Lecture 2: The Quest for Causality

ECON 480 - Econometrics - Fall 2018

Ryan Safner

August 29, 2018

The Two Big Problems with Data

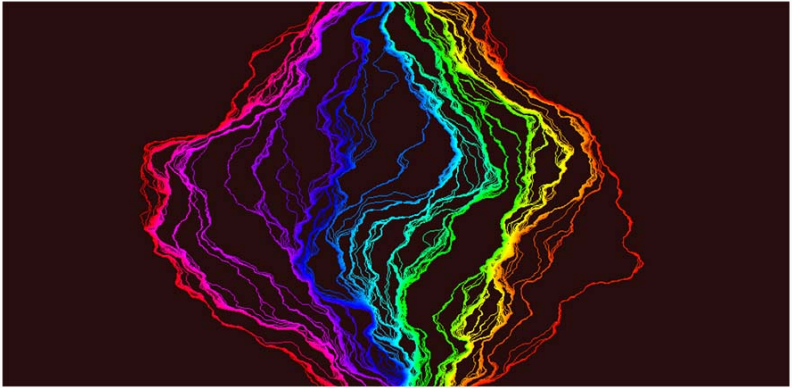
Identifying Causal Effects: Randomized Controlled Trials

Some Theory by Example

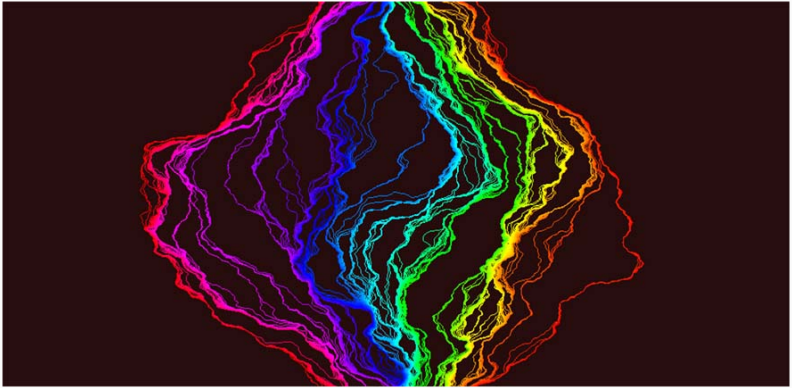
Natural Experiments

The Two Big Problems with Data

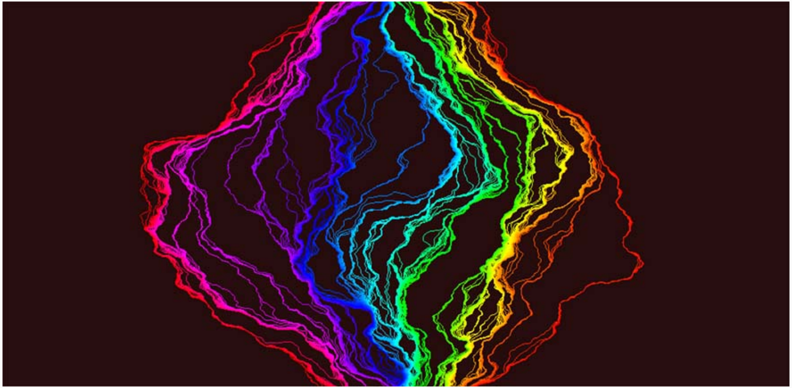
- We want to use econometrics to **identify** causal relationships and make **inferences** about them



- We want to use econometrics to **identify** causal relationships and make **inferences** about them
    1. Problem for **identification**: endogeneity



- We want to use econometrics to **identify** causal relationships and make **inferences** about them
    1. Problem for **identification**: endogeneity
    2. Problem for **inference**: randomness

