08 - Causal Inference

ml4econ, HUJI 2020

Itamar Caspi

May 18, 2019 (updated: 2020-05-18)

Replicating this presentation

Use the pacman package to install and load packages:

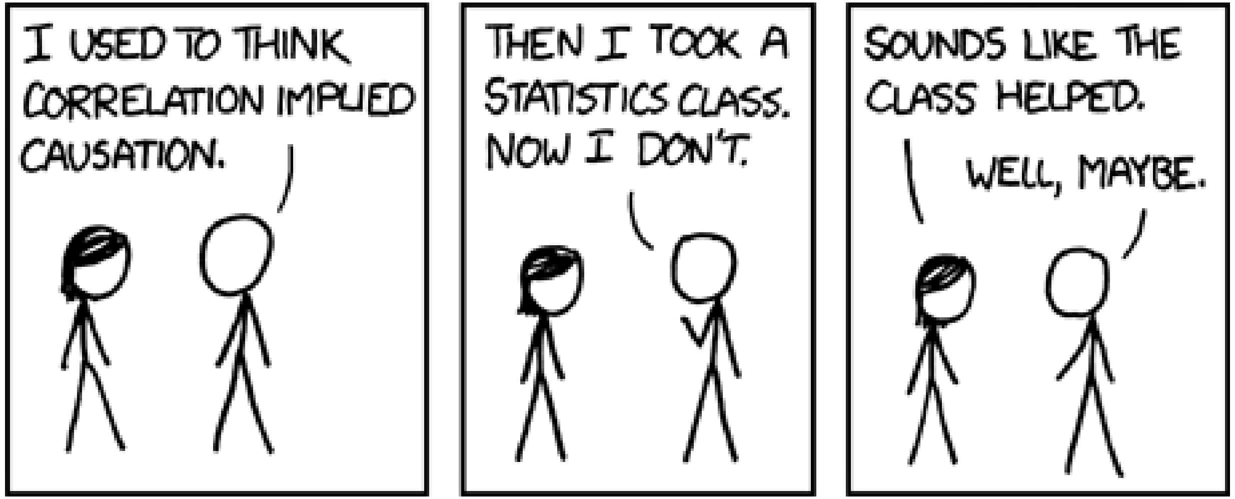
```r
if (!require("pacman"))
  install.packages("pacman")

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


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Source: XKCD

Looking forward

- Until now, our focus what on prediction.
- However, what we economists mostly care about is causal inference:
  - What is the effect of class size on student performance?
  - What is the effect of education on earnings?
  - What is the effect of government spending on GDP?
  - etc.
- Before we learn how to adjust and apply ML method to causal inference problems, we need to be explicit about what causal inference is.
- This lecture will review two dominant approaches to causal inference, the statistical/econometric approach and the computer science approach.

Pearl and Rubin

Source: The Book of Why (Pearl and Mackenzie)

A note on identification

- The primary focus of this lecture is on identification, as opposed to prediction, estimation and inference.

- In short, identification is defined as

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

- More specifically, think about identifying the parameter of interest when you have unlimited data (the entire population).

Potential Outcomes

The road not taken

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

Notation

- Y is a random variable
- X is a vector of attributes
- $\mathbf{X}$ is 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}.

Treatment effect and observed outcomes

- Individual treatment effect: 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 is the *control* group.

Selection bias

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

However, this might be misleading:

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

> **Causal inference is mostly about eliminating selection-bias**

EXAMPLE: Individuals who go to private universities probably have different characteristics than those who go to public universities.

Randomized control trial (RCT) solves selection bias

In an RCT, the treatments are randomly assigned. This means entails that D_i is *independent* of potential outcomes, namely

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

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

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

EXAMPLE: In theory, randomly assigning students to private and public universities would allow us to estimate the ATE going to private school have on future earnings. Clearly, RCT in this case is infeasible.

Estimands and regression

Assume for now that the treatment effect is constant across all individuals, i.e.,

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

Accordingly, we can express Y_i as

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

Or more conveniently

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

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

Unconfoundedness

Typically, in observational studies, treatments are not randomly assigned. (Think of $D_i = \{\text{private}, \text{public}\}$.)

In this case, identifying causal effects depended on the *Unconfoundedness* assumption (also known as "selection-on-observable"), which is defined as

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

In words: treatment assignment is independent of potential outcomes *conditional* on observed X_i, i.e., selection bias *disappears* when we control for X_I.

Adjusting for confounding factors

The most common approach for controlling for X_i is by adding them to the regression:

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

COMMENTS:

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

2. If D_i is randomly assigned, adding X_i to the regression **might** increases the accuracy of ATE.

3. If D_i is assigned conditional on X_i (e.g., in observational settings), adding X_i to the regression eliminates selection bias.

Illustration: the OHIE data

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

Load the OHIE data

In this illustration we will join 3 seperate (stata) files and load them to 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 entire OHIE data can be found here.

Preprocessing: Joining datasets

Join 3 data frames and remove empty responses:

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

Preprocessing: Refinement

Select the relevant variables and re-level `numhh_list` (household size)

```r
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]
## # ... with 23,482 more rows
```

Distribution of treated-control

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

treatment	n
0	11811
1	11681

Estimating ATE

The estimated model

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

In R:

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

Results

```r
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 that you visit primary care physician in the following year by 5.72 [4.47, 7.97] percentage points.

Adjustments

One issue with OHIE is that people are able to apply for Medicaid for their entire household.

