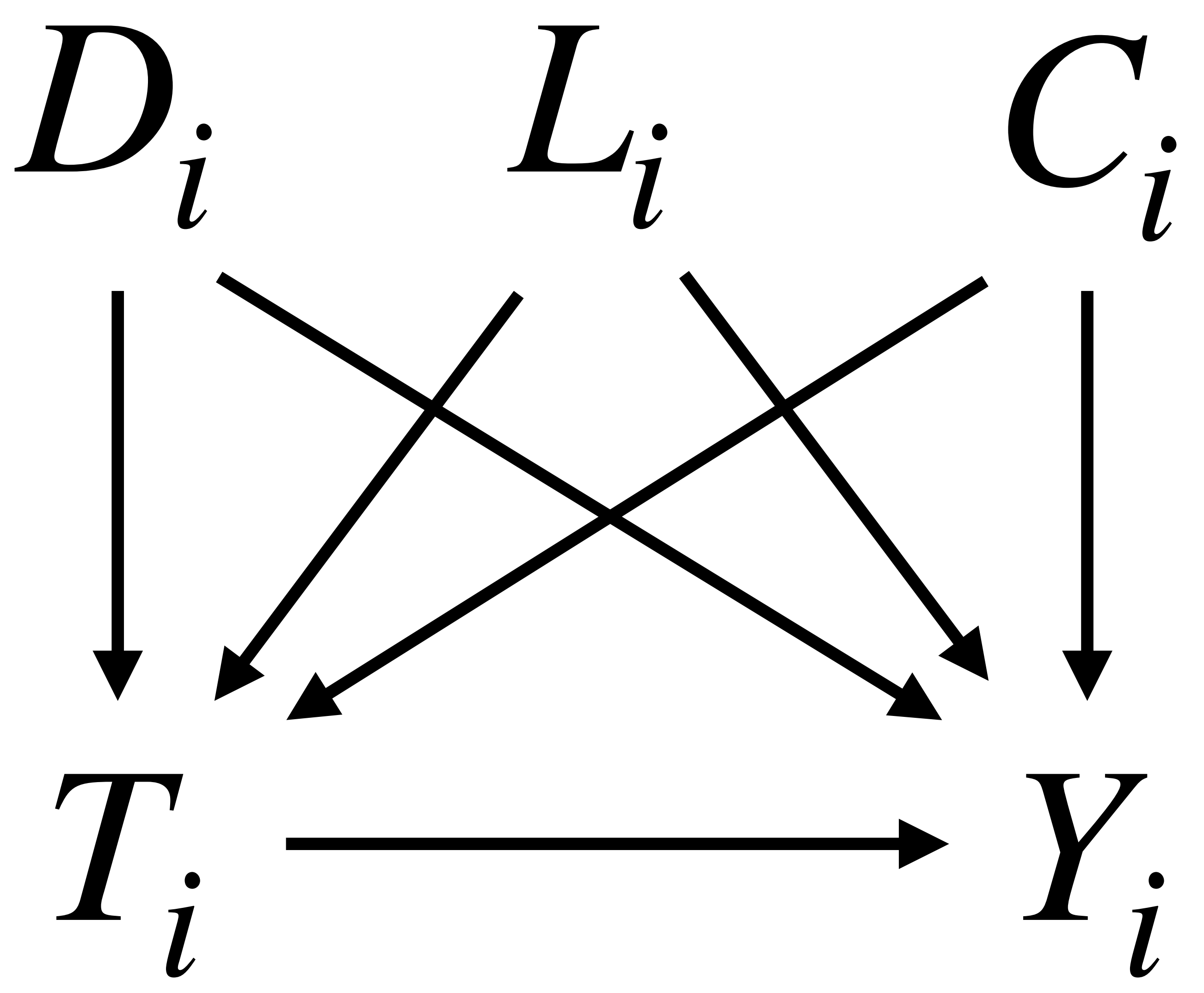
- β_0: Y_i *when* $C_i = 0, T_i = 0.$
- β_1: Difference in Y_i between $T_i = 1$ and $T_i = 0$ holding C_i fixed!
- β_2: Difference in Y_i from increasing C_i by one unit holding T_i fixed!

Interpretation now is in percentage points!