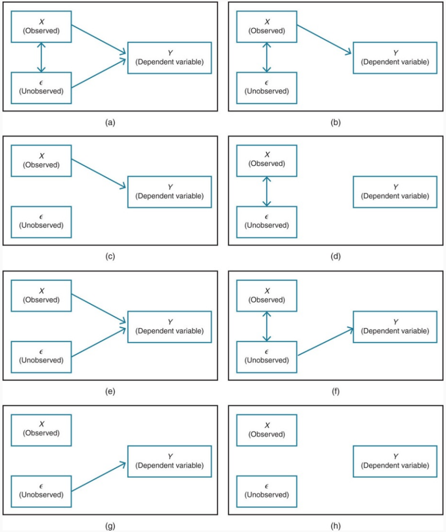


- An independent variable (x) is exogenous if its variation is *unrelated* to other factors that affect the dependent variable (y)

Identification Problem: Endogeneity

- An independent variable (x) is exogenous if its variation is *unrelated* to other factors that affect the dependent variable (y)
- An independent variable (x) is endogenous if its variation is *related* to other factors that affect the dependent variable (y)

Identification Problem: Endogeneity II



Bailey (2017: 12)

- Data is random due to **natural sampling variation**

Inference Problem: Randomness

- Data is random due to **natural sampling variation**
    - Taking one sample of a population will yield slightly different information than another sample of the same population

Inference Problem: Randomness

- Data is random due to **natural sampling variation**
    - Taking one sample of a population will yield slightly different information than another sample of the same population
- Common in statistics, *easy to fix*

Inference Problem: Randomness

- Data is random due to **natural sampling variation**
    - Taking one sample of a population will yield slightly different information than another sample of the same population
- Common in statistics, *easy to fix*
- **Inferential Statistics**: making claims about a wider population using sample data

Inference Problem: Randomness

- Data is random due to **natural sampling variation**
  - Taking one sample of a population will yield slightly different information than another sample of the same population
- Common in statistics, *easy to fix*
- **Inferential Statistics**: making claims about a wider population using sample data
  - We use common tools and techniques to deal with randomness

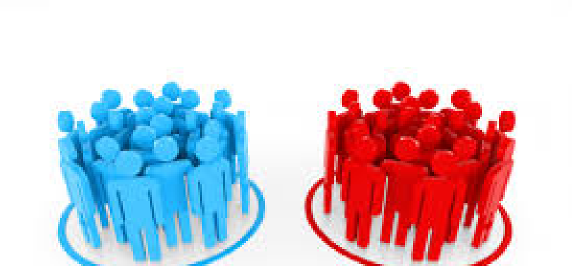
Identifying Causal Effects: Randomized Controlled Trials

- The *ideal* way to demonstrate causation is through a randomized controlled trial (RCT) or "random experiment"

- The *ideal* way to demonstrate causation is through a randomized controlled trial (RCT) or "random experiment"
  - *Randomly* assign experimental units (e.g. people, firms, etc.) into groups

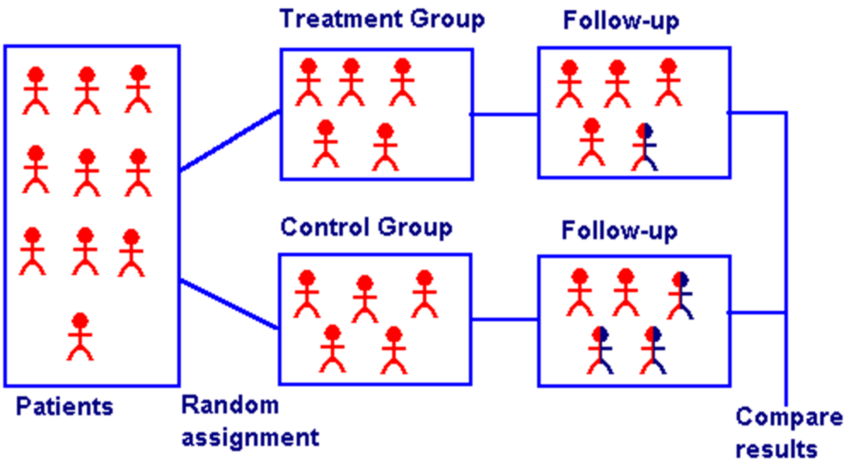
- The *ideal* way to demonstrate causation is through a randomized controlled trial (RCT) or "random experiment"
    - *Randomly* assign experimental units (e.g. people, firms, etc.) into groups
    - **Treatment group(s)** get a (level/kind of) treatment

