# Causality in statistical inference

Michał Jakubczyk
Michał Kot

# Goals (and Agenda)

- Why is causality interesting?
- There is some vocabulary to learn
- We need causality in statistical inference
- And we can infer about causality, too

# Introduction, motivation

# What an econometric model really is?

- $Y_i = \alpha_0 + \alpha_1 \cdot X_{1,i} + \alpha_2 \cdot X_{2,i} + \ldots$

- What does this equation mean?
  - Simplified data set summary?
  - Equation used to create predictions?
  - Relationship useful in planning results of taken decisions?

- Inspiration – how does it differ in machine learning(e.g decision trees)?

# Why causality?

Democritus: *[I would] rather discover one cause than gain the kingdom of Persia*

- Knowledge
- Consequences of decisions
  - (e.g what will happen if minimum wage rises?)
- World controlling methods
  - (e.g how to improve demographic situation?)

- Questions
  - What influences what
  - Influence strength

# Causality – definitions

- Correlation & time sequence
- Time&space processes
- Bradford-Hill criteria (medicine)
- *Human agency*
- **Manipulating**
- Counterfactuals
- Invariance
- **Probabilistic approach**
- Eclecticism
- … (e.g  Aristotle, Kant, …, Cartwright, …)

# Confounding

| Treatment | # patients | # live | # die |
|---|---|---|---|
| expensive | 100 | 46 | 54 |
| none | 100 | 80 | 20 |

| Patient status | # patients | # live | # die |
|---|---|---|---|
| Expensive treatment | | | |
| Advanced | 90 | 36 | 54 |
| Initial | 10 | 10 | 0 |
| No treatment | | | |
| Advanced | 20 | 0 | 20 |
| Initial | 80 | 80 | 0 |

# Applications (University of California, Berkeley)

- Are women being discriminated?

| | # applications | % hired |
|---|---|---|
| male | 8442 | 44% |
| female | 4321 | 35% |

- But…

| Department | # male applications | % hired men | # female applications | % hired women |
|---|---|---|---|---|
| A | 825 | 62% | 108 | 82% |
| B | 560 | 63% | 25 | 68% |
| C | 325 | 37% | 593 | 34% |
| D | 417 | 33% | 375 | 35% |
| E | 191 | 28% | 393 | 24% |
| F | 272 | 6% | 341 | 7% |

- Simpson paradox

# Simpson Paradox

- Two batting (baseball) averages

| Batter\Year | 1995 | | 1996 | | Average | |
|---|---|---|---|---|---|---|
| Derek Jeter | 12/48 | .250 | 183/582 | .314 | 195/630 | **.310** |
| David Justice | 104/411 | **.253** | 45/140 | **.321** | 149/551 | .270 |

- Justice, despite having better season averages, had worse career one!

# Our problem

- We have observational data
- We estimate model
  - frequencies, econometrics, DM, …
- Yet we want to do interventions!
  - We believe, that causal relationships are stronger

- Solutions
  - Stratified analysis (e.g CMH test over $\chi^2$)
  - Additional variables in econometric model (DM)
  - Models of *propensity score* type
  - …

- This lecture – how causal thinking can improve process of analysis development

# Graphical representation

# Used notation

- We observe three binary variables: X, Y, Z
- Simplifications
  - We ignore estimation uncertainty while using frequency analysis
  - We assume >0 probability of any combination

| x | y | z | Pr(X=x,Y=y,Z=z) |
|---|---|---|---|
| 0 | 0 | 0 | 5,6% |
| 0 | 0 | 1 | 22,4% |
| 0 | 1 | 0 | 8,4% |
| 0 | 1 | 1 | 33,6% |
| 1 | 0 | 0 | 2,4% |
| 1 | 0 | 1 | 9,6% |
| 1 | 1 | 0 | 3,6% |
| 1 | 1 | 1 | 14,4% |

# Previous table - shortened

- $P(X=1)=30\%$, $P(Y=1)=60\%$, $P(Z=1)=80\%$; independent and conditionally independent variables
- For example:
  - $X \perp\!\!\!\perp Y$
    - $P(X=1 \mid Y=1) = P(X=1), \ldots$
  - $X \perp\!\!\!\perp Y,Z$
    - $P(X=1 \mid Y=1, Z=1) = P(X=1), \ldots$
  - And $X \perp\!\!\!\perp Y \mid Z$
    - $P(X=1 \mid Y=1, Z=1) = P(X=1 \mid Z=1), \ldots$
  - (let's check if those hold for previous table)
- Such reduction may not be possible!