This fact undermines the critical random assignment assumption since belonging to larger households increases the chances of being selected to Medicade.

```
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:

```r
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`, ATE has increased from 5.72 to 6.35 percentage points. (Can you guess why?)

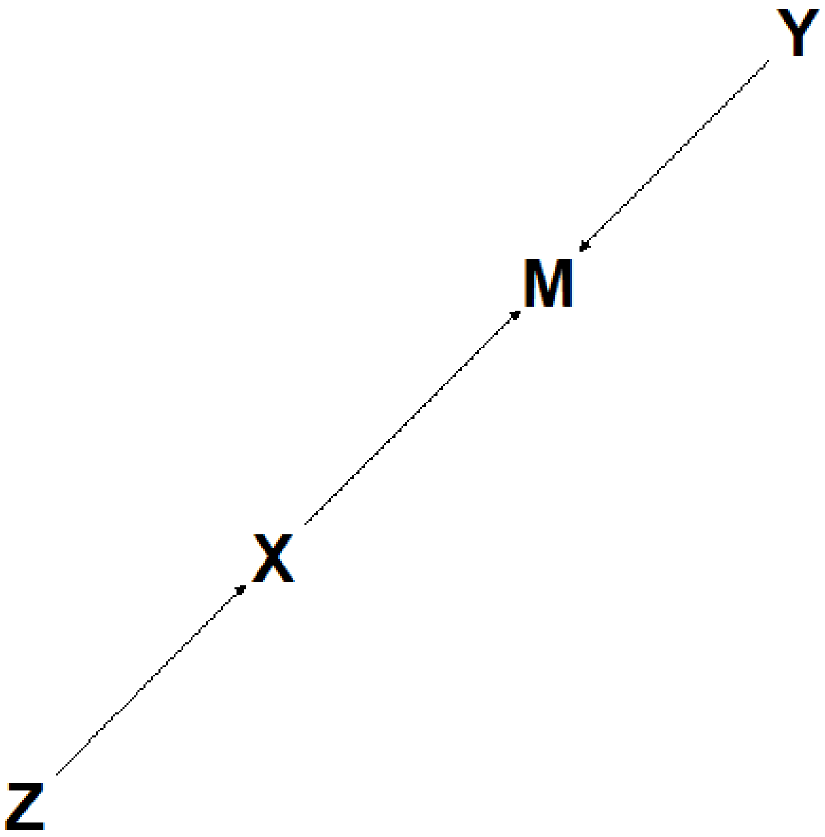
Directed Acyclic Graphs

DAGs

What are DAGs?

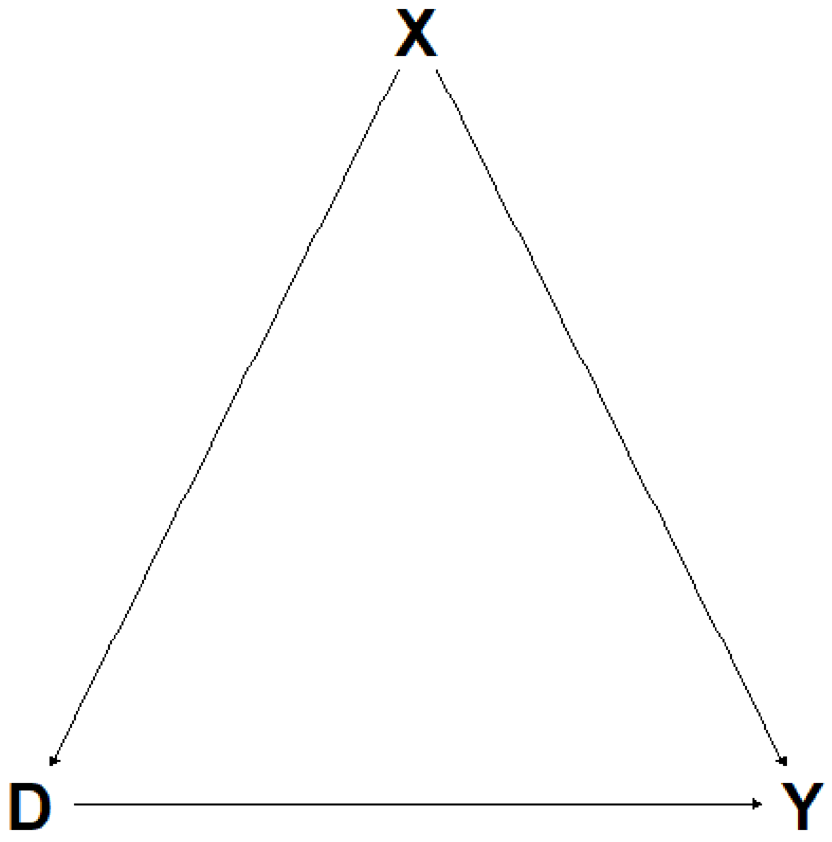
A DAG (directed acyclic graph) is a way to model a system of causal interactions using graphs.

- **Nodes** represents random variables, e.g., X, Y, etc.
- **Arrows** (or directed edges) represent "from $\rightarrow$ to" causal effects. For example, $Z \rightarrow X$ reads " Z causes X".
- A **path** is a sequence of edges connecting two nodes. For example, $Z \rightarrow X \rightarrow M \leftarrow Y$ describes a path from Z to Y.
- In a **direct path** arrows point to the same direction: $Z \rightarrow X \rightarrow M$



Confounder DAG

- X is a common cause of D and Y.
- conditioning on X removes dependency between D and Y through X.
- In DAG terms, controlling for X "closes the backdoor path" between D and Y, and leaves open the direct path.
- The notion of closing the backdoor path is related to the notion of omitted variable bias.