- The *ideal* way to demonstrate causation is through a randomized controlled trial (RCT) or "random experiment"
  - *Randomly* assign experimental units (e.g. people, firms, etc.) into groups
  - **Treatment group(s)** get a (level/kind of) treatment
  - **Control group** gets no treatment



- The *ideal* way to demonstrate causation is through a randomized controlled trial (RCT) or "random experiment"
  - *Randomly* assign experimental units (e.g. people, firms, etc.) into groups
  - **Treatment group(s)** get a (level/kind of) treatment
  - **Control group** gets no treatment
  - Compare results of treatment and control groups to observe the average treatment effect (ATE)

- The *ideal* way to demonstrate causation is through a randomized controlled trial (RCT) or "random experiment"
  - *Randomly* assign experimental units (e.g. people, firms, etc.) into groups
  - **Treatment group(s)** get a (level/kind of) treatment
  - **Control group** gets no treatment
  - Compare results of treatment and control groups to observe the average treatment effect (ATE)
- **We will understand "causality" to mean the ATE from an ideal RCT**



Classic (simplified) procedure of a randomized control trial (RCT) from medicine

- Random assignment to groups ensures that the *only* differences between members of the treatment(s) and control groups is *receiving treatment or not*

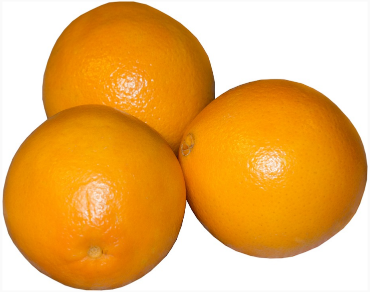
Treatment Group

Control Group

- Random assignment to groups ensures that the *only* differences between members of the treatment(s) and control groups is *receiving treatment or not*

- Random assignment to groups ensures that the *only* differences between members of the treatment(s) and control groups is *receiving treatment or not*
- Selection bias: (pre-existing) differences between members of treatment and control groups *other* than treatment, that affect the outcome

Treatment Group



Control Group

Selection Bias

Some Theory by Example

Example

What is the impact of having health insurance on health outcomes?

Example: The Effect of Having Health Insurance

Example

What is the impact of having health insurance on health outcomes?

- National Health Interview Survey (NHIS) asks “Would you say your health in general is excellent, very good, good, fair, or poor?”

Example

What is the impact of having health insurance on health outcomes?

- National Health Interview Survey (NHIS) asks "Would you say your health in general is excellent, very good, good, fair, or poor?"
  - Outcome variable (Y): Index of health (1-poor to 5-excellent) in a sample of married NHIS respondents in 2009 who may or may not have health insurance

Example
What is the impact of having health insurance on health outcomes?

- National Health Interview Survey (NHIS) asks "Would you say your health in general is excellent, very good, good, fair, or poor?"
  - Outcome variable (Y): Index of health (1-poor to 5-excellent) in a sample of married NHIS respondents in 2009 who may or may not have health insurance
  - Treatment (X): Having health insurance (vs. not)

Example: The Effect of Having Health Insurance II

	Husbands			Wives		
	Some HI (1)	No HI (2)	Difference (3)	Some HI (4)	No HI (5)	Difference (6)
A. Health						
Health index	4.01 [.93]	3.70 [1.01]	.31 (.03)	4.02 [.92]	3.62 [1.01]	.39 (.04)
B. Characteristics						
Nonwhite	.16	.17	−.01 (.01)	.15	.17	−.02 (.01)
Age	43.98	41.26	2.71 (.29)	42.24	39.62	2.62 (.30)
Education	14.31	11.56	2.74 (.10)	14.44	11.80	2.64 (.11)
Family size	3.50	3.98	−.47 (.05)	3.49	3.93	−.43 (.05)
Employed	.92	.85	.07 (.01)	.77	.56	.21 (.02)
Family income	106,467	45,656	60,810 (1,355)	106,212	46,385	59,828 (1,406)
Sample size	8,114	1,281		8,264	1,131	

Health and demographics of insured and uninsured couples in NHIS (Angrist & Pischke 2015)