"Linear Probability Model"

$$E[Y_i \mid T_i, C_i] = \Pr(Y_i \mid T_i, C_i)$$

Interactions



$$Y_i = \beta_0 + \beta_1 C_i + \beta_2 D_i + \beta_3 C_i D_i + \epsilon_i$$

- $Y_i \in [0,10]$ — Correctness score
- $C_i \in \{\text{Low, High}\}$ — Difficulty
- $D_i \in \{\text{Math, History}\}$ — Domain

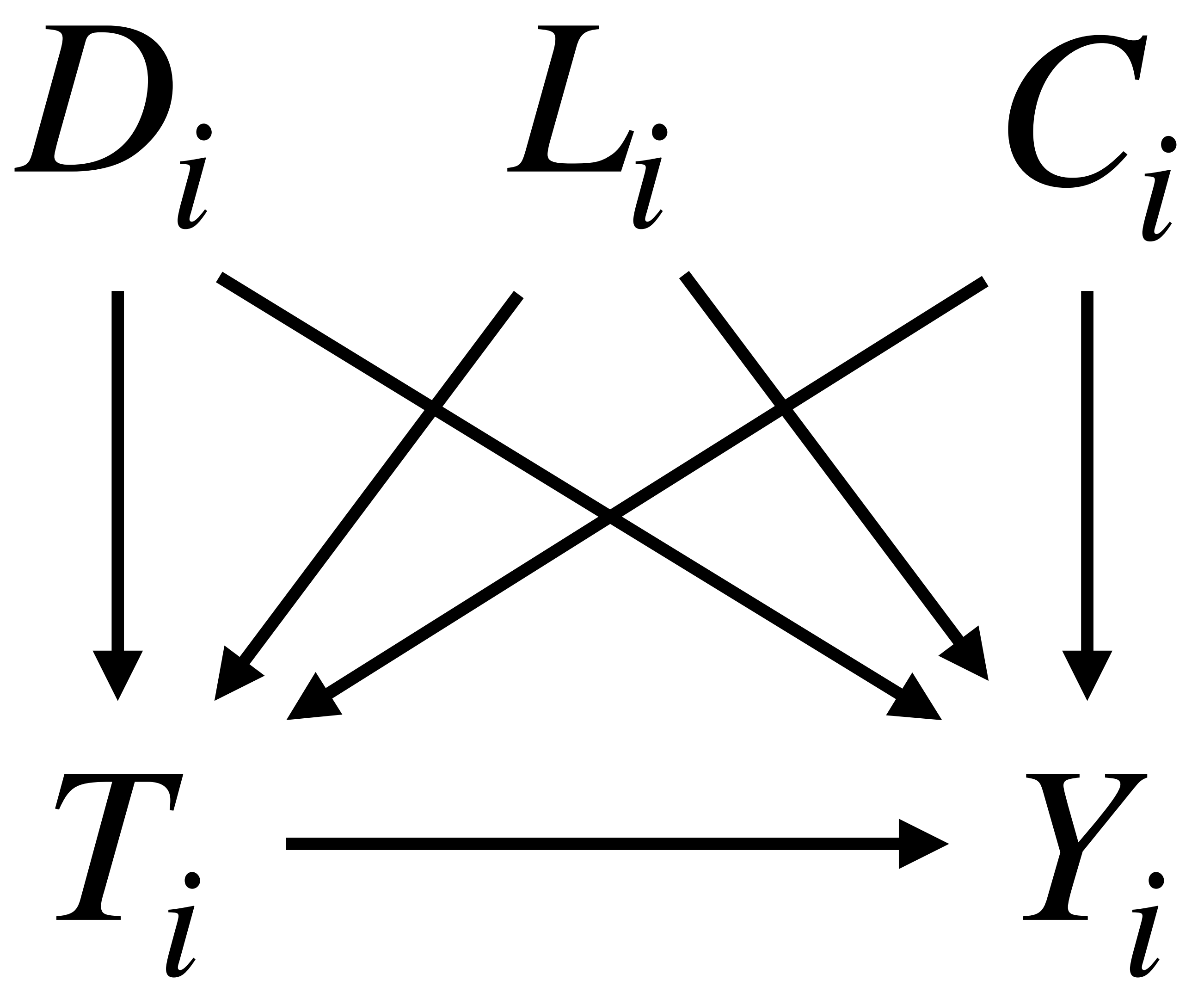
- β_0: Y_i when $C_i =$ Low, $D_i =$ Math.
- β_1: effect of high vs. low difficulty within the baseline subject (Math)
- β_2: effect of history vs math within the baseline difficulty (Low)

$E[Y_i \mid C_i = \text{Low}, D_i = \text{Math}] = \beta_0$

$E[Y_i \mid C_i = \text{High}, D_i = \text{Math}] - E[Y_i \mid C_i = \text{Low}, D_i = \text{Math}] = \beta_1$

$E[Y_i \mid C_i = \text{Low}, D_i = \text{History}] - E[Y_i \mid C_i = \text{Low}, D_i = \text{Math}] = \beta_2$

Interactions



$$Y_i = \beta_0 + \beta_1 C_i + \beta_2 D_i + \beta_3 C_i D_i + \epsilon_i$$

- $Y_i \in [0,10]$ — Correctness score
- $C_i \in \{\text{Low, High}\}$ — Difficulty
- $D_i \in \{\text{Math, History}\}$ — Domain

- β_3: is the interaction term, it captures the extra bump (positive or negative) that appears only when both indicators are 1, beyond what you'd expect from adding them separately.