DAGs and SEM

- Another way to think about DAGs is as non-parametric **structural equation models** (SEM)

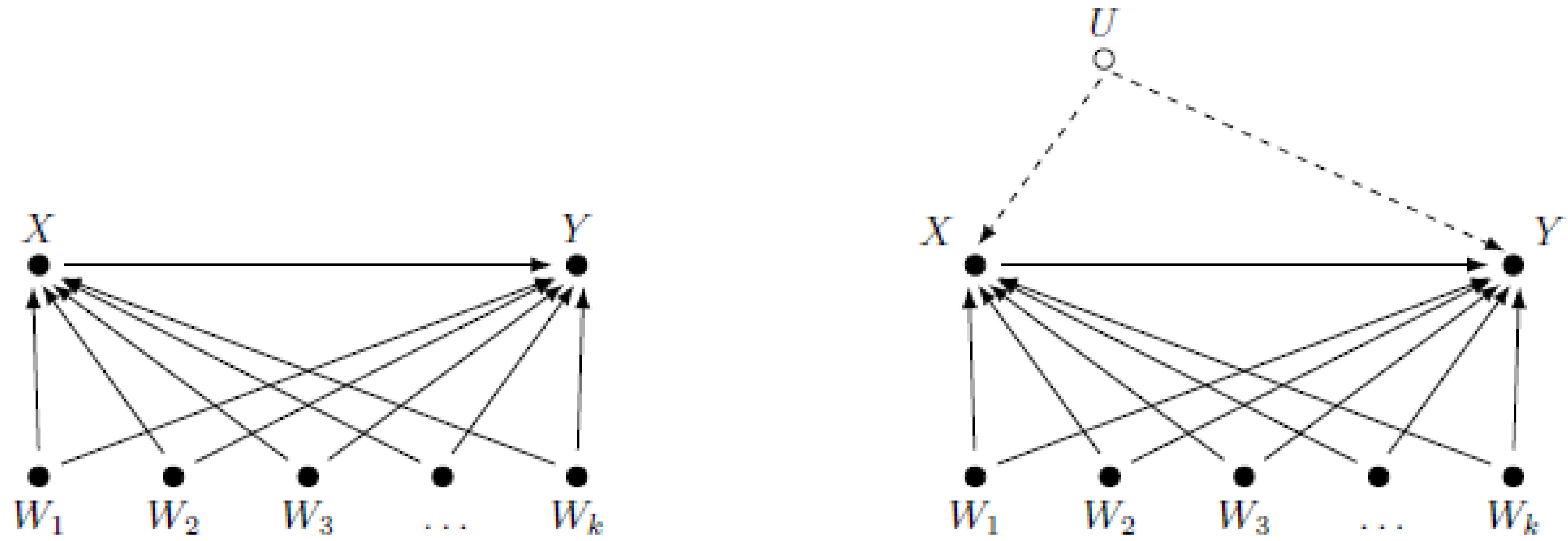
- For example, the single-confounder DAG we've just seen can be represented by a set of three equations:

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

where

- The f_i's denote the causal mechanisms in the model. Are not restricted to be linear.
- u_X, u_D,and u_Y denote independent background factors that the we chooses not to include in the analysis.
- Assignment operator ($\leftarrow$) captures asymmetry of causal relationships.

Unconfoundedness in DAGs



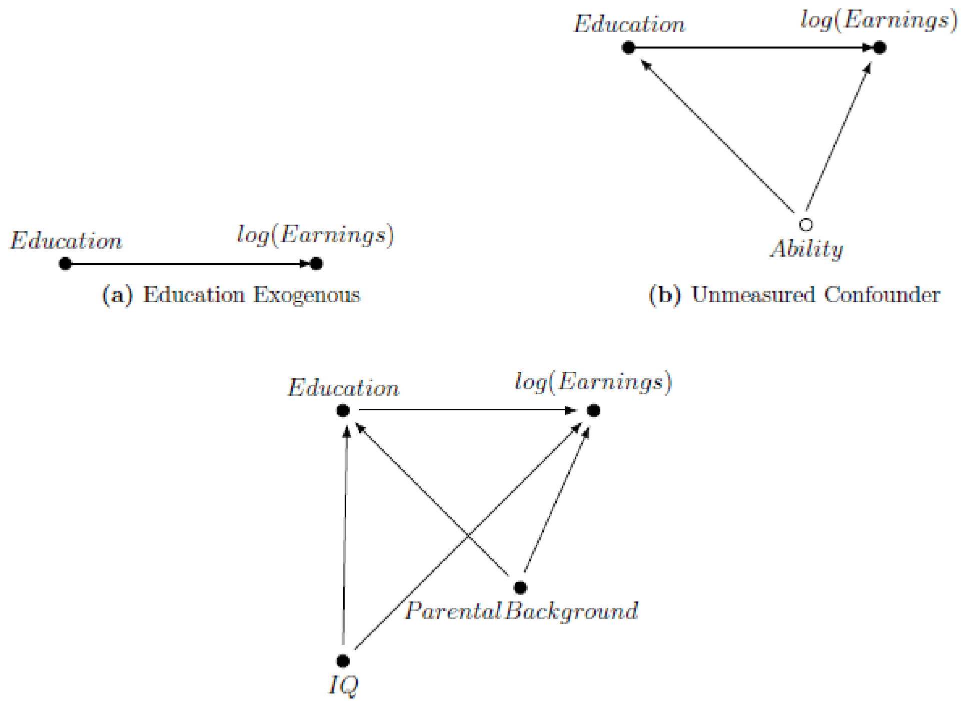
(a) Confoundedness

(b) Violation of Unconfoundedness

Figure 8: Unconfoundedness with Multiple Observed Confounders

Source: Imbens (forthcoming).