- Y: outcome variable (health index score, 1-5)

Example: The Effect of Having Health Insurance III

- Y: outcome variable (health index score, 1-5)
- Y_i: health score of an individual i

Example: The Effect of Having Health Insurance III

- Y: outcome variable (health index score, 1-5)
- Y_i: health score of an individual i
- Individual i has a choice between two outcomes:

Example: The Effect of Having Health Insurance III

- Y: outcome variable (health index score, 1-5)
- Y_i: health score of an individual i
- Individual i has a choice between two outcomes:
  - $Y_{0,i}$: individual i with *no* health insurance

HOOD COLLEGE

Example: The Effect of Having Health Insurance III

- Y: outcome variable (health index score, 1-5)
- Y_i: health score of an individual i
- Individual i has a choice between two outcomes:
  - $Y_{0,i}$: individual i with *no* health insurance
  - $Y_{1,i}$: individual i with health insurance

Example: The Effect of Having Health Insurance III

- Y: outcome variable (health index score, 1-5)
- Y_i: health score of an individual i
- Individual i has a choice between two outcomes:
  - $Y_{0,i}$: individual i with *no* health insurance
  - $Y_{1,i}$: individual i with health insurance
- $(Y_{1,i} - Y_{0,i})$: **causal effect of having insurance for individual i**

Example: A Hypothetical Comparison

- John (Y_J)
  - $Y_{0,J} = 3$
  - $Y_{1,J} = 4$

- John (Y_J)
    - $Y_{0,J} = 3$
    - $Y_{1,J} = 4$
    - $(Y_{1,J} - Y_{0,J}) = 1$

Example: A Hypothetical Comparison

- John (Y_J)
  - $Y_{0,J} = 3$
  - $Y_{1,J} = 4$
  - $(Y_{1,J} - Y_{0,J}) = 1$
  - Will buy insurance

- John (Y_J)
  - $Y_{0,J} = 3$
  - $Y_{1,J} = 4$
  - $(Y_{1,J} - Y_{0,J}) = 1$
  - Will buy insurance
  - We observe

    $Y_J = Y_{1,J} = 4$

Example: A Hypothetical Comparison

- John (Y_J)
  - $Y_{0,J} = 3$
  - $Y_{1,J} = 4$
  - $(Y_{1,J} - Y_{0,J}) = 1$
  - Will buy insurance
  - We observe

    $Y_J = Y_{1,J} = 4$
- Maria (Y_M)
  - $Y_{0,M} = 5$
  - $Y_{1,M} = 5$

Example: A Hypothetical Comparison

- John (Y_J)
  - $Y_{0,J} = 3$
  - $Y_{1,J} = 4$
  - $(Y_{1,J} - Y_{0,J}) = 1$
  - Will buy insurance
  - We observe

    $Y_J = Y_{1,J} = 4$

- Maria (Y_M)
  - $Y_{0,M} = 5$
  - $Y_{1,M} = 5$
  - $(Y_{1,M} - Y_{0,M}) = 0$

Example: A Hypothetical Comparison

- John (Y_J)
  - $Y_{0,J} = 3$
  - $Y_{1,J} = 4$
  - $(Y_{1,J} - Y_{0,J}) = 1$
  - Will buy insurance
  - We observe

    $Y_J = Y_{1,J} = 4$

- Maria (Y_M)
  - $Y_{0,M} = 5$
  - $Y_{1,M} = 5$
  - $(Y_{1,M} - Y_{0,M}) = 0$
  - Will *not* buy insurance

Example: A Hypothetical Comparison

- John (Y_J)
  - $Y_{0,J} = 3$
  - $Y_{1,J} = 4$
  - $(Y_{1,J} - Y_{0,J}) = 1$
  - Will buy insurance
  - We observe

    $Y_J = Y_{1,J} = 4$

- Maria (Y_M)
  - $Y_{0,M} = 5$
  - $Y_{1,M} = 5$
  - $(Y_{1,M} - Y_{0,M}) = 0$
  - Will *not* buy insurance
  - We observe

