Lecture 2. Introduction

Vinish Shrestha

Towson University

Motivation

A Lab Experiment

Challenges

Directed Acyclic Graph (DAG)

A simulated DGP

Discussion

Motivation

Cause and Effect

- A fundamental pursuit in science
- theory
- experiments
- computational models

Examples

1. Does increasing the mass of an object cause it to fall faster under gravity in the absence of air resistance?
2. Does light (sound) travel in a vaccum?
3. Intensity:
    - What's the magnitude of gravitational force?
4. Can tarrif reduce inflation?

Experimental setting

- The researcher has control over key aspects of the approach

1. Assignment of the treatment
2. Controls
3. Measurement of outcomes

- Limited luxury of conducting controlled experiments

The purpose of lectures

- Focus on challenges concerned with causal inference
- Address experimental and Observational setting
- Emphasis on empirical approach, simulations, and application to the real world data

A Lab Experiment

Jade: A thought experiment

- Suppose you want to investigate the importance of sunlight for the growth of a jade plant

You might

1. Obtain two baby jade plants from the same seller
2. Place jade plant A near the window sill, where it can receive ample sunlight, and place jade plant B in a closet within the same room
3. Keep the room temperature uniform and water both plants equally, once a week
4. After some time, you evaluate the growth of both plants

you wouldn't actually do this because we already know the plant in the closet would not survive

The construct of the throught experiment

1. Treatment assignment: exposure to sunlight
2. Controlling for other variables: water, humidity
3. Measurement: growth, number of leaves

Challenges

A fundamental challenge

- Say, we want to study the relationship between having a college degree and health outcomes
- Ideally, you'd want to observe an individual, say person A, in two states *(potential outcomes)*
  - with college level education
  - without college level education
- Measure her health outcomes in these two states and make comparison

Another challenge

- Limited lab experiments.

Observational data

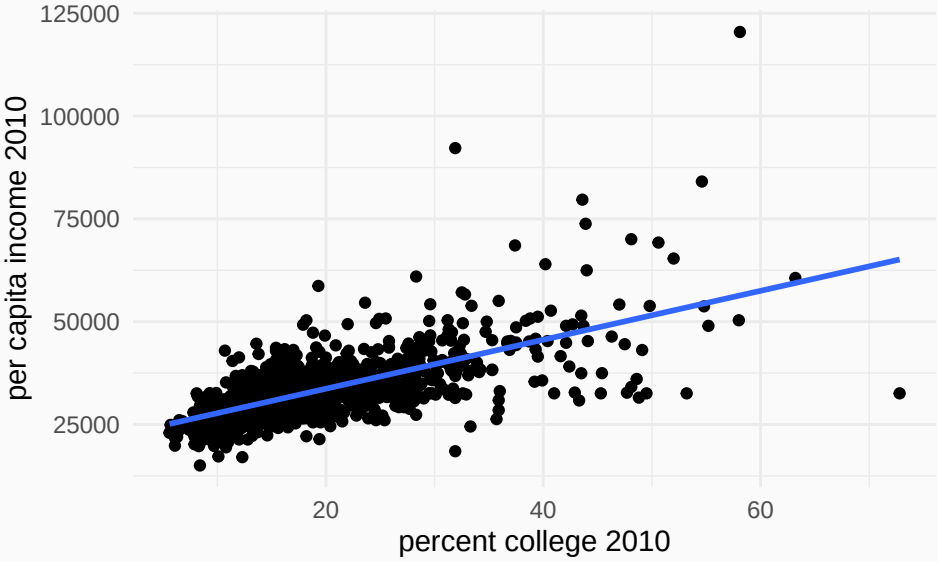
- You'll have to tackle your research questions with observational data
  - surveys
  - administrative data
- Let's use the county level data for 2010:

