INFO371: Causal Inference

Ott Toomet

April 12, 2018

What is *Cause*?

- Link between two events, *cause* and *effect*
- Effect (partly) dependent on cause
- Cause precedes the effect in time
- 3 Cases
    - Sufficient and necessary
    - Sufficient but not necessary
    - Necessary but not sufficient

Does Flu Shot Help to Avoid Illness?

- Collect data (get medical data):

gotFlu	fluShot
1	0
0	1
...	...

- Run a regression:

$$\text{gotFlu}_i = \alpha + \beta \cdot \text{fluShot}_i + \epsilon_i$$

- We want to know β
  - What does the regression tell?

A Causal Question

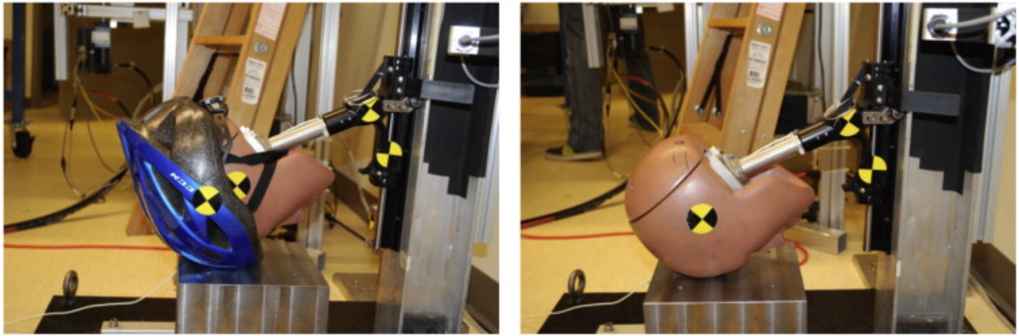
- Many important questions are causal
  - Does college pay off?
  - Which career path should I choose?
  - Does the drug cure illness?
  - Does the advertisement work?
- These models can be written as

$$Y_i = \alpha + \beta \cdot T_i + \epsilon_i$$

What are *outcome* and *treatment*?

Example:

> *Bicycle helmets are highly effective at precenting head injury [. . .] (Cripton et al, 2014).*



Should we make helmets mandatory?

- In the paper, *treatment* is helmet and outcome is head injury.
- It is not *helmet law* and *public health*
- There are many additional factors:
  - Helmets may hurt in certain cases (rotational injury)
  - Helmeted cyclists may opt for less safe riding (risk compensation)
  - Motorists may opt for less safe techniques
  - When arguing for helmets, one argues that cycling is unsafe
    - Less cycling, more driving
    - Lower *safety in numbers*
    - *Worse* public health outcomes
  - Ditto with inconvenience of helmets

Reflexivity

Want to earn *a lot* of money?

1. Find stocks that are rising
2. Buy!
3. Sell later
4. Enjoy life ☺

Problem:

- Everyone else is also going to buy the same stocks
- Stocks that were profitable a short while ago are not any more
- Have to re-assess the strategy all the time

Humans adapt to interventions and render these less effective

How to measure outcome?

We want *numeric estimate* of the causal effect of our intervention

- $Y(1)$: outcome when treated $(T = 1)$
- $Y(0)$: when not treated $(T = 0)$
- The *treatment effect*:

$$Y(1) - Y(0) = \beta$$

- Only observe $Y(0)$ or $Y(1)$ but never both!

Ordinary Data

Estimate the regression model:

$$Y_i(T_i) = \alpha + \beta \cdot T_i + \epsilon_i$$

The regression coefficient is the average difference over $T = 1$ and $T = 0$ cases:

$$\begin{aligned}
\mathbb{E}[Y(1) - Y(0)] = & \\
& = \mathbb{E}\,Y(1) - \mathbb{E}\,Y(0) = \\
& = \mathbb{E}[\alpha + \beta \cdot 1 + \epsilon | T = 1] - \mathbb{E}[\alpha + \beta \cdot 0 + \epsilon | T = 0] = \\
& = \beta + \mathbb{E}[\epsilon | T = 1] - \mathbb{E}[\epsilon | T = 0]
\end{aligned}$$

This is not β

- unless $\mathbb{E}[\epsilon | T = 1] - \mathbb{E}[\epsilon | T = 0] = 0$

What Does Regression Do

Regression compares Y for two groups: $T = 0$ and $T = 1$.