# Another table

- Is $X \perp\!\!\!\perp Y$?
- Is $X \perp\!\!\!\perp Z$?
- Or is $X \perp\!\!\!\perp Y \mid Z$?

| x | y | z | Pr(X=x,Y=y,Z=z) |
|---|---|---|---|
| 0 | 0 | 0 | 9% |
| 0 | 0 | 1 | 36% |
| 0 | 1 | 0 | 12% |
| 0 | 1 | 1 | 3% |
| 1 | 0 | 0 | 24% |
| 1 | 0 | 1 | 6% |
| 1 | 1 | 0 | 2% |
| 1 | 1 | 1 | 8% |

# Chain rule in probability calculus

- $P(X=x, Y=y, Z=z) =$
  $$= P(X=x) \cdot P(Y=y \mid X=x) \cdot P(Z=z \mid X=x, Y=y)$$
- If any independencies (conditional) apply, equaton may be simplified

- E.g we have $X \perp\!\!\!\perp Y$, hence
- $P(X=x, Y=y, Z=z) =$
  $= P(X=x) \cdot P(Y=y \mid X=x) \cdot P(Z=z \mid X=x, Y=y)$
  $= P(X=x) \cdot P(Y=y) \cdot P(Z=z \mid X=x, Y=y)$
- Further reduction is impossible, since below don't hold:
  - $Z \perp\!\!\!\perp X,Y$
  - $Z \perp\!\!\!\perp X \mid Y$
  - $Z \perp\!\!\!\perp Y \mid X$

# Final table

- Data generating process (DGP):
  - X – cut while shaving (10%)
  - Y – coffee burn during breakfast (20%)
  - Z – bad mood during day
    - 90%, if an accident in the morning;
    - 5% otherwise

| x | y | z | Pr(X=x,Y=y,Z=z) |
|---|---|---|---|
| 0 | 0 | 0 | 68,4% |
| 0 | 0 | 1 | 3,6% |
| 0 | 1 | 0 | 0,9% |
| 0 | 1 | 1 | 17,1% |
| 1 | 0 | 0 | 0,4% |
| 1 | 0 | 1 | 7,6% |
| 1 | 1 | 0 | 0,1% |
| 1 | 1 | 1 | 1,9% |

# Chain rule application

- $P(X, Y, Z) = P(X) \cdot P(Y) \cdot P(Z \mid X,Y)$
- Graphically equivalent to:

$$X \rightarrow Z \leftarrow Y$$

- Yet we can analyze variables in different order, e.g $P(Z, Y, X)$
- Then we obtain

$$P(Z, Y, X) = P(Z) \cdot P(Y \mid Z) \cdot P(X \mid Y,Z)$$

- And graphically

$$\begin{tikzcd} Z \arrow[r] \arrow[rr, bend left] & Y \arrow[r] & X \end{tikzcd}$$

# Which image is proper?

- How to present interventions on such graphs?
  - What an intervention really is?
  - Typical interventions?
- Now we assume, that graph structure is known
  - During next lecture we will try to retreive it from data

# Graphical notation (Bayes Networks)

- Notation
  - Vertices are variables (some of them latent)
  - Edges are mechanisms (lack of arrow is also important)
  - **Intervention removes arrows coming in**
  - Chain of arrows (not necessarilly →→→) is path
  - Some paths are not causal (e.g →←→→)
  - Some paths transport information (end features dependant)
- We have
  - X and Y are dependant

