08 - Causal Inference

ml4econ, HUJI 2024

Itamar Caspi
July 7, 2024 (updated: 2024-07-07)

Reproducing This Presentation

Use the pacman package to install and load necessary packages.

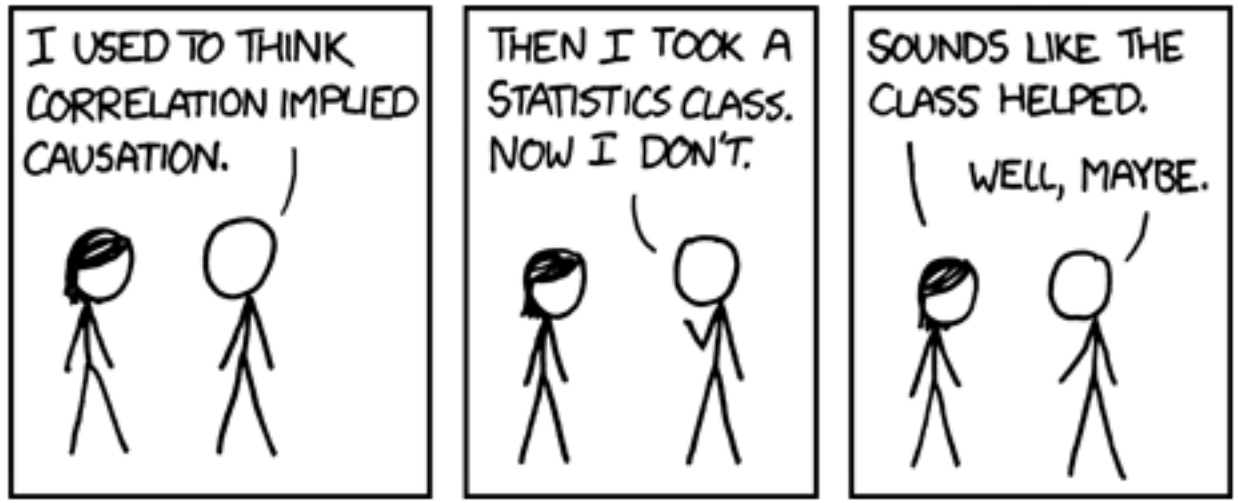
```
if (!require("pacman"))
  install.packages("pacman")

pacman::p_load(
  tidyverse,   # for data wrangling and visualization
  tidymodels,  # for modeling
  haven,       # for reading dta files
  here,        # for referencing folders
  dagitty,     # for generating DAGs
  ggdag,       # for drawing DAGs
  knitr        # for printing html tables
)
```

Outline

- Causal Inference
- Potential Outcomes
- Directed Acyiclic Graphs
- Simulations

Causal Inference

Predicting vs. Explaining



Source: XKCD

Shifting Gears: From Prediction to Causal Inference

- So far, our focus has centered on prediction.
- But as economists, we're primarily interested in *causal inference*, such as:
  - How does class size impact student performance?
  - How does education influence earnings?
  - What is the effect of government spending on GDP?
  - And so on.
- Before diving into how to modify and apply ML methods to causal inference issues, we need to clarify what we mean by causal inference.
- This lecture will cover two prevalent approaches to causal inference: the statistical/econometric approach and the computer science approach.

Pearl and Rubin

Source: The Book of Why (Pearl and Mackenzie)

Spotlight on Identification

- This lecture mainly zeroes in on *identification*, rather than prediction, estimation, or inference.
- To put it briefly, identification refers to:

> *"Model parameters or features being uniquely determined from the observable population that generates the data."* - (Lewbel, 2019)

- To be more specific, consider identifying the parameter of interest when you have access to unlimited data (the entire population).

Potential Outcomes

The Road Not Taken (Counterfactuals)

Source: https://mru.org/courses/mastering-econometrics/ceteris-paribus

Understanding Notation

- Y represents a random variable.
- X signifies a vector of attributes.
- $\mathbf{X}$ stands for a design matrix.

Treatment and potential outcomes (Rubin, 1974, 1977)

- Treatment

$$D_i = \begin{cases} 1, & \text{if unit } i \text{ received the treatment} \\ 0, & \text{otherwise.} \end{cases}$$

- Treatment and potential outcomes

$$\begin{aligned} Y_{i0} &\quad \text{is the potential outcome for unit } i \text{ with } D_i = 0 \\ Y_{i1} &\quad \text{is the potential outcome for unit } i \text{ with } D_i = 1 \end{aligned}$$

- Observed Outcome: Under the Stable Unit Treatment Value Assumption (SUTVA), the realization of unit i's outcome is

$$Y_i = Y_{1i} D_i + Y_{0i}(1 - D_i)$$

Fundamental problem of causal inference (Holland, 1986): We cannot observe *both* Y_{1i} and Y_{0i}.

Exploring Treatment Effect and Observed Outcomes

- Individual Treatment Effect: This is the difference between unit i's potential outcomes:

$$\tau_i = Y_{1i} - Y_{0i}$$

- *Average treatment effect* (ATE)

$$\mathbb{E}[\tau_i] = \mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$$

- *Average treatment effect for the treatment group* (ATT)

$$\mathbb{E}[\tau_i | D_i = 1] = \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1] = \mathbb{E}[Y_{1i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 1]$$

NOTE: The complement of the treatment group forms the *control* group.

Guarding Against Selection Bias

A straightforward estimand for ATE is the difference between average outcomes based on treatment status.

However, tread carefully as this approach might lead you astray:

$$\mathbb{E}\left[Y_i|D_i=1\right]-\mathbb{E}\left[Y_i|D_i=0\right]=\underbrace{\mathbb{E}\left[Y_{1i}|D_i=1\right]-\mathbb{E}\left[Y_{0i}|D_i=1\right]}_{\text{ATT}}+\underbrace{\mathbb{E}\left[Y_{0i}|D_i=1\right]-\mathbb{E}\left[Y_{0i}|D_i=0\right]}_{\text{selection bias}}$$

> **Most of causal inference is about eradicating selection bias.**

EXAMPLE: Individuals choosing private universities likely possess distinct characteristics compared to those opting for public universities.

How Randomized Control Trials (RCTs) Solve Selection Bias

In an RCT, treatments get assigned randomly. Consequently, D_i is *independent* of potential outcomes, namely:

$$\{Y_{1i}, Y_{0i}\} \perp D_i$$

RCTs allow us to estimate ATE using the average difference in outcomes by treatment status:

$$\begin{aligned}
\mathbb{E}\left[Y_i|D_i=1\right] - \mathbb{E}\left[Y_i|D_i=0\right] &= \mathbb{E}\left[Y_{1i}|D_i=1\right] - \mathbb{E}\left[Y_{0i}|D_i=0\right] \\
&= \mathbb{E}\left[Y_{1i}|D_i=1\right] - \mathbb{E}\left[Y_{0i}|D_i=1\right] \\
&= \mathbb{E}\left[Y_{1i} - Y_{0i}|D_i=1\right] \\
&= \mathbb{E}\left[Y_{1i} - Y_{0i}\right] \\
&= \mathbf{ATE}
\end{aligned}$$

EXAMPLE: In theory, randomly assigning students to private and public universities would enable us to estimate the ATE of attending a private school on future earnings. Clearly, executing an RCT in this situation is unfeasible.