a. **Dependent variable.** prevalence of poor health, life expectancy

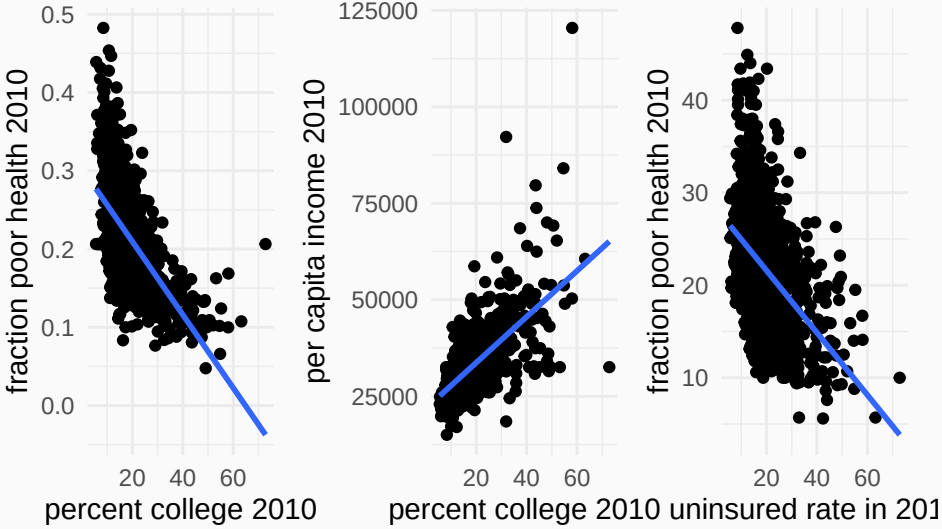
b. **Independent variable.** percentage of population with a college degree

c. What about controls?

Relationship between percent college and per capita income



Let's take a look at more wholesome illustrations.



Directed Acyclic Graph (DAG)

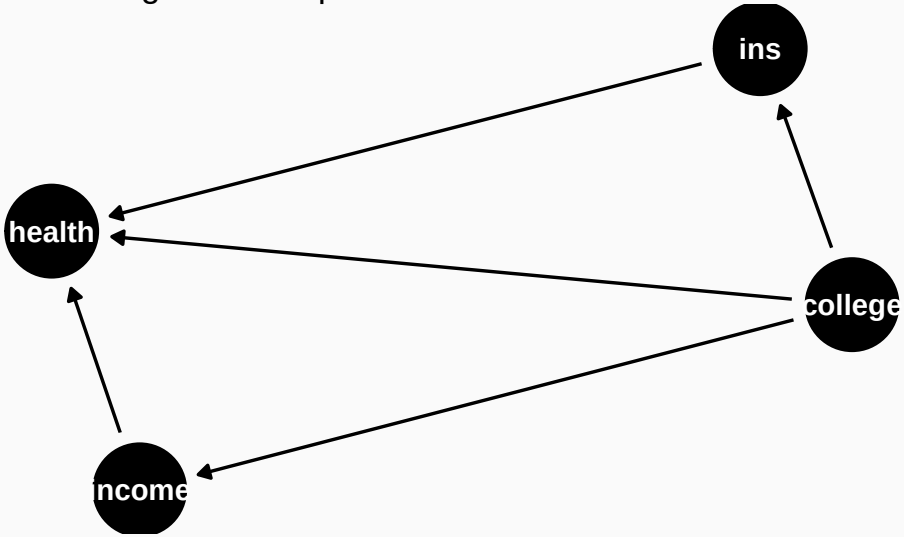
DAG

- Causal diagram: A simple way to keep track of whats going on.
- Directed Acyclic Graph (DAG): in various fields like statistics, computer science, and epidemiology
- Used to depict the relationship between variables

Before we move on let's discuss the DGP

- **Data Generating Process.** *represents the underlying set of mechanisms or laws of the universe that produce the data we observe.*
- These mechanisms are not immediately apparent to us
- Our goal is to uncover and understand the phenomena governing the DGP
- DAG can be used to depict the DGP

Causal Diagram Example A.



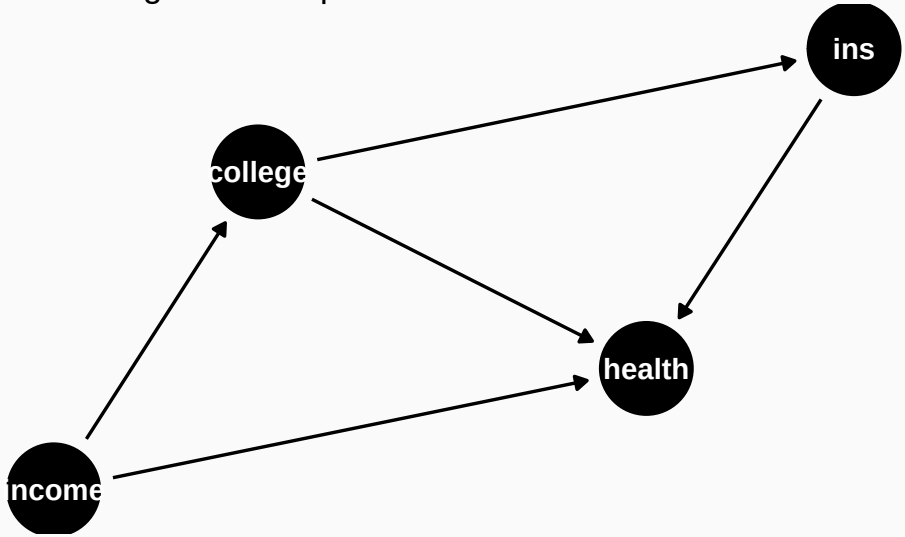
In the DGP as illustrated by DAG

1. College affects Health
2. College affects insurance status. Insurance affects health. (mechanism through how college affects health).
3. College affects income. Income then affects health. (another mechanism through which college affects health).