**X** $\longleftarrow$ **Third cause** $\longrightarrow$ **Y**

  - Intervention on X does not affect Y

**Intervention** $\longrightarrow$ **X** **Third cause** $\longrightarrow$ **Y**

# *Confounding* graphically

# Our problems in terms of graphs

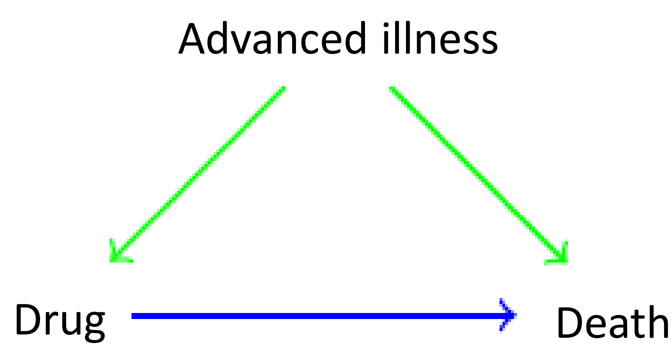
- Divorces ← time → Margarine consumption
- Patients state→ Treatment→ Result
- Smoking ← Genes → Lung cancer
  - (according to industry, yellow arrow does not exist)
- Gender → Department → Result
  - (Discrimination is yellow arrow)
  - (How to consider abilities on the graph?)

# What we want to be able to do?

- Assuming known graph structure
- Estimate causal impact strength
- (strength of relationships between variables along causal paths →…→)
- Taking under consideration, that dependencies n observational data may result from non-causal paths as well (e.g ←…←, ←……→)

# Graphical solution

- During analysis of observational data we have to block non-causal paths (e.g stratified analysis, etc.)
  - In order to get causal interpretation of analysis results
- In our examples:
  - Zero in regression model
  - Expensive drug ↗ % survival
- Why does randomization in an experiment helps?



# Attention (Berkson paradox)

- Sometimes introduction of a variable may open path!

Cut while shaving $\longrightarrow$ Bad day $\longleftarrow$ Coffee burned

- Another example

| | handsome | no |
|---|---|---|
| intelligent | 25% | 25% |
| no | 25% | 25% |

25% = 50% $\times$ 50%

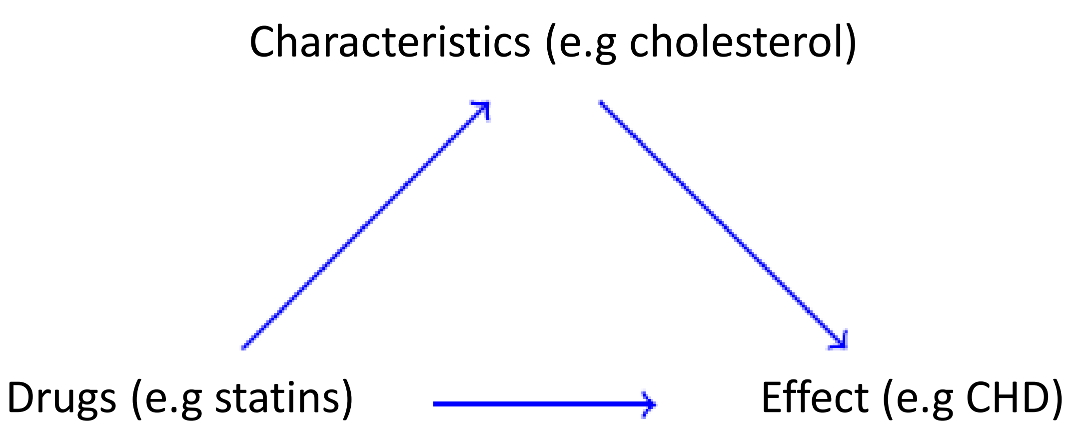
| | handsome | no |
|---|---|---|
| intelligent | 33,(3)% | 33,(3)% |
| no | 33,(3)% | (no recall) |

1/3 < 2/3 $\times$ 2/3

- For proofs and assmptions, see Pearl

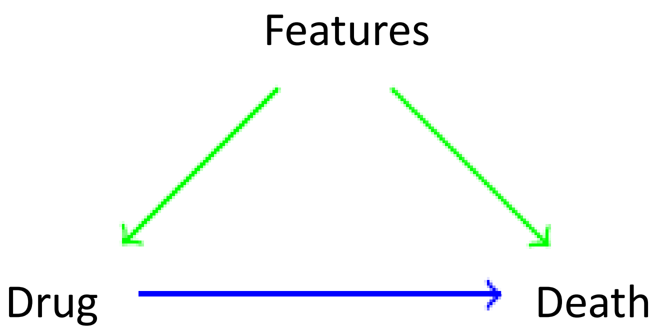
# Too many controlled variables

- We block one of causal paths
  - (we get direct causal effect)
  - (if exists)
- In terms of correlation this example is identical to the previous one!
  - It is not possible to define confounding in the field of statistics