Interpreting Estimands and Regression

Let's make an assumption that the treatment effect is constant across all individuals, i.e.,

$$\tau = Y_{1i} - Y_{0i}, \quad \forall i.$$

Given this, we can formulate Y_i as follows:

$$\begin{aligned}
Y_i &= Y_{1i}D_i + Y_{0i}(1 - D_i) \\
&= Y_{0i} + D_i(Y_{1i} - Y_{0i}), \\
&= Y_{0i} + \tau D_i, && \text{since } \tau = Y_{1i} - Y_{0i} \\
&= \mathbb{E}[Y_{0i}] + \tau D_i + Y_{0i} - \mathbb{E}[Y_{0i}], && \text{add and subtract } \mathbb{E}[Y_{0i}]
\end{aligned}$$

Or, more simply:

$$Y_i = \alpha + \tau D_i + u_i,$$

where $\alpha = \mathbb{E}[Y_{0i}]$ and $u_i = Y_{0i} - \mathbb{E}[Y_{0i}]$ denotes the random component of Y_{0i}.

The Role of Unconfoundedness

In most observational studies, treatments aren't randomly assigned (Consider $D_i = \{\mathbf{private}, \mathbf{public}\}$).

In such situations, identifying causal effects relies on the *Unconfoundedness* assumption, also known as "selection-on-observable". This is defined as:

$$\{Y_{1i}, Y_{0i}\} \perp D_i | X_i$$

This means treatment assignment is independent of potential outcomes *conditional* on observable X_i. In other words, selection bias *vanishes* when we control for X_i.

Adjusting for Confounding Factors

We typically control for X_i by incorporating it into the regression:

$$Y_i = \alpha + \tau D_i + X_i'\boldsymbol{\beta} + u_i.$$

Comments:

1. Strictly speaking, the regression model above is valid only if we truly *believe* that the "real" model is $Y_i = \alpha + \tau D_i + X_i'\boldsymbol{\beta} + u_i$.

2. If D_i is randomly assigned, including X_i in the regression **might** increase the accuracy of ATE.

3. If D_i is assigned based on X_i (as often happens in observational settings), including X_i in the regression eliminates selection bias.

Illustration: The OHIE Data

- The Oregon Health Insurance Experiment (OHIE) is a randomized controlled trial designed to measure the treatment effect of Medicaid eligibility.
- Treatment group: Individuals selected in the Medicaid lottery.
- The outcome, `doc_any_12m`, equals 1 for patients who saw a primary care physician, and zero otherwise.

Load the OHIE Data

In this illustration, we will merge three separate (Stata) files and import them into R using the {haven} package:

```
descr <-
  here("08-causal-inference/data",
       "oregonhie_descriptive_vars.dta") %>%
  read_dta()

prgm <-
  here("08-causal-inference/data",
       "oregonhie_stateprograms_vars.dta") %>%
  read_dta()

s12 <-
  here("08-causal-inference/data",
       "oregonhie_survey12m_vars.dta") %>%
  read_dta()
```

The complete OHIE dataset can be accessed here.

Preprocessing: Joining Datasets

To join the three data frames and remove empty responses, you can follow these steps:

```
ohie_raw <-
  descr %>%
  left_join(prgm) %>%
  left_join(s12) %>%
  filter(sample_12m_resp == 1) %>%
  drop_na(doc_any_12m)
```

Preprocessing: Refinement

To refine the data, follow these steps:

1. Select the relevant variables that are of interest.
2. Re-level the `numhh_list` variable (household size) according to your specific needs.

```
ohie <-
  ohie_raw %>%
  dplyr::select(numhh_list, treatment, doc_any_12m) %>%
  mutate(
    numhh_list = factor(numhh_list, levels = c("1", "2", "3"))
  )
```

The Final Dataset

```
ohie
```

```
## # A tibble: 23,492 x 3
##    numhh_list treatment        doc_any_12m
##    <fct>      <dbl+lbl>        <dbl+lbl>
##  1 1          1 [Selected]     0 [No]
##  2 1          1 [Selected]     0 [No]
##  3 1          1 [Selected]     0 [No]
##  4 1          1 [Selected]     1 [Yes]
##  5 2          0 [Not selected] 0 [No]
##  6 1          0 [Not selected] 1 [Yes]
##  7 2          0 [Not selected] 1 [Yes]
##  8 1          1 [Selected]     1 [Yes]
##  9 1          1 [Selected]     0 [No]
## 10 2          1 [Selected]     1 [Yes]
## # i 23,482 more rows
```

Distribution of Treated vs. Control

```
ohie %>%
  count(treatment) %>%
  kable(format = "html")
```

treatment	n
0	11811
1	11681

Estimating ATE

To estimate the Average Treatment Effect (ATE), you can use the following model:

$$doc\_any\_12m_i = \alpha + \tau \times selected_i + \varepsilon_i$$

In R:

```
fit <- lm(doc_any_12m ~ treatment, data = ohie)
```

Results

```
fit %>%
  tidy(conf.int = TRUE) %>%
  filter(term != "(Intercept)") %>%
  dplyr::select(term, estimate, starts_with("conf.")) %>%
  kable(digits = 4, format = "html")
```

term	estimate	conf.low	conf.high
treatment	0.0572	0.0447	0.0697

Interpretation: Being selected in the lottery increases the probability of visiting a primary care physician in the following year by 5.72 [4.47, 6.79] percentage points.

Adjustments

One concern with the OHIE dataset is that individuals can apply for Medicaid for their entire household.

This fact undermines the crucial random assignment assumption, as belonging to larger households increases the chances of being selected for Medicaid.

```
ohie %>%
  count(treatment, numhh_list) %>%
  kable(format = "html")
```

treatment	numhh_list	n
0	1	8824
0	2	2981
0	3	6
1	1	7679
1	2	3950
1	3	52

ATE Under Adjustment for numhh

The model with adjustment:

$$doc\_any\_12m_i = \alpha + \tau \times selected_i + \beta \times numhh_i + \varepsilon_i$$

Estimation:

```
fit_adj <- lm(doc_any_12m ~ treatment + numhh_list, data = ohie)
```

Results

```
fit_adj %>%
  tidy(conf.int = TRUE) %>%
  dplyr::select(term, estimate, starts_with("conf.")) %>%
  kable(digits = 4, format = "html")
```

term	estimate	conf.low	conf.high
(Intercept)	0.5925	0.5831	0.6020
treatment	0.0635	0.0510	0.0760
numhh_list2	-0.0654	-0.0792	-0.0517
numhh_list3	-0.1839	-0.3097	-0.0582