- Mechanisms through which college affects health are *good pathways*
- *DAG in this example falls way too short in explaining the actual DGP. WHY?*

DAG: Allow income to cause health

Causal Diagram Example B.

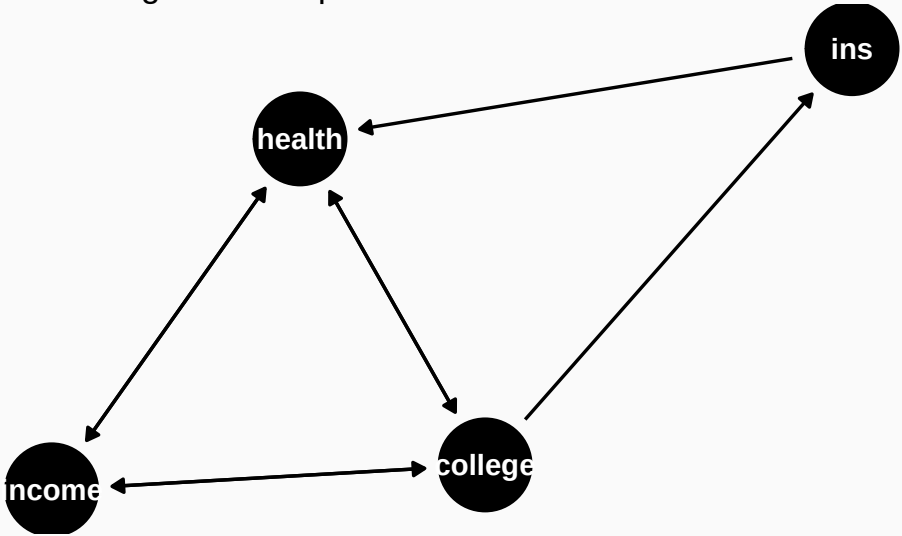


How to identify the effect of college education on health?

- Income causes both health and education
  - creates confounding effects
- Need to isolate the effect of college education on health
- Look at individuals with the same income and utilize the variation in college education
  - income of $50,000
  - some will have college education and some won't
- This variation in college education can be fruitful in identification

We've accounted (controlled) for income. However, we'r agnostic about the functional form.

Causal Diagram Example C.



Description of the new DAG

Note that now arrows are facing both ways for income, health, and college. We have a feedback loop between income, health, and college.

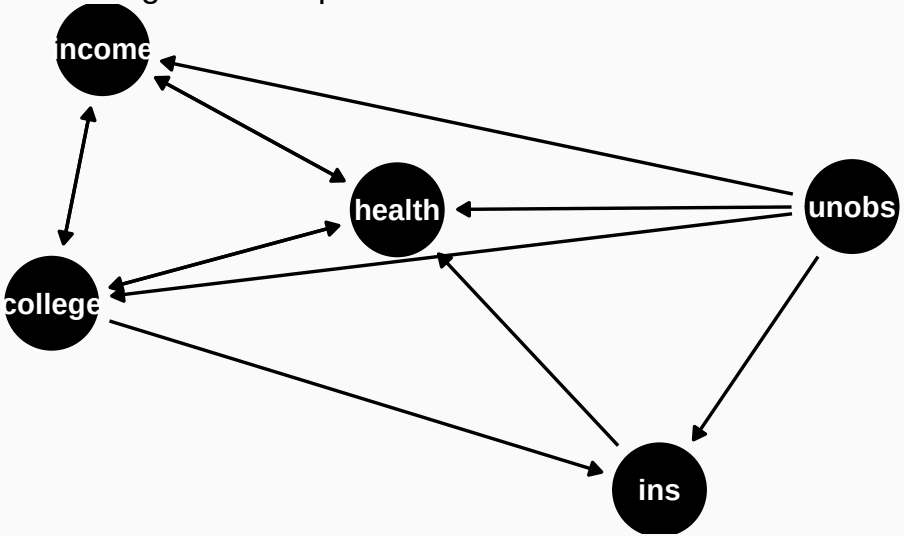
This means:

1. Income can affect health; health can affect income.
2. Income can affect college education; college can affect income.
3. College education can affect health; health can affect college education.
    - **reverse causality**

- Block unnecessary backdoors: i) income to college; ii) health to income; iii) health to college

DAG with unobserved component

Causal Diagram Example D.