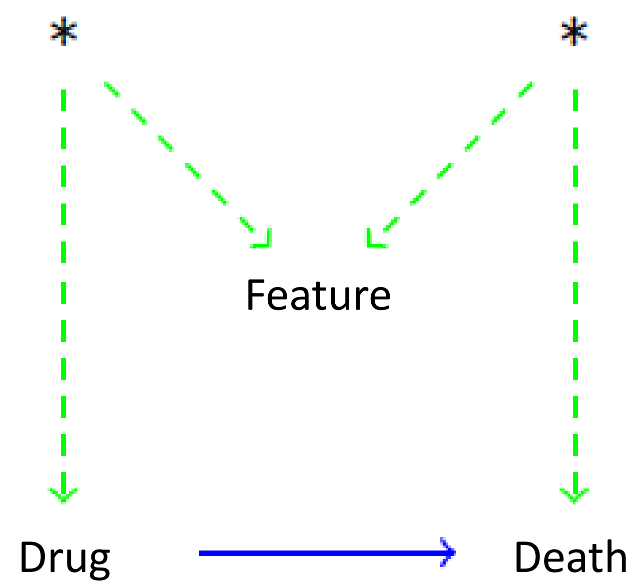


# Demographic features controlling

When YES

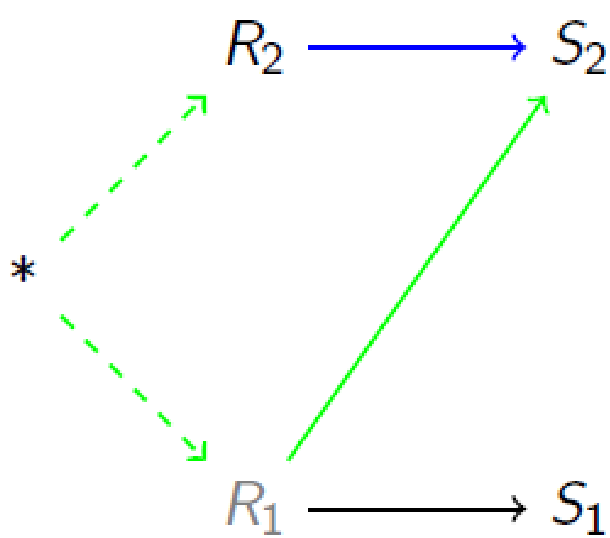



When NO

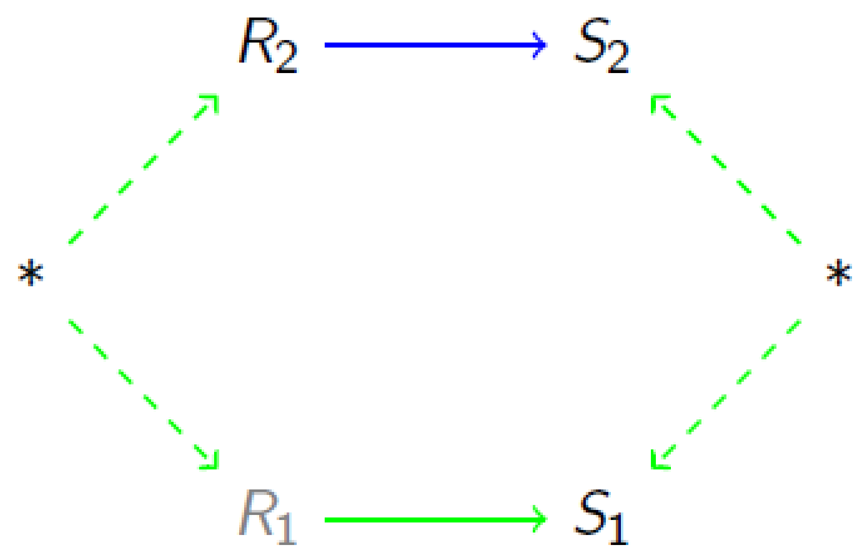