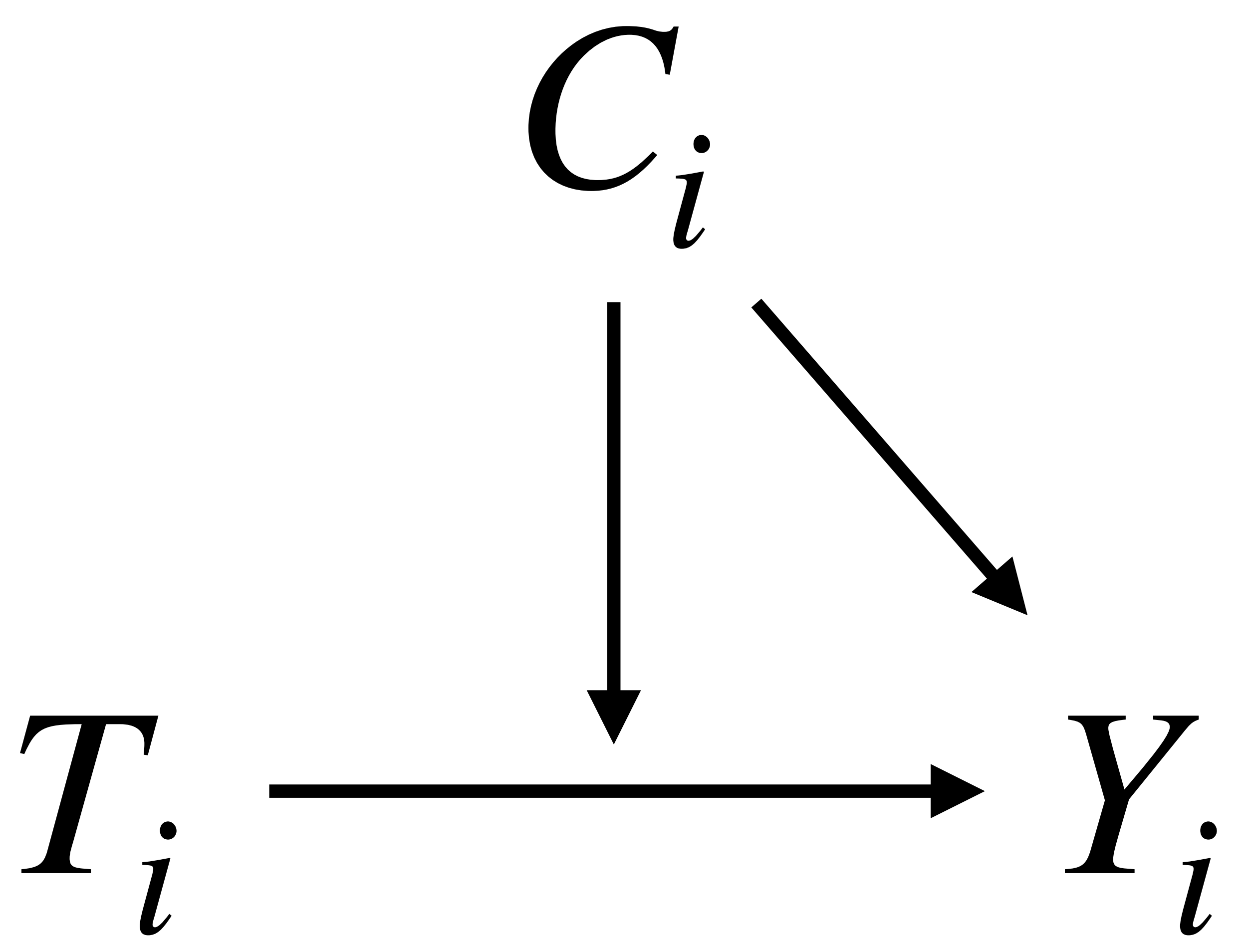
$$\beta_3 = E[Y \mid C=1, D=1] - E[Y \mid C=1, D=0] - E[Y \mid C=0, D=1] + E[Y \mid C=0, D=0]$$

$$\beta_3 = E[Y \mid C=1, D=1] - \beta_2 - \beta_1 + \beta_0$$

Effect Modification

$Y_i = \beta_0 + \beta_1 T_i + \beta_2 C_i + \beta_3 C_i T_i$

- $Y_i \in [0,10]$ — Correctness score
- $T_i \in \{0,1\}$ — Prompt-type
- $C_i \in \{\text{Low, High}\}$ — Difficulty



- β_3: captures whether the effect of the prompt type depends on the difficulty.

- C_i is an effect modifier on T_i if $\beta_3 \neq 0$.

We say that effects that differ significantly across subpopulations are *heterogeneous.*

Note*: We use the term* *effect interaction* *to refer to the interaction of two treatments.*