- And that's it.
- It is "just correlation"
- In order to interpret it causally, we have to ensure $T \perp\!\!\!\perp \epsilon$
  - i.e. treatment is unrelated to the disturbance term

3 Stories

There are always 3 stories:

1. T causes Y
   "Because there are so many guns, there is so much violence"
2. Y causes T
   "Because of the violence, people get guns"
3. something else causes both T and Y
   "Because US history of being a frontier land, people want to have guns, and are used to only rely on themselves when seeking justice"

(sometimes some of the stories not convincing)

Exercise

Amaros et al (2011)

- analyze French medical registries
- compare bicycle accident data with/without helmets
- compare 'similar' injuries outside of the head region
- control for road type, accident type, age, gender
- "…confirm the protective effect for head and facial injuries…"

Can you tell 3 stories?

Answer

1. Helmet causes less head injuries
2. Head injuries cause not wearing helmets
   Does not sound convincing as we *know* the decision to (not to) wear helmet strictly preceded the crash. However, we also have to assume that riders did not want to get injured. These are pieces of extra information.
3. Those who care about their health both wear helmets and ride carefully.
   Although the injuries outside of head region was controlled for, it may still be possible to take more care of your head than rest of the body

Note that this does not address the social psychology issues.

Need Additional Information

We have to know that $\mathbb{E}[\epsilon|T = 1] - \mathbb{E}[\epsilon|T = 0] = 0$

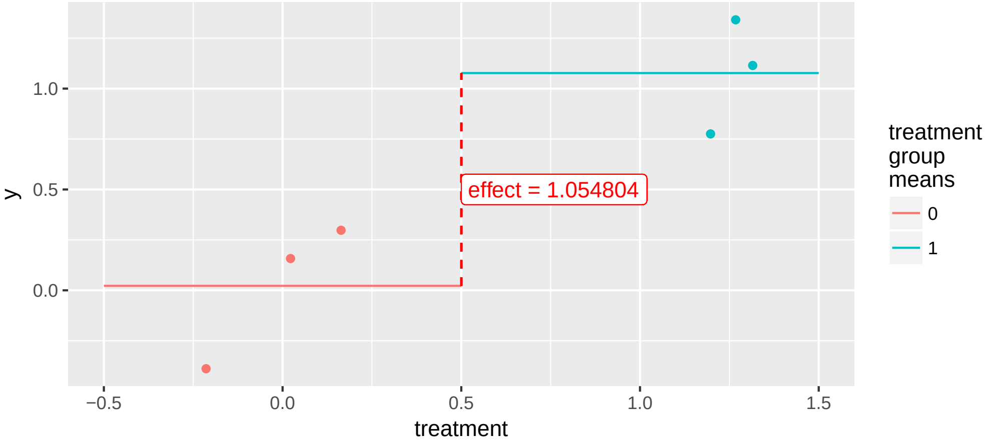
- This information is usually not called "data"
- Examples:
  - This was a randomized experiment
  - Treatment was caused by an external factor (natural experiment)
    - natural event
    - age-dependent rules
    - decision to change rules for some people
  - We know what was treatment into selection based on
  - We know functional form of ϵ.

Note: "big" data does not help here

Cross-Sectional Estimator

```
##            y treatment
## 1  0.1575982         0
## 2  0.2978230         0
## 3 -0.3886009         0
## 4  1.1147280         1
## 5  1.3410852         1
## 6  0.7754202         1
```

Cross-Sectional Estimator



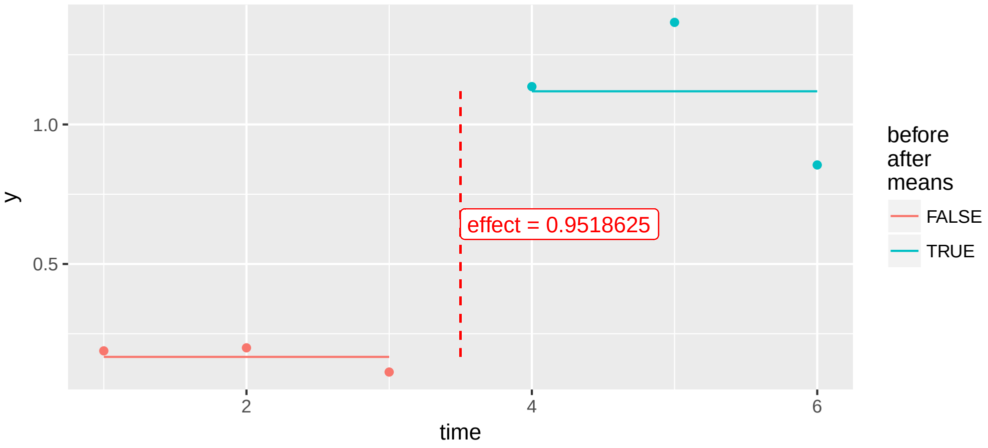
As Regression

```
##
## Call:
## lm(formula = y ~ treatment, data = df)
##
## Residuals:
##        1        2        3        4        5        6
##  0.13532  0.27555 -0.41087  0.03765  0.26401 -0.30166
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.02227    0.18823   0.118   0.9115
## treatment    1.05480    0.26620   3.962   0.0166 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Before-After Estimator