Example: Identifying the Returns to Education



(a) Education Exogenous

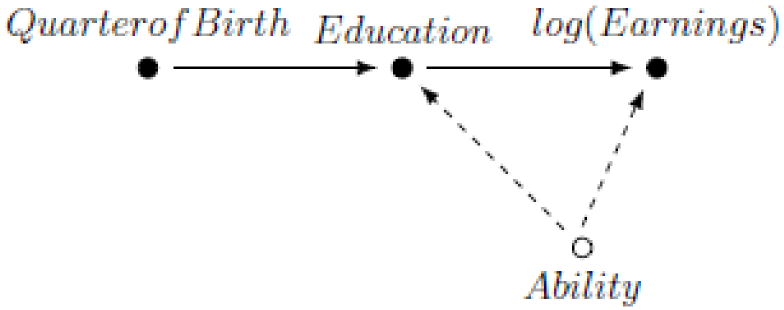
(b) Unmeasured Confounder

(c) Unconfoundedness

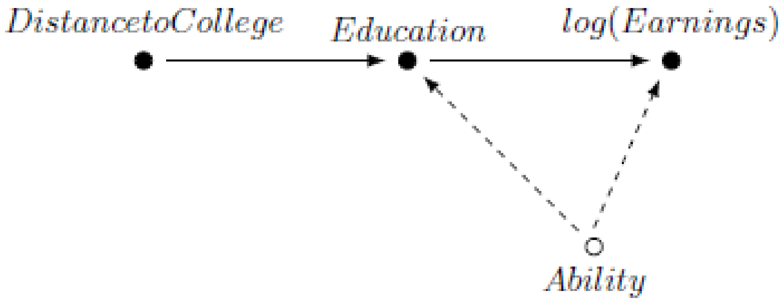
Figure 15: DAGs for the Returns to Education (I)

Source: Imbens (forthcoming).

Instrumental variables in DAGs



(a) Instrumental Variables: Quarter of Birth

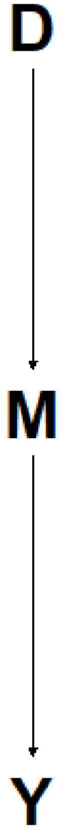


(b) Instrumental Variables: Distance to College

Source: Imbens (forthcoming).

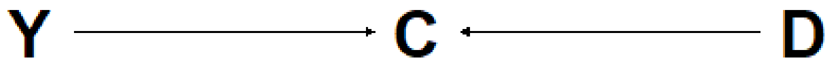
A mediator

- D causes M causes Y.
- M mediates the causal effect of D on Y
- conditioning on M removes dependency between D and Y
- We've essentially closed a direct path (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

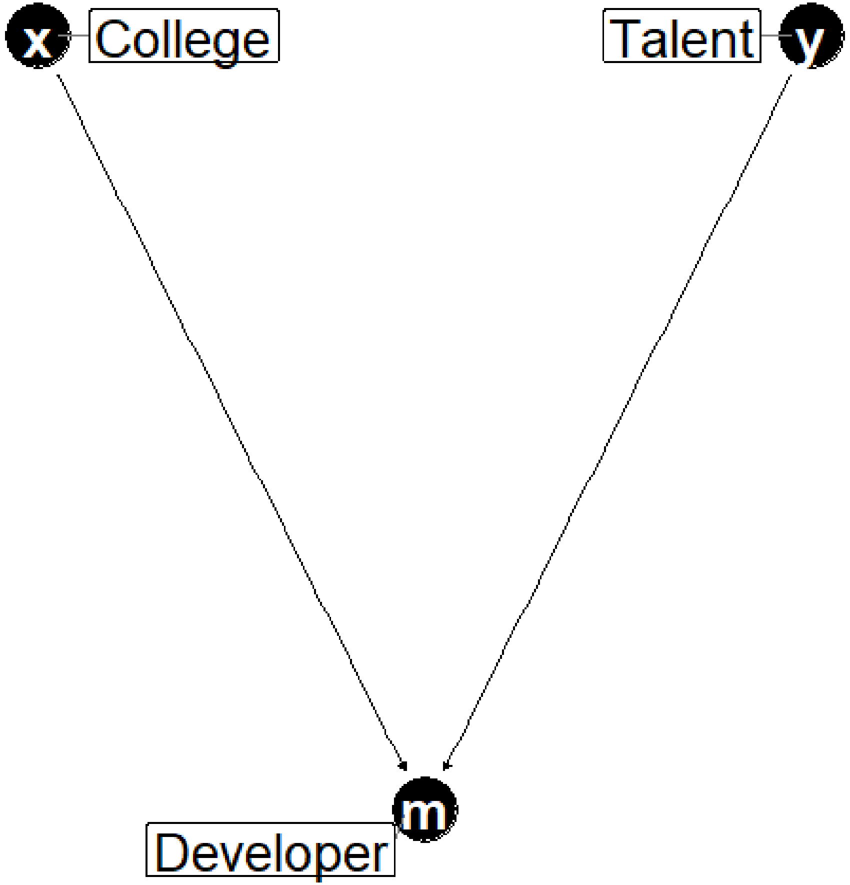


Example: "Bad controls"