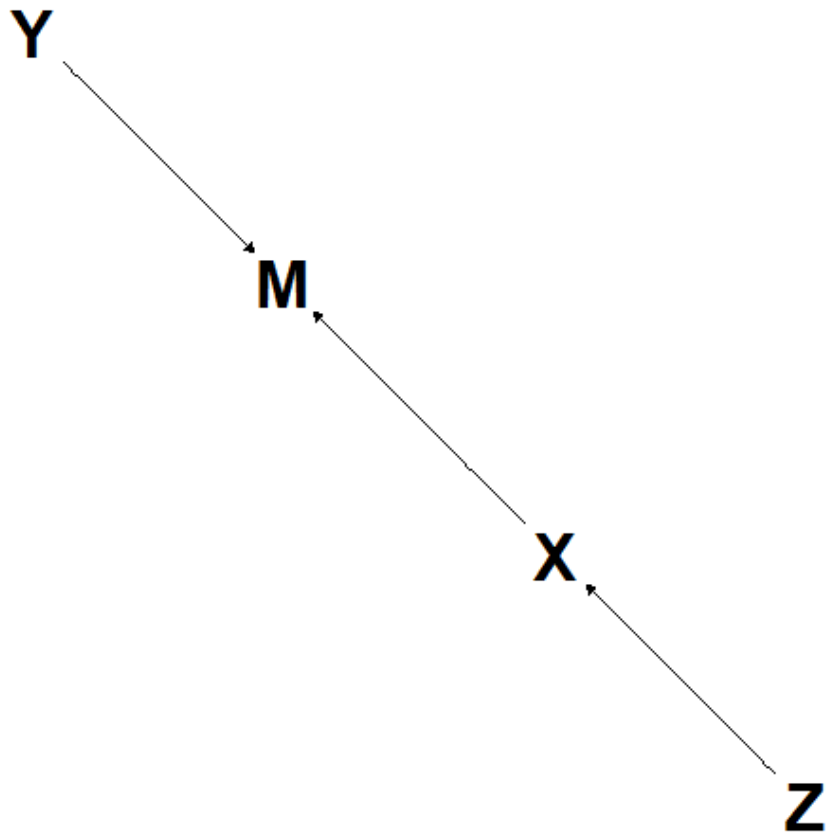
After adjusting for `numhh` (household size), the Average Treatment Effect (ATE) has increased from 5.72 to 6.35 percentage points (why?).

Directed Acyclic Graphs

Understanding DAGs

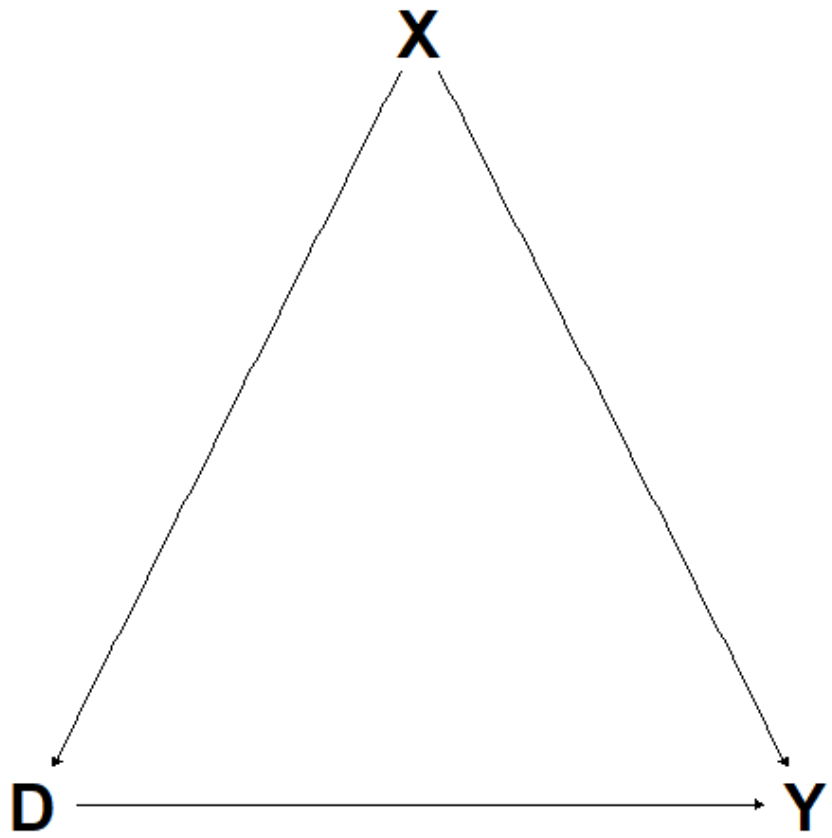
A DAG (Directed Acyclic Graph) is a graphical representation used to model a system of causal interactions.

- **Nodes** represent random variables, such as X, Y, etc.
- **Arrows** (or directed edges) indicate causal effects. For example, $Z \rightarrow X$ denotes that "Z causes X".
- A **path** is a sequence of edges connecting two nodes. For instance, $Z \rightarrow X \rightarrow M \leftarrow Y$ describes a path from Z to Y.
- In a **direct path**, arrows point in the same direction, like $Z \rightarrow X \rightarrow M$.



Confounder DAG

- X acts as a common cause of both D and Y.
- When we condition on X, we eliminate the dependency between D and Y through X.
- In terms of DAGs, controlling for X "closes the backdoor path" between D and Y while leaving the direct path open.
- The concept of closing the backdoor path is connected to the notion of omitted variable bias.



DAGs and SEM (Structural Equation Models)