```
##           y time
## 1 0.1887248    1
## 2 0.1996113    2
## 3 0.1127663    3
## 4 1.1357981    4
## 5 1.3659416    5
## 6 0.8549502    6
```

Cross-Sectional Estimator



As Regression

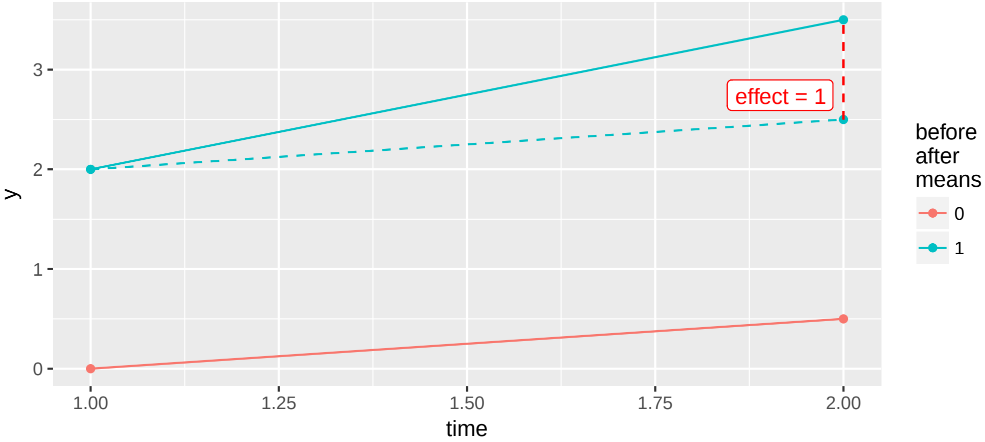
```
##
## Call:
## lm(formula = y ~ time > 3, data = df)
##
## Residuals:
##        1        2        3        4        5        6
##  0.02169  0.03258 -0.05427  0.01690  0.24704 -0.26395
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.1670     0.1062   1.572  0.19102
## time > 3TRUE   0.9519     0.1503   6.335  0.00318 **
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Diff-in-Diff

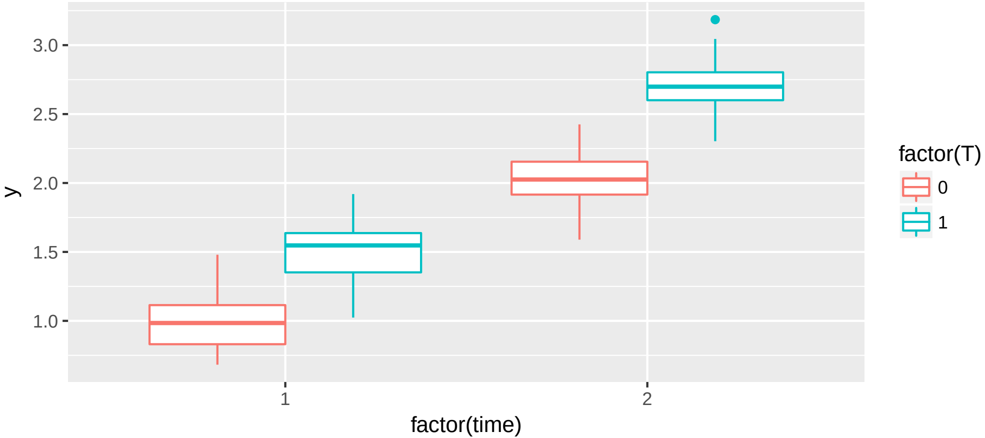
Differences-in-differences

```
##       y time treatment
## 1 0.0    1         0
## 2 0.5    2         0
## 3 2.0    1         1
## 4 3.5    2         1
```

Diff-in-dff



Random Data



Regression

```
##
## Call:
## lm(formula = y ~ time * T, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49302 -0.13115  0.00572  0.12300  0.48529
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.03149    0.06317  -0.499  0.61868
## time         1.02598    0.03995  25.680  < 2e-16 ***
## T            0.36226    0.08518   4.253 3.27e-05 ***
## time:T       0.16002    0.05387   2.970  0.00335 **
## ---
## Signif. codes:
```