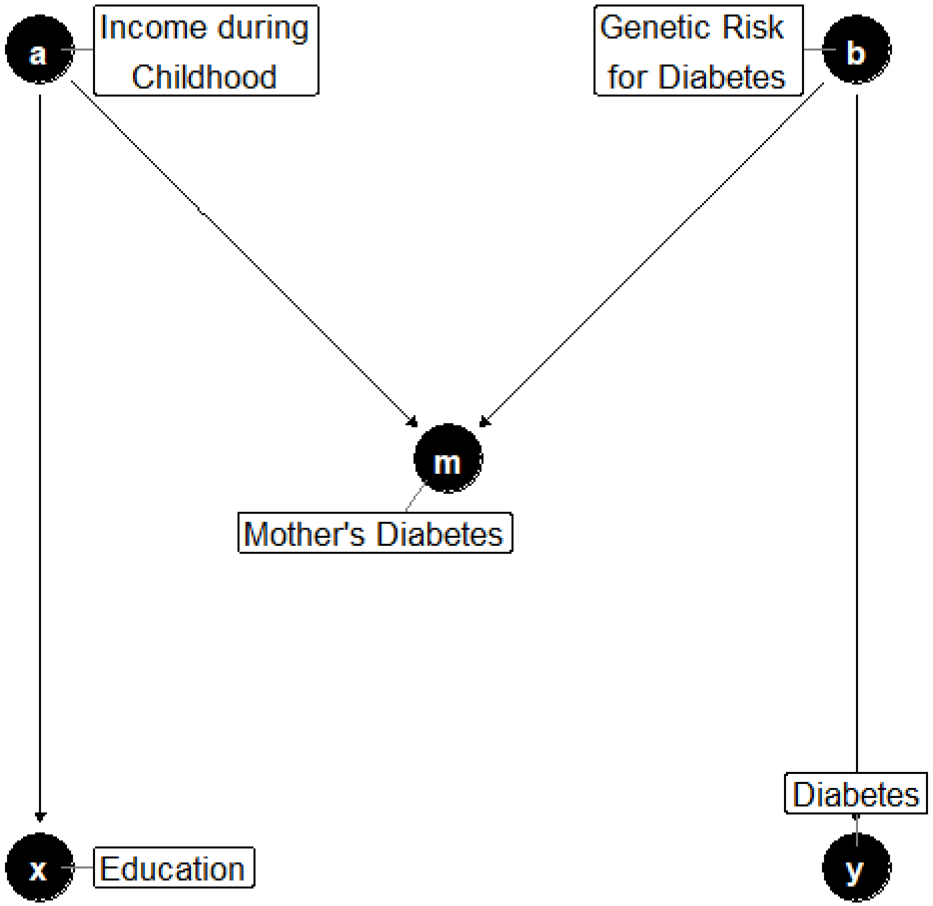
- "Bad controls" are variables that are themselves outcome variables.
- This distinction becomes important when dealing with high-dimensional data

EXAMPLE: Occupation as control in a return to years of schooling regression.

Discovering that a person works as a developer in a high-tech firm changes things; knowing that the person does not have a college degree tells us immediately that he is likely to be very talented.



Collider: M-bias



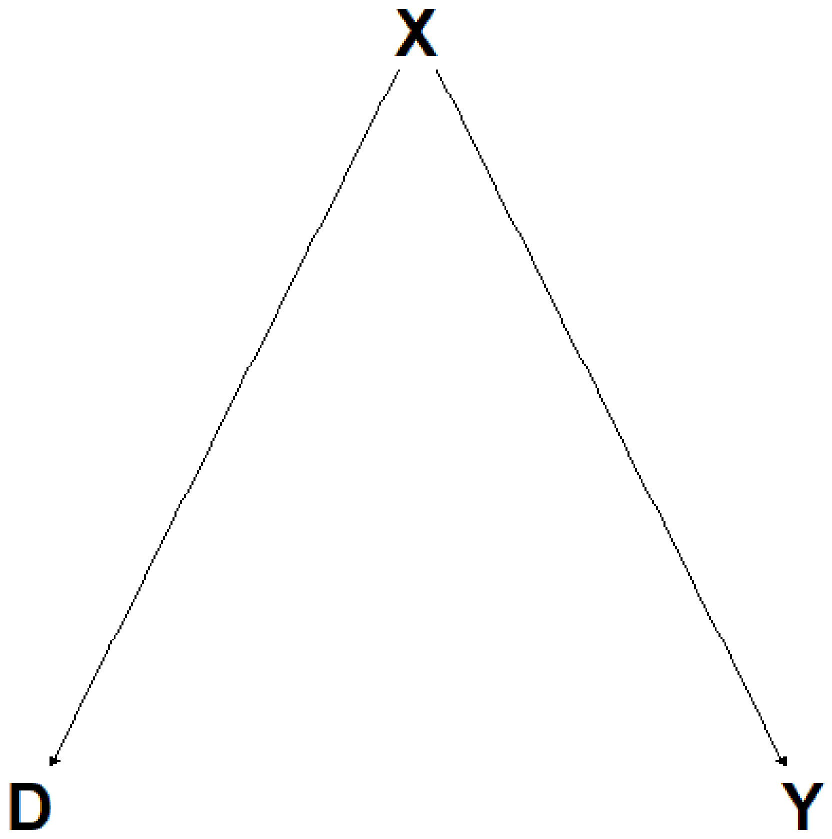
Simulations

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 $D \to Y$ is zero (i.e., $ATE = 0$).