# Medical documentation controling

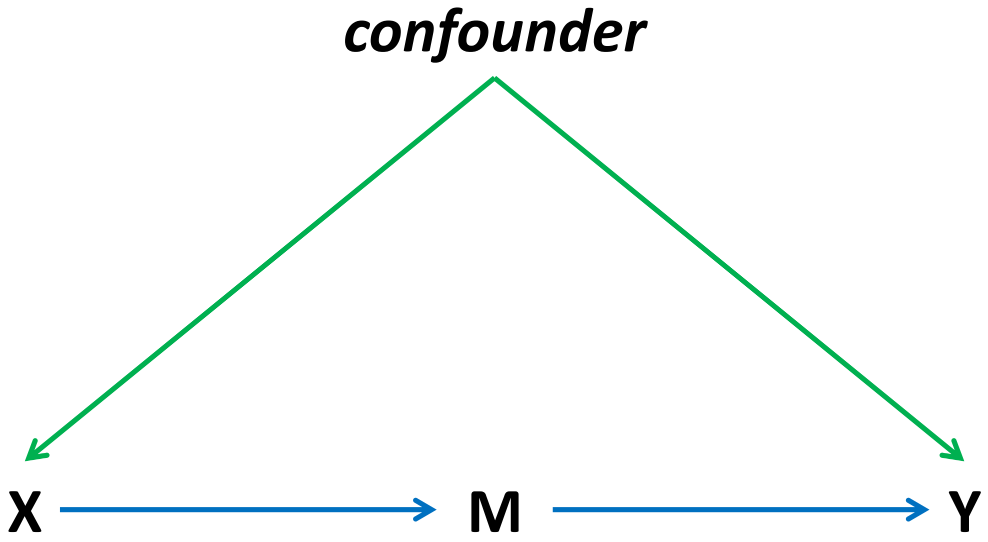
When YES



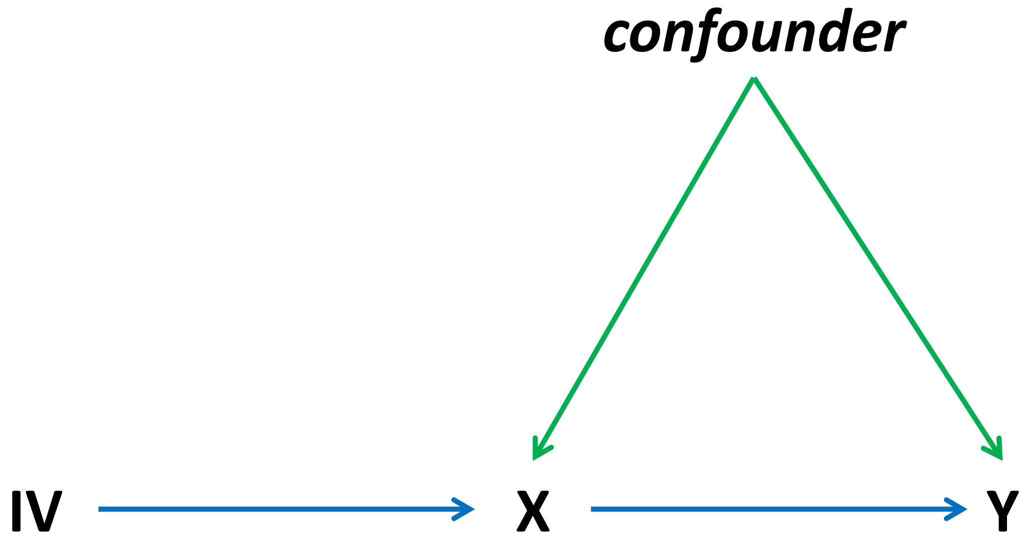

When NO



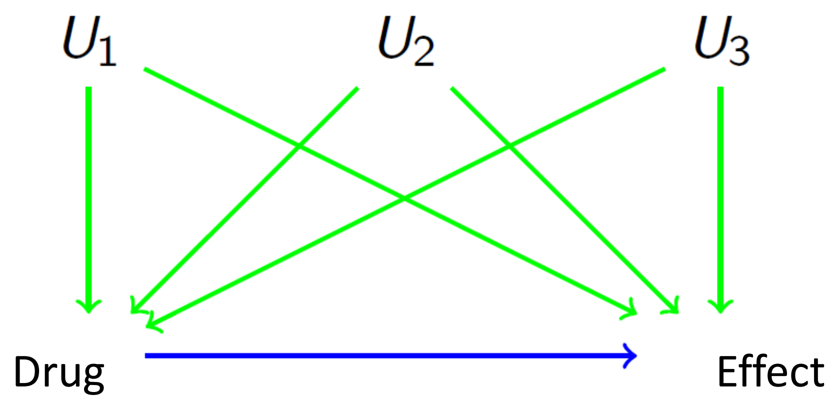
# Difficult situations (and smoking)