DAG with unobserved component

- This DAG perhaps most closely represent the DGP?

There are two limitations here

1. Don't have data for unobserved variables
   - Belongs to the data generating process; but you don't have data for them
   - Data limitations $\rightarrow$ **omitted variables**
2. **Reverse causality**
   - the effect runs borth from health to college and college to health

- A significant portion of causal inference is about alleviating the concerns of omitted varables and reverse causality.

A simulated DGP

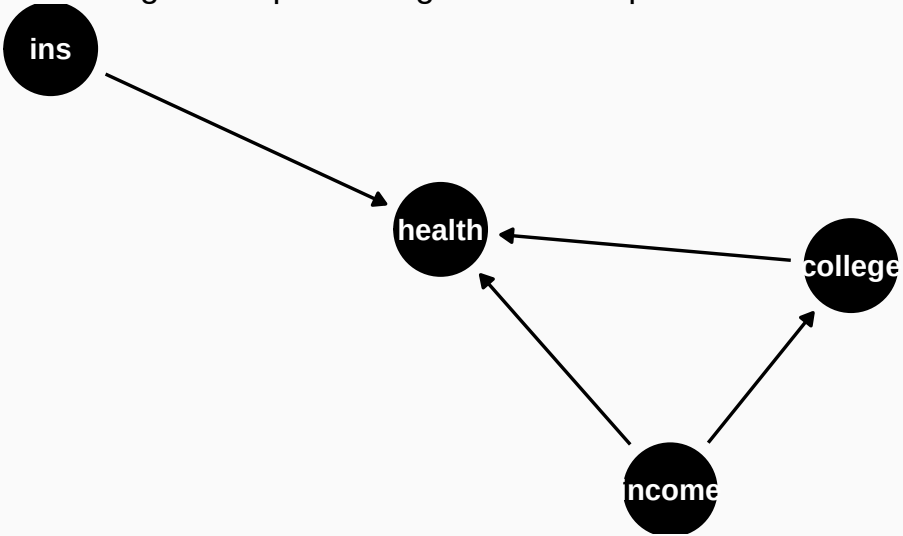
Let's simulate a DGP

For our understanding

1. College education boosts health by 10 percent.
2. Income boosts health by 20 percent.
3. 40 percent more people from higher income households have college education.
4. Having insurance boosts health by 5 percent.

DAG for the simulated DGP

Causal Diagram Representing the Made−up DGP.



Code to simulate the DGP

- Let's start with generating Log normal income

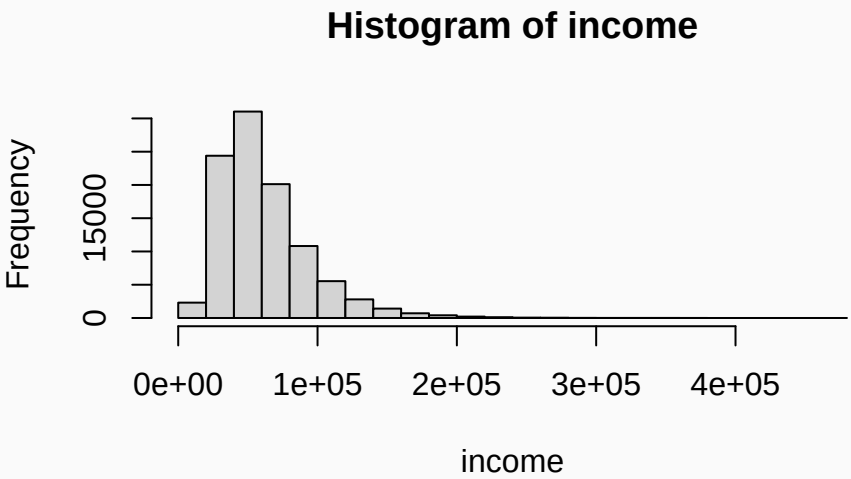
```
# number of observations
n   <- 100000

# income follows the log normal distributing
income  <- rlnorm(n, meanlog = 1, sdlog = 0.5)

# multiplying the log normal dist with 20000
income  <-  income * 20000
```