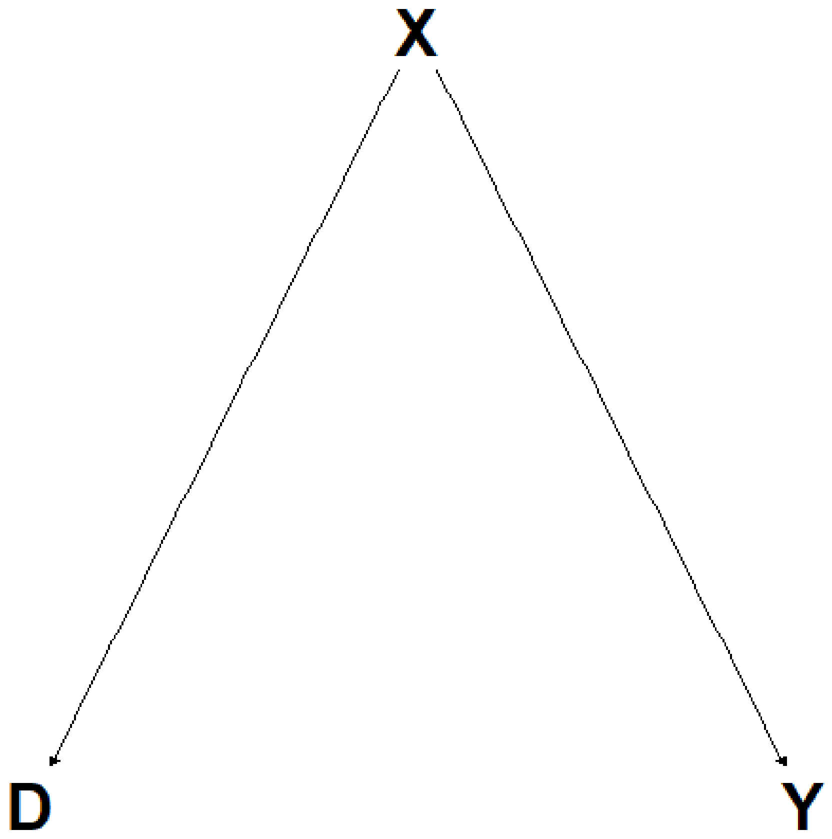


Simulation I: De-counfounding (cont.)

Raw correlation matrix:

	y	x	d
y	1.0	0.1	0.1
x	0.1	1.0	0.8
d	0.1	0.8	1.0

Note: Y and D are correlated even though there is no direct arrow between them. This is due to the confounder X which opens a backdoor path between Y and D.



Simulation I: De-counfounding (cont.)

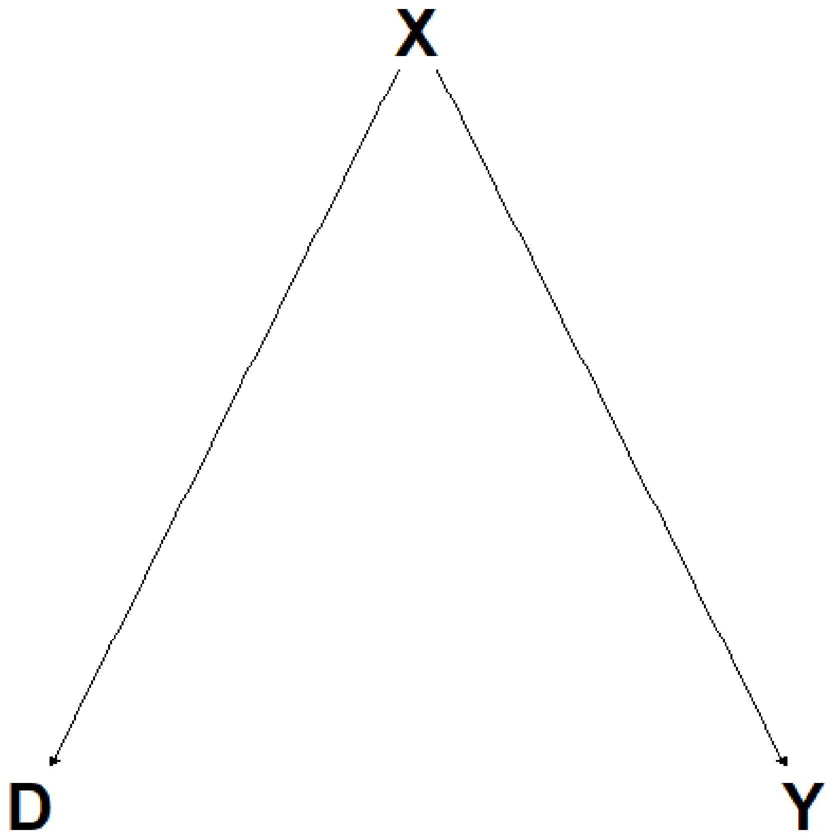
Let's estimate the model with X on the right hand side:

term	estimate	p.value
d	0.01	0.81
x	0.15	0.01

and without X

term	estimate	p.value
d	0.12	0

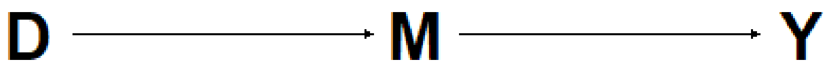
BOTTOM LINE: Controlling for X provides the correct answer.