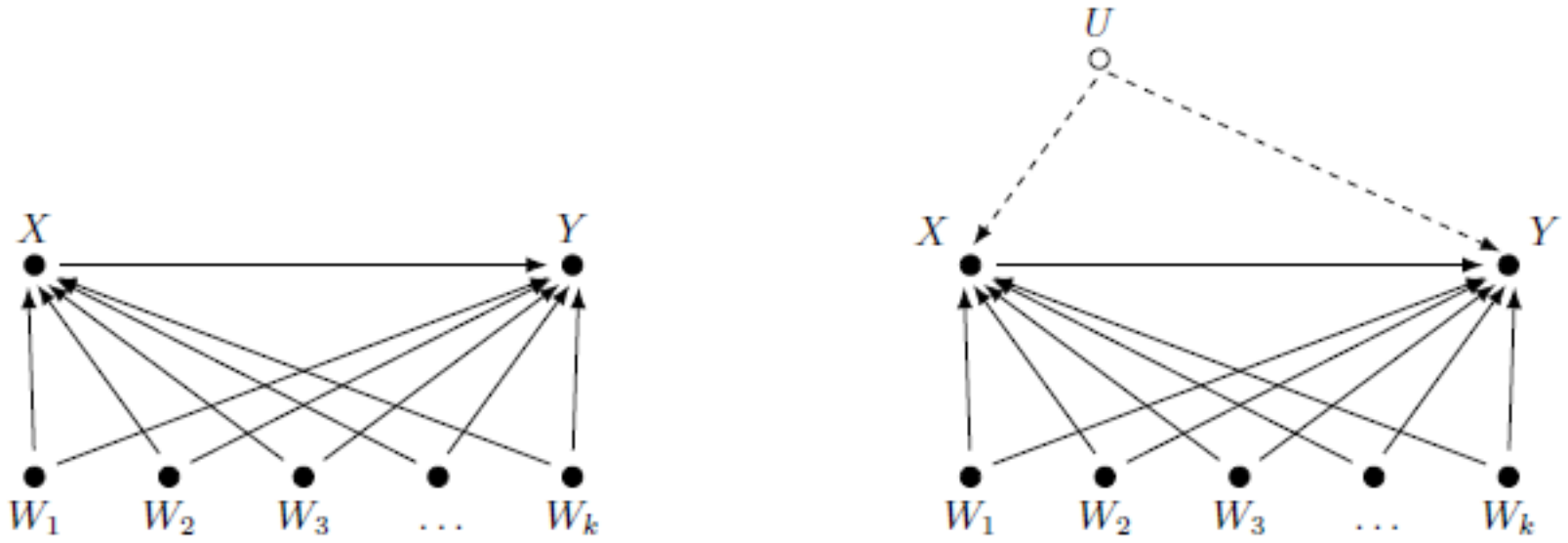
- Another way to conceptualize DAGs is as non-parametric **Structural Equation Models** (SEM).
- For instance, the single-confounder DAG we just explored can be represented by a set of three equations:

$$
\begin{aligned}
X &\leftarrow f_X(u_X) \\
D &\leftarrow f_D(X, u_D) \\
Y &\leftarrow f_Y(D, X, u_Y)
\end{aligned}
$$

where:

- The f_i functions represent the causal mechanisms in the model and are not restricted to being linear.
- u_X, u_D, and u_Y denote independent background factors that we choose not to include in the analysis.
- The assignment operator ($\leftarrow$) captures the asymmetry of causal relationships.

Unconfoundedness in DAGs

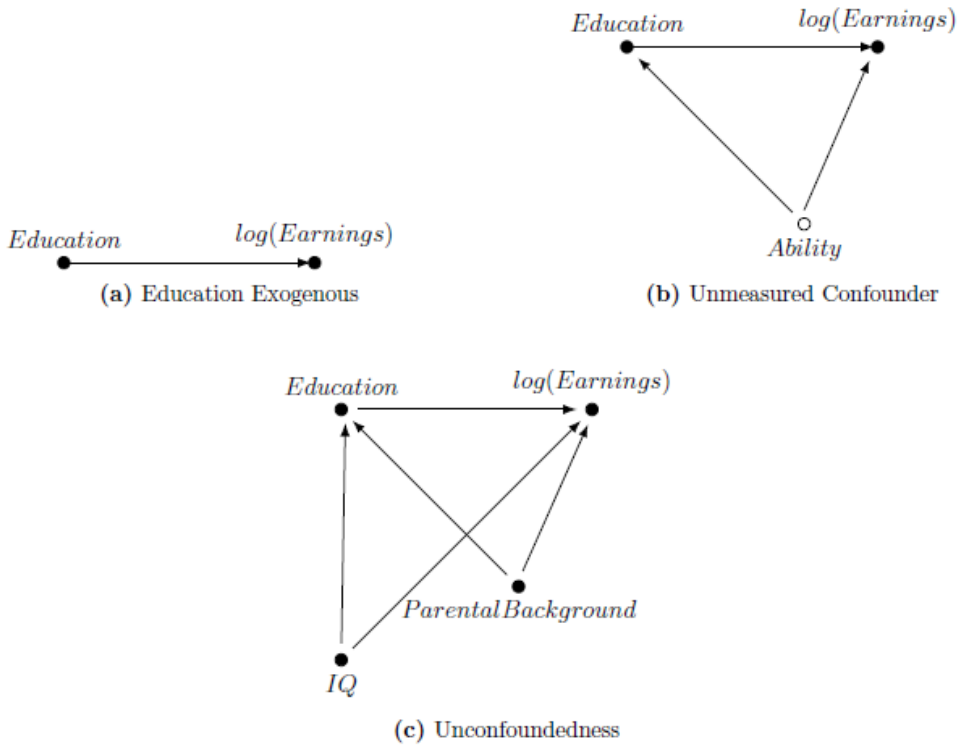


(a) Confoundedness **(b)** Violation of Unconfoundedness

Figure 8: Unconfoundedness with Multiple Observed Confounders

Source: Imbens (2019).

Example: Identifying the Returns to Education



(a) Education Exogenous

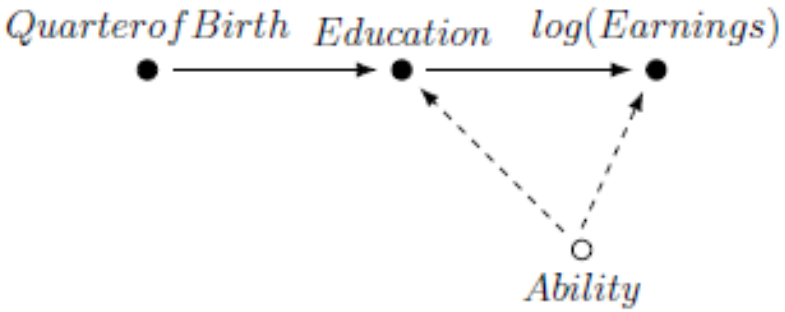
(b) Unmeasured Confounder

(c) Unconfoundedness

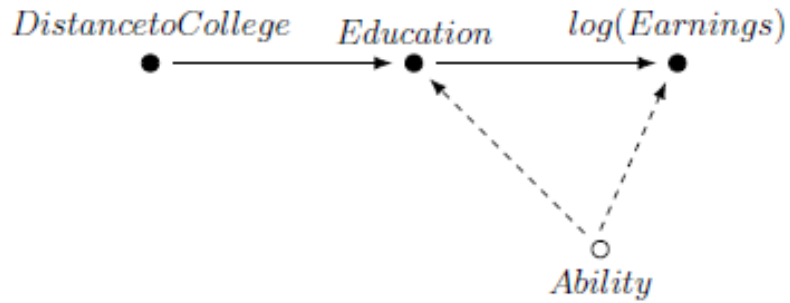
Figure 15: DAGs for the Returns to Education (I)

Source: Imbens (2019).