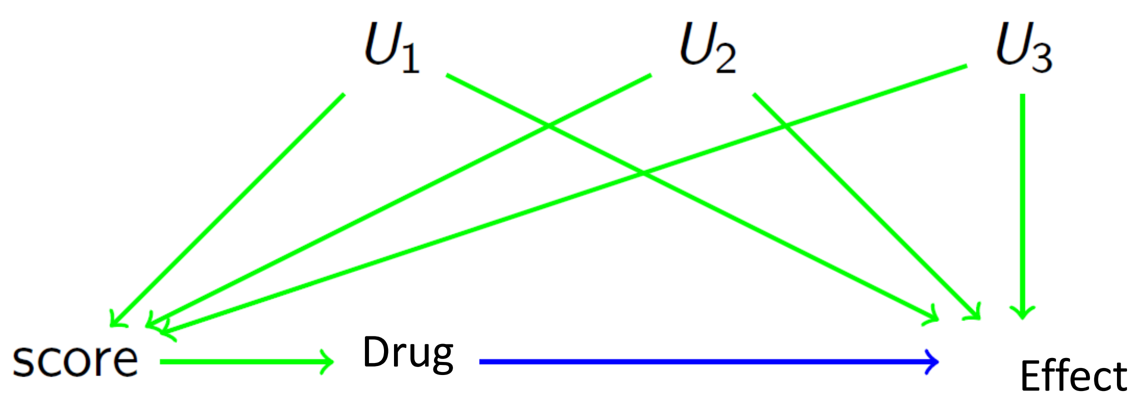


# Paths vs instumental variables



# *Propensity score method*



# Appendix

# Independence in samples

- We assumed that probability distribution is known
- We have samples and have to deal with random errors
- Tests exist
  - Independence, e.g $\chi^2$, LIS-test
  - Conditional independence, e.g CMH, KCI-test

# Counterfactuals

- Three levels of inference:
  - observational
    - „patient takes drug A, will he recover?”
  - interventional
    - „what will happen if I will prescribe patient drug A?”
  - counterfactual
    - „patient took drug A and recovered, what would happen if took drug B?”
    - Recovery resulting from A / independently of A / despite A
- Are counterfactuals necessary?
  - Repeatable decisions
  - What would happen without treatment? Question linked to another one regarding % of patients who will die when treated

# Counterfactual inference dilemma. Neyman-Rubin model

- Patients have been randomly assigned to treatment (X=1) or not (X=0)
- Recovery has been observed (Y=1) or not (Y=0)

| | Y=0 | Y=1 |
|---|---|---|
| X=0 | 15% | 10% |
| X=1 | 30% | 45% |

- For patient X=0, Y=1, what would happen if X=1?
- Extreme cases

| Y for X=0 | Y for X=1 | % patients |
|---|---|---|
| 0 | 0 | 40% |
| 0 | 1 | 20% |
| 1 | 1 | 40% |

| Y for X=0 | Y for X=1 | % patients |
|---|---|---|
| 1 | 0 | 40% |
| 0 | 1 | 60% |

# Summary

- Observational data (non-experimental) ≠ Interventional data
  - e.g forgotten OLS assumption (not random X)!
- Statistics alone is unable to conduct causal inference
  - How to say ‚puddle does not cause rain'?
- Black-box approach is not sufficient
  (not always, the more variables the better!)
- Causal inference is important!

# References


- D. Colombo et al. „Learning High-Dimensional Directed Acyclic Graphs with Latent and Selection Variables”, *The Annals of Statistics*, 2012
- J. Pearl, *Causality. Models, Reasoning, and Inference*, 2009
- J. Pearl, Causal inference in statistics: An overview. *Statistics Surveys*, 2009
- Woodward J. *Making Things Happen. A Theory of Causal Explanation*. 2003

# Please think if below is true?

- When a country's debt rises above 90% of GDP, growth slows.
- Therefore, high debt causes slow growth.

# Please think if below is true?

- As ice cream sales increase, the rate of drowning deaths increases sharply.
- Therefore, ice cream consumption causes drowning.

# Thank you!