Simulation II: Mediator

The DGP:

```r
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]
```



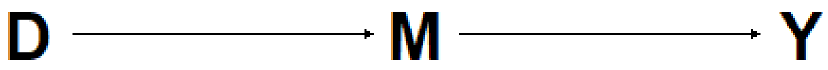
True effect of $D \to Y$ is $1.3 \times 0.1 = 0.13$.

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.



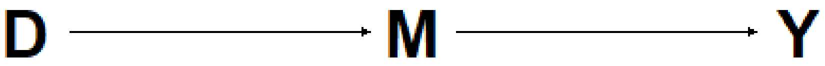
Simulation II: Mediator (cont.)

Estimate the model with M:

term	estimate	p.value
d	-0.03	0.59
m	0.11	0.00

and without M:

term	estimate	p.value
d	0.12	0



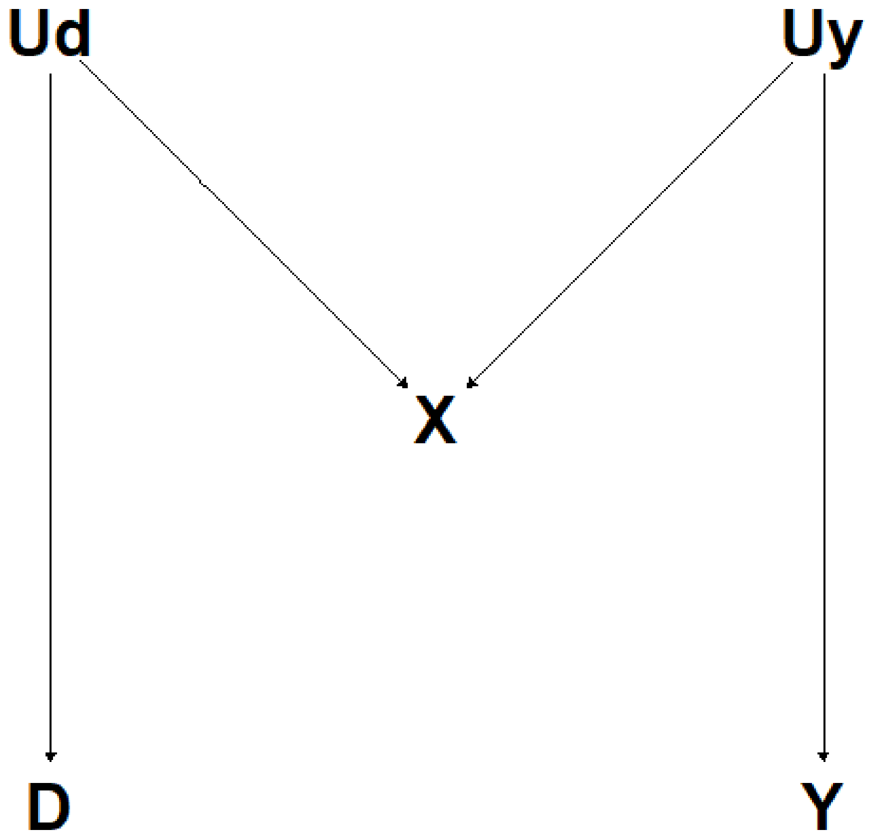
BOTTOM LINE: Controlling for M in this case biases the total effect of D on Y downward since it blocks the path from D to Y.