Instrumental variables in DAGs



(a) Instrumental Variables: Quarter of Birth



(b) Instrumental Variables: Distance to College

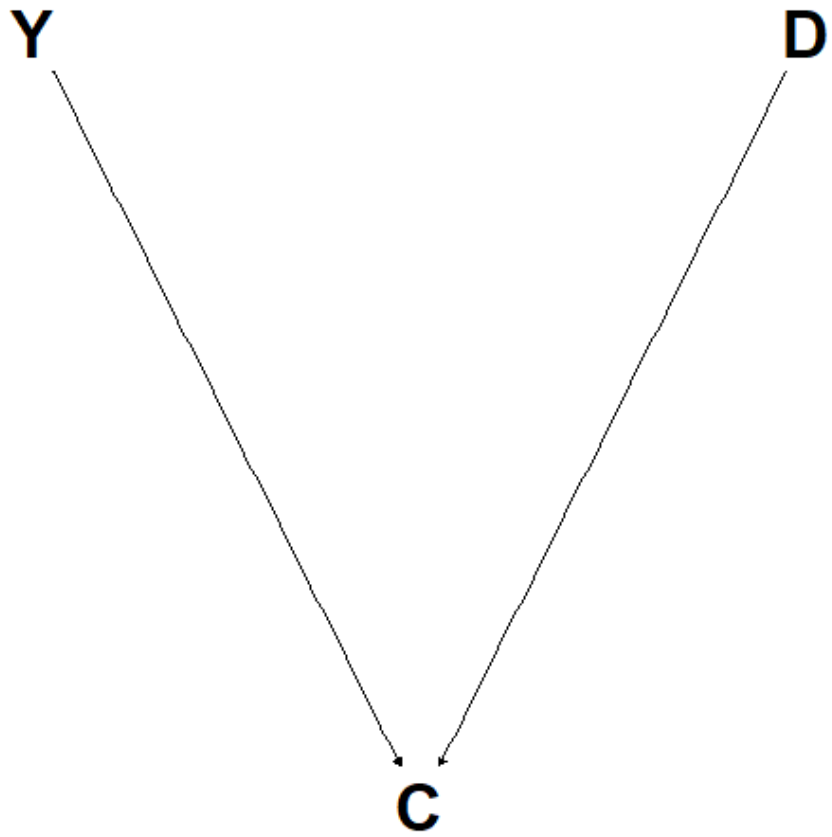
Source: Imbens (2019).

Mediator DAG

- D influences M which, in turn, influences Y.
- M acts as a mediator of the causal effect between D and Y.
- Conditioning on M eliminates the dependency between D and Y.
- Essentially, we have closed the direct path, which is the only direct path between D and Y.

A Collider

- D are Y are independent.
- D and Y jointly cause C.
- conditioning on C creates dependency between D and Y

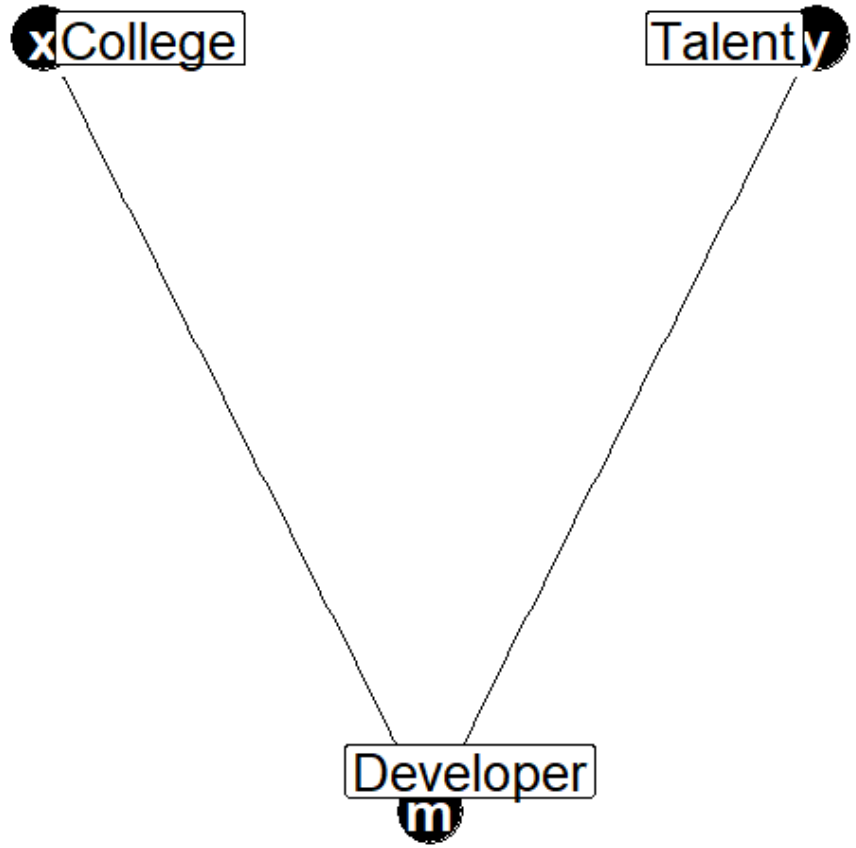


Example: "Bad Controls"

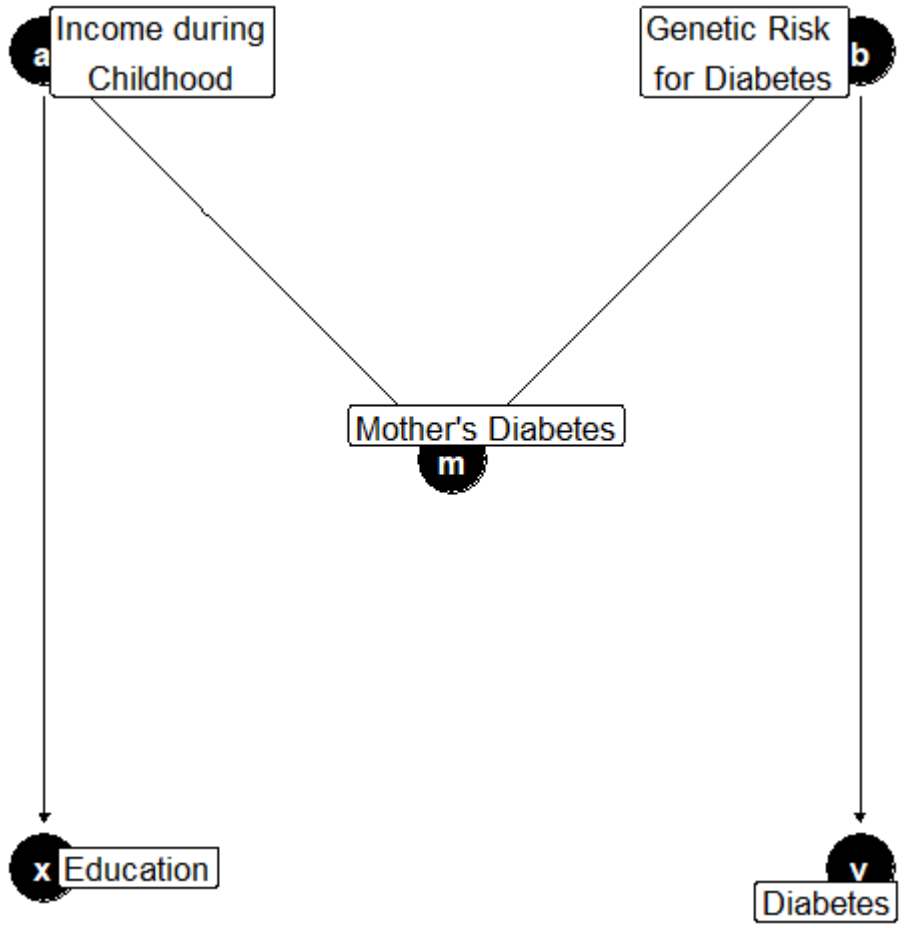
- "Bad controls" refer to variables that are outcome variables themselves.
- This distinction becomes important, particularly when working with high-dimensional data.