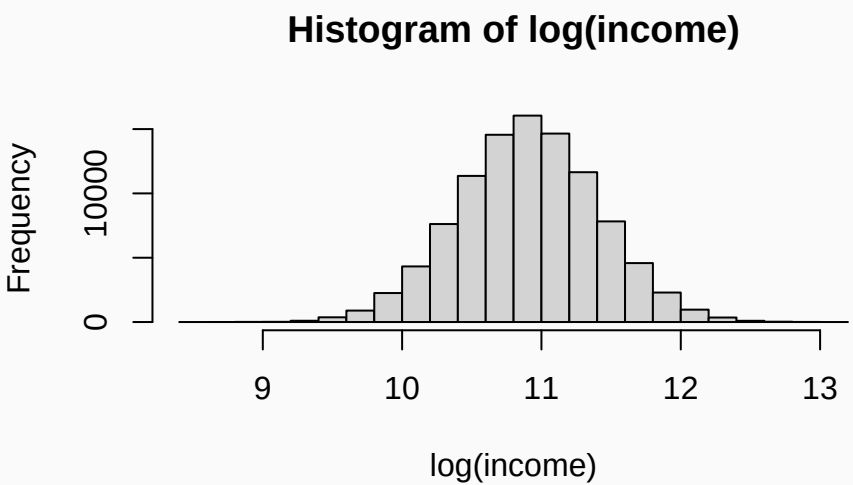
Histogram of income

```
# a right skewed distribution
hist(income)
```



Histogram of log income

```
# a log normal distribution
hist(log(income))
```



A code to simulate DGP

```r
# high income
high_income  <- ifelse(income > median(income), 1, 0)

# college education
college  <-  rbinom(n, 1, 0.3 + 0.4 * high_income)

# proportion of college graduates by income status
college_lowinc  <- table(college[high_income == 0])
college_highinc  <-  table(college[high_income == 1])
```

College graduates by income status

```
cat(paste("Number with no college for low income:", college_lowinc[[1]]))
```

```
## Number with no college for low income: 34916
```

```
cat(paste("Number with college for low income:", college_lowinc[[2]]))
```

```
## Number with college for low income: 15084
```

```
cat(paste("Number with no college for high income:", college_highinc[[1]]))
```

```
## Number with no college for high income: 15031
```

```
cat(paste("Number with college for high income:", college_highinc[[2]]))
```

```
## Number with college for high income: 34969
```

A code to simulate DGP (treatment effect = 0.1 or 10 pp)

```
# insurance status
ins  <-  rbinom(n, 1, 0.5)

# health (good health 1, poor health 0)
# 60 percent of people with no college, low income, and no insurance have
# 10 percent more of people with college have good health and so on.
health  <- rbinom(n, 1, 0.6 + 0.1 * college + 0.2 *
            high_income + 0.05 * ins)

table(health)
```

```
## health
##     0     1
## 22426 77574
```

A glimpse of the dataframe

```
data  <- data.frame(good_health = health, income = income, highincome = hig
                    college = college, insurance = ins)
head(data)
```

```
##   good_health   income highincome college insurance
## 1           1 90585.00          1       1         0
## 2           1 60593.72          1       1         0
## 3           1 91343.45          1       0         1
## 4           0 56944.08          1       1         0
## 5           0 36702.60          0       0         0
## 6           1 50477.65          0       1         1
```

Building Models

```
reg1  <- lm(good_health ~ college, data = data)

reg2  <- lm(good_health ~ college + income, data = data)

reg3  <- lm(good_health ~ college + highincome, data = data)

reg4  <- lm(good_health ~ college + income + highincome, data = data)

reg5  <- lm(good_health ~ college + highincome + ins, data = data)
```

Results

Results

Table 1: Regression Results

	Dependent variable:				
	good health				
	(1)	(2)	(3)	(4)	(5)
college	0.177***	0.140***	0.099***	0.099***	0.099***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
income		0.00000***		0.00000	
		(0.00000)		(0.00000)	
high income			0.196***	0.195***	0.197***
			(0.003)	(0.004)	(0.003)
insurance					0.048***
					(0.003)
Constant	0.687***	0.584***	0.628***	0.627***	0.603***
	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
Observations	100,000	100,000	100,000	100,000	100,000

Note: *p<0.1; **p<0.05; ***p<0.01

Discussion

Discussion

- Discussed some fundamental blocks for starting a research project
- DGP might be constructed in the real world
- *good* versus *bad* pathways
- omitted variable bias and reverse causality

Ultimately, we are trying to understand something (new) about the DGP. In reality, the laws governing the DGP may not be that simple. That means we should try harder with persistence and creativity.