Simulation III: M-bias

Generate the data:

```r
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 $D \to Y$ is zero (i.e., $ATE = 0$).

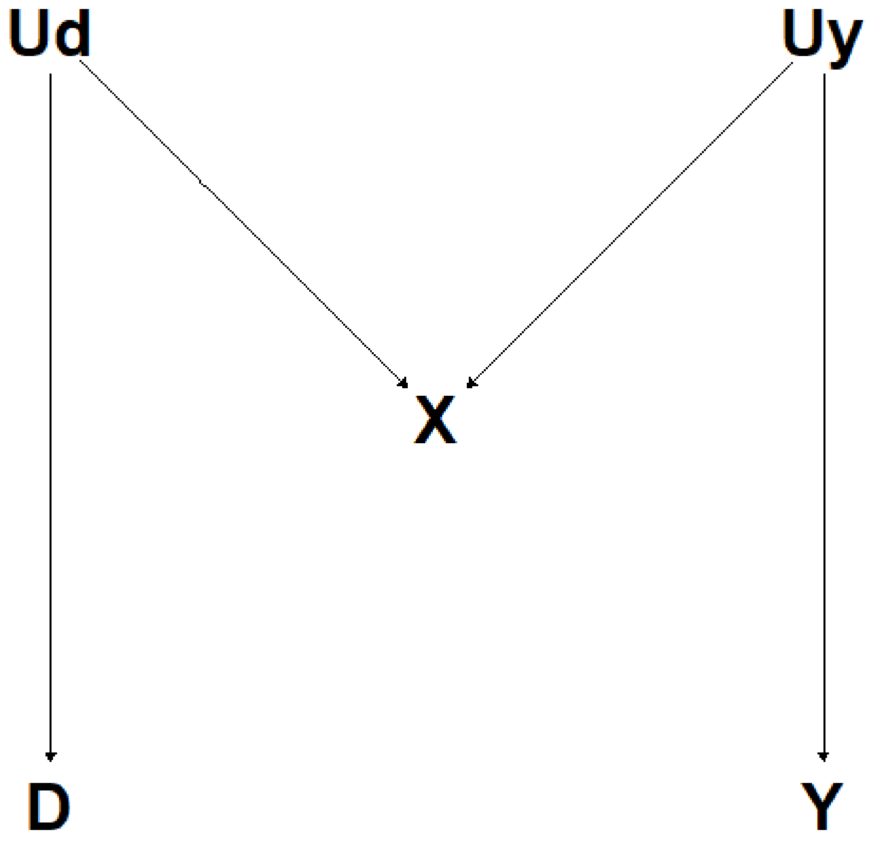


Simulation III: M-bias (cont.)

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

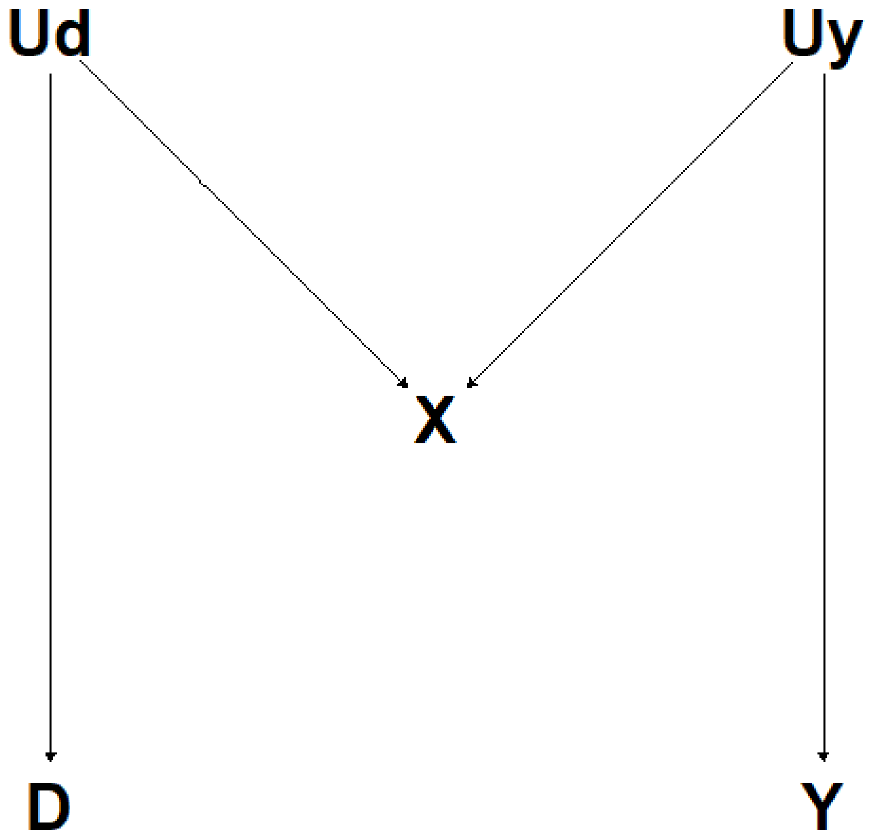
Estimate the model with X

term	estimate	p.value
d	-0.37	0
x	0.49	0

and without X

term	estimate	p.value
d	0.01	0.69

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

- Hard to write down a DAG for complicated (econometric) structural models.
- Need to specify the entire DGP (it REALY a limitation?)
- Simultaneity: *"In fact it is not immediately obvious to me how one would capture supply and demand models in a DAG"* (Imbens, forthcoming)

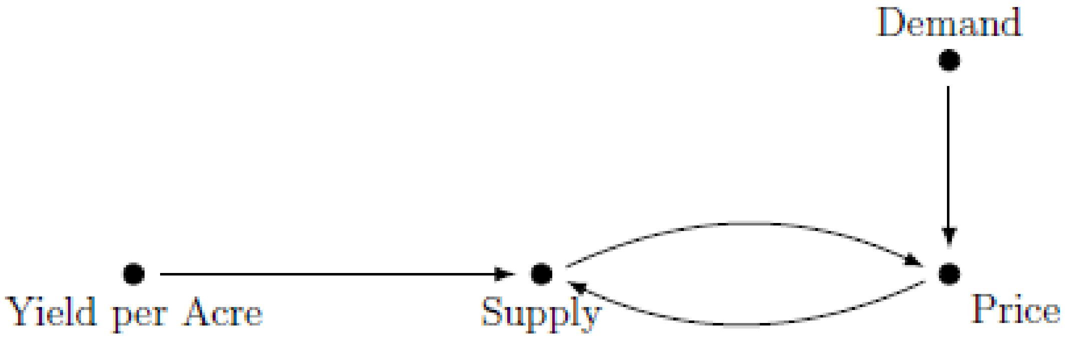


Figure 11: Based on Figure 7.10 in TBOW, p. 251.

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 setting

Consider again the standard "treatment effect regression":

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

Our object of interest is $\hat{\tau}$, the estimated *average treatment effect* (ATE).

In high-dimensional settings $k \gg n$.

```
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. (forthcoming). 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.