EXAMPLE: Using occupation as a control in a regression estimating the return to years of schooling.

Considering that an individual works as a developer in a high-tech firm can change the interpretation of the results. Knowing that the person does not have a college degree immediately indicates their likely exceptional talent.



Collider: M-bias



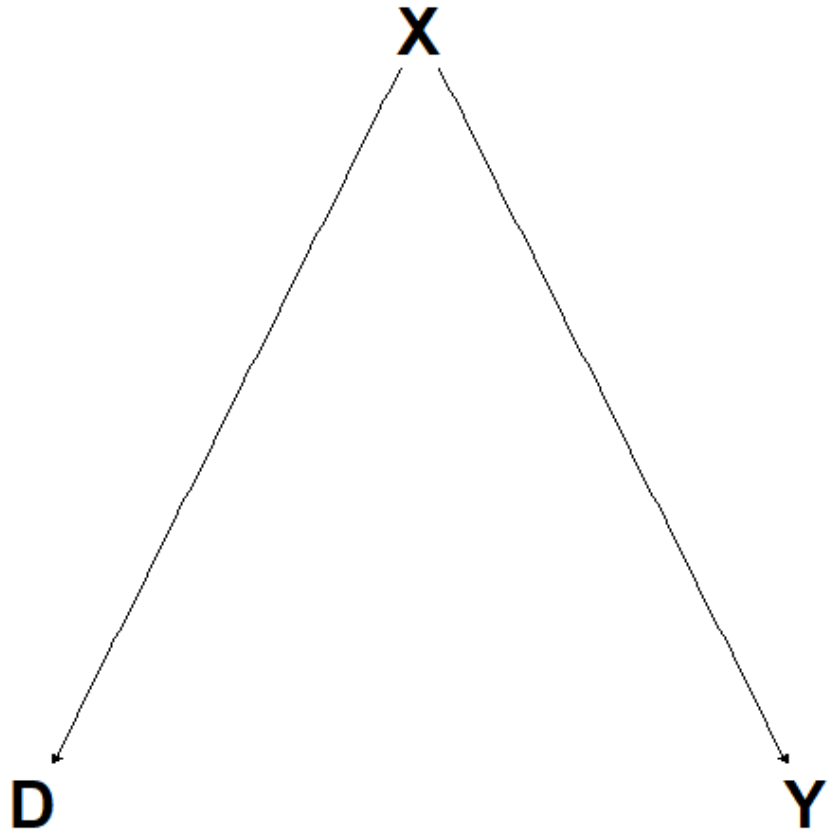
Simulations

Simulation I: De-counfounding

Simulate the DGP:

```
n <- 1000
p <- 3

u <- matrix(rnorm(n * p), n, p)

x <- u[,2]
d <- 0.8 * x + 0.6 * u[,1]
y <- 0 * d + 0.2 * x + u[,3]
```

Note that the "true" effect $D \rightarrow Y$ is zero (i.e., $ATE = 0$).

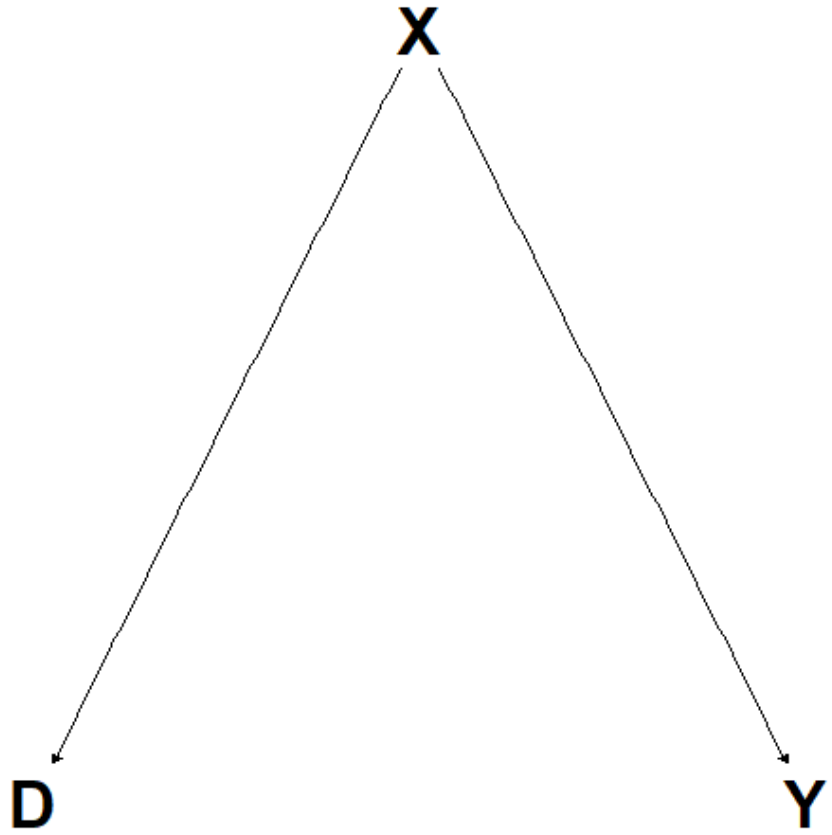


Simulation I: De-confounding (Cont.)

Raw correlation matrix:

	y	x	d
y	1.0	0.2	0.2
x	0.2	1.0	0.8
d	0.2	0.8	1.0

Note: It is important to highlight that Y and D are correlated even though there is no direct arrow between them in the DAG. This correlation arises due to the presence of the confounder X, which opens a backdoor path between Y and D.



Simulation I: De-confounding (cont.)

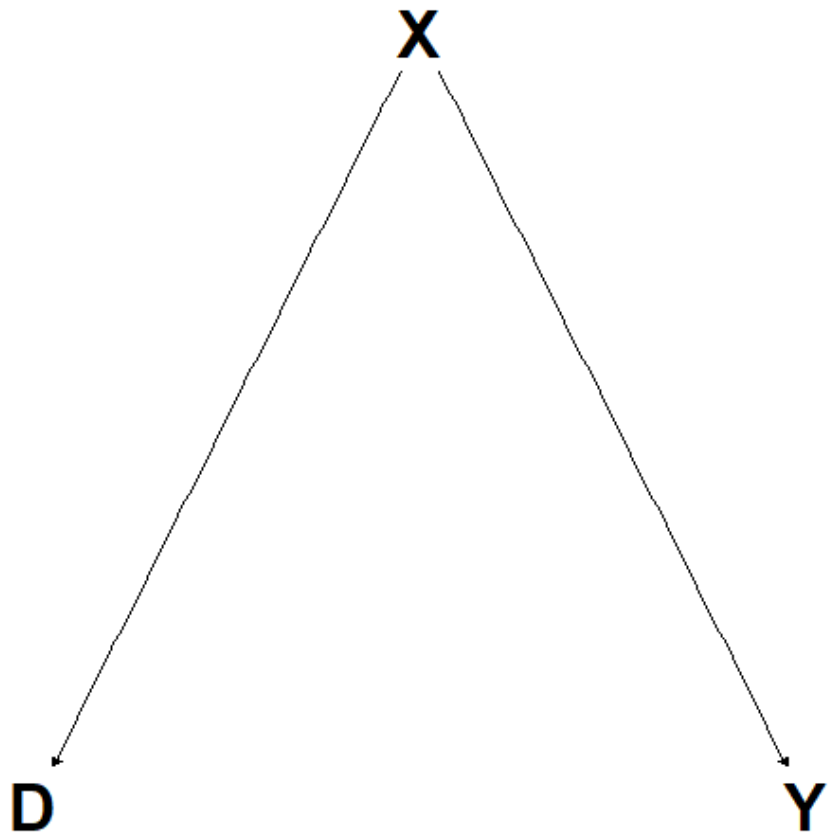
Now, let's estimate the model with X included on the right-hand side:

term	estimate	p.value
d	0.08	0.14
x	0.14	0.01

and without X

term	estimate	p.value
d	0.2	0

BOTTOM LINE: Controlling for X provides the correct answer.



Simulation II: Mediator

The DGP:

```
n <- 1000
p <- 3

u <- matrix(rnorm(n * p), n, p)

d <- u[,1]
m <- 1.3 * d + u[,2]
y <- 0.1 * m + u[,3]
```

True effect of $D \rightarrow Y$ is $1.3 \times 0.1 = 0.13$.

Simulation II: Mediator (cont.)

Raw correlation matrix:

	y	m	d
y	1.0	0.1	0.1
m	0.1	1.0	0.8
d	0.1	0.8	1.0

In this case, both the mediator M and the treatment D are correlated with the outcome Y.

Simulation II: Mediator (Cont.)

Estimate the model with M:

term	estimate	p.value
d	0.00	0.94
m	0.08	0.01

and without M:

term	estimate	p.value
d	0.1	0

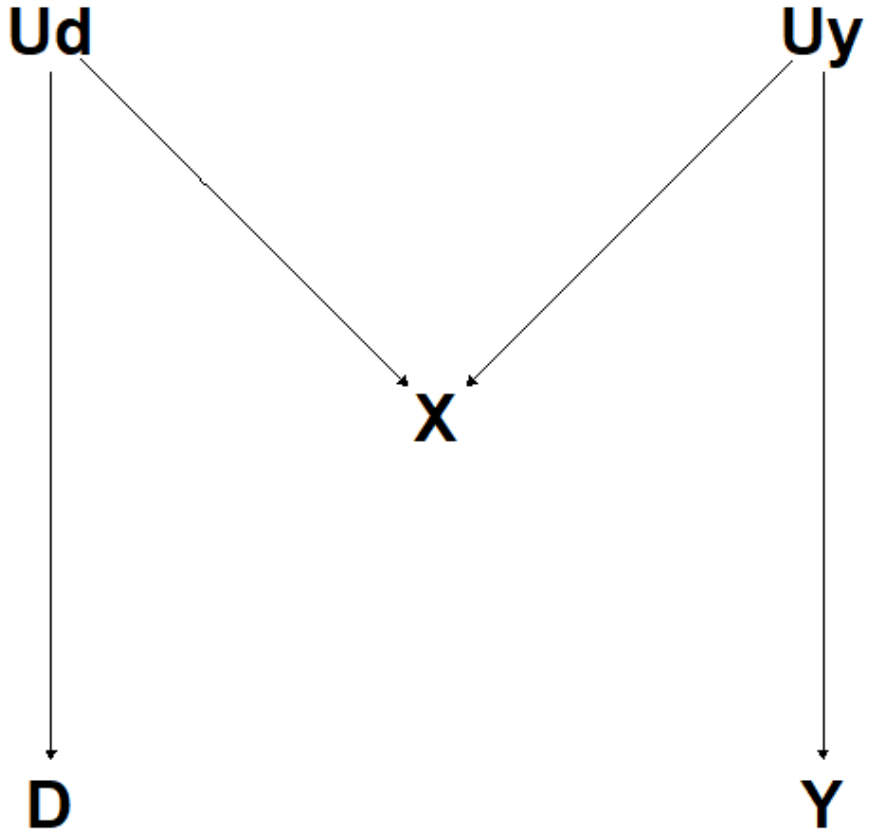
BOTTOM LINE: Controlling for M in this case biases the total effect of D on Y downward since it blocks the path from D to Y.

Simulation III: M-bias

Generate the data:

```
n <- 1000
p <- 3

u <- matrix(rnorm(n * p), n, p)

d <- u[,1]
x <- 0.8 * u[,1] + 0.2 * u[,2] + 0.6 *
y <- 0 * d + u[,2]
```

Note that X is a collider, and that the "true" effect $D \rightarrow Y$ is zero (i.e., $ATE = 0$).

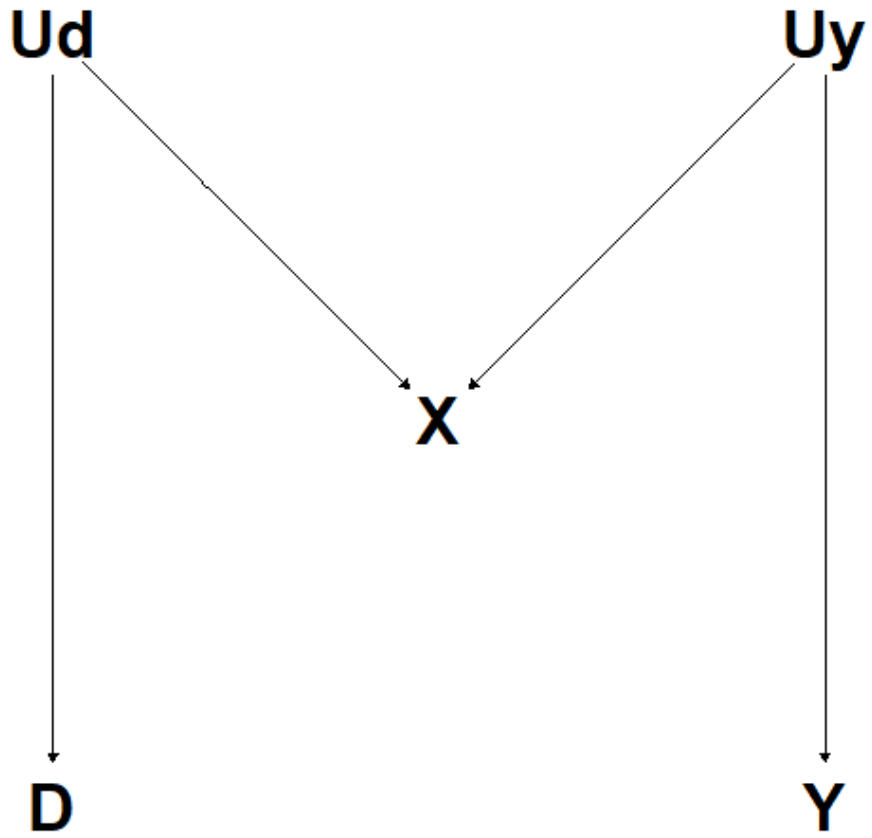


Simulation III: M-bias (cont.)

Raw correlation matrix:

	y	x	d
y	1.0	0.2	0.0
x	0.2	1.0	0.8
d	0.0	0.8	1.0

Notice how Y is uncorrelated with D and X is correlated with both D and Y.



Simulation III: M-bias

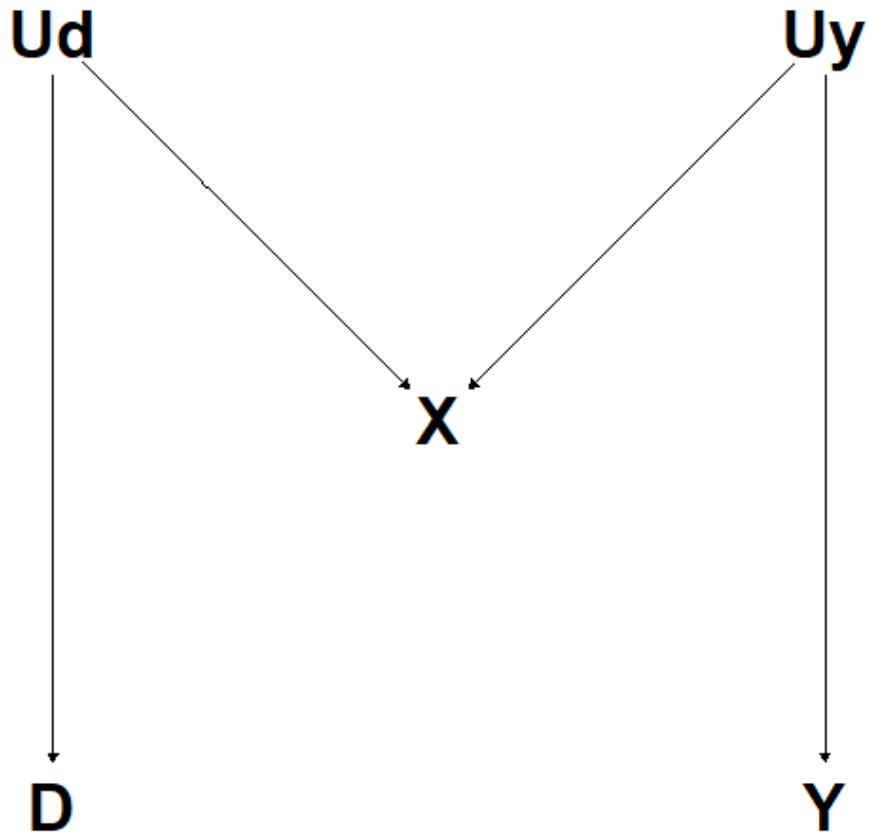
Estimate the model with X

term	estimate	p.value
d	-0.46	0
x	0.54	0

and without X

term	estimate	p.value
d	-0.03	0.33

BOTTOM LINE: Controlling for X in this case results in finding a spurious effect of D on Y since it opens a backdoor path between D to Y.



Limitations of DAGs

- It can be challenging to construct a DAG for complex (econometric) structural models.
- The need to specify the entire Data Generating Process (DGP) raises questions about whether this is truly a limitation.
- Simultaneity: *"In fact it is not immediately obvious to me how one would capture supply and demand models in a DAG"* Imbens, GW. (2020, JEL)

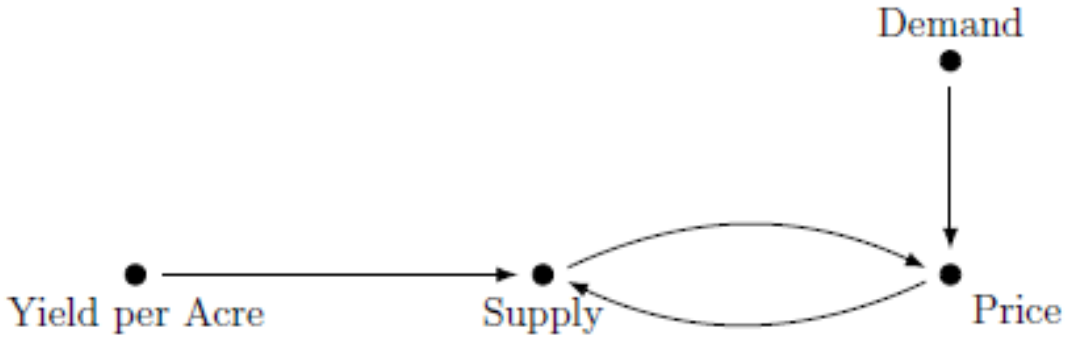


Figure 11: Based on Figure 7.10 in TBOW, p. 251.

Recommended introductory level resources on DAGs

- The Book of Why by Pearl and Mackenzie.
- Causal Inference in Machine Learning and Al by Paul Hünermund.
- Causal Inference: The Mixtape (pp. 67-80) by Scott Cunningham.
- Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics by Guido W. Imbens
- A Crash Course in Good and Bad Controls by Cinelli, Forney, and Pearl, J. (2020).

Next Time: Causal Inference in High-Dimensional Settings

We'll revisit the standard "treatment effect regression" equation:

$$Y_i = \alpha + \underbrace{\tau D_i}_{\text{low dimensional}} + \underbrace{\sum_{j=1}^{k} \beta_j X_{ij}}_{\text{high dimensional}} + \varepsilon_i, \quad \text{for } i = 1, \ldots, n$$

Our focus is on estimating $\hat{\tau}$, the estimated *average treatment effect* (ATE).

In high-dimensional settings, we encounter the scenario where $k \gg n$ (i.e., the number of covariates is much larger than the number of observations).

```
slides %>% end()
```

Source code

Selected References

- Hünermund, P., & Bareinboim, E. (2019). Causal Inference and Data-Fusion in Econometrics. arXiv preprint arXiv:1912.09104.
- Imbens, W. G. (2020). Potential Outcome and Directed Acyclic Graph Approaches to Causality: Relevance for Empirical Practice in Economics. *Journal of Economic Literature*.
- Lewbel, A. (2019). The identification zoo: Meanings of identification in econometrics. *Journal of Economic Literature*, 57(4), 835-903.