    $Y_M = Y_{0,M} = 5$

Example: A Hypothetical Comparison II

- John (Y_J)
  - $Y_{0,J} = 3$
  - $Y_{1,J} = 4$
  - $(Y_{1,J} - Y_{0,J}) = 1$
  - Will buy insurance
  - We only observe
    $Y_J = Y_{1,J} = 4$
- Maria (Y_M)
  - $Y_{0,M} = 5$
  - $Y_{1,M} = 5$
  - $(Y_{1,M} - Y_{0,M}) = 5$
  - Will *not* buy insurance
  - We only observe
    $Y_M = Y_{0,M} = 5$
- *Observed* difference between John & Maria: $Y_J - Y_M = -1$

HOOD COLLEGE

- Note that what we *actually* see is:

$$Y_J - Y_M = \underbrace{Y_{1,J} - Y_{0,M}}_{=-1}$$

Example: A Hypothetical Comparison III

- Note that what we *actually* see is:

$$Y_J - Y_M = \underbrace{Y_{1,J} - Y_{0,M}}_{=-1}$$

- Algebra trick: add and subtract $Y_{0,J}$:

Example: A Hypothetical Comparison III

- Note that what we *actually* see is:

$$Y_J - Y_M = \underbrace{Y_{1,J} - Y_{0,M}}_{=-1}$$

- Algebra trick: add and subtract $Y_{0,J}$:

$$Y_J - Y_M = \underbrace{Y_{1,J} - Y_{0,J}}_{=1} + \underbrace{Y_{0,J} - Y_{0,M}}_{=-2}$$

Example: A Hypothetical Comparison III

- Note that what we *actually* see is:

$$Y_J - Y_M = \underbrace{Y_{1,J} - Y_{0,M}}_{=-1}$$

- Algebra trick: add and subtract $Y_{0,J}$:

$$Y_J - Y_M = \underbrace{Y_{1,J} - Y_{0,J}}_{=1} + \underbrace{Y_{0,J} - Y_{0,M}}_{=-2}$$

- Causal effect of buying health insurance for John

- Note that what we *actually* see is:

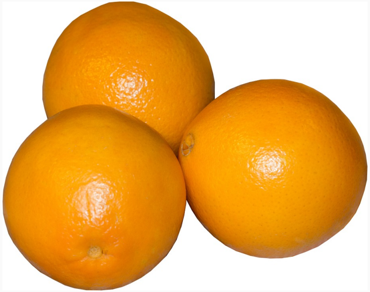
$$Y_J - Y_M = \underbrace{Y_{1,J} - Y_{0,M}}_{=-1}$$

- Algebra trick: add and subtract $Y_{0,J}$:

$$Y_J - Y_M = \underbrace{Y_{1,J} - Y_{0,J}}_{=1} + \underbrace{Y_{0,J} - Y_{0,M}}_{=-2}$$

- Causal effect of buying health insurance for John
- Difference between John & Maria before treatment: called **selection bias**

Treatment Group



Control Group

Selection Bias

Ideal (but impossible) data

i	$Y_{1,i}$	$Y_{0,i}$	$(Y_{1,i} - Y_{0,i})$
J	4	3	1
M	5	5	0
	$avg(Y_{1,i})$	$avg(Y_{0,i})$	$avg(Y_{1,i}) - avg(Y_{0,i})$
	4.5	4	0.5

Example: Thinking About the Data

Ideal (but impossible) data

i	$Y_{1,i}$	$Y_{0,i}$	$(Y_{1,i} - Y_{0,i})$
J	4	3	1
M	5	5	0
	$avg(Y_{1,i})$	$avg(Y_{0,i})$	$avg(Y_{1,i}) - avg(Y_{0,i})$
	4.5	4	0.5

Actually observed data

i	$Y_{1,i}$	$Y_{0,i}$	$(Y_{1,i} - Y_{0,i})$
J	4	?	?
M	?	5	?
	$avg(Y_{1,i})$	$avg(Y_{0,i})$	$avg(Y_{1,i}) - avg(Y_{0,i})$
	?	?	?

HOOD COLLEGE

- For entire groups (of size n), the average treatment effect (ATE) is:

$$ATE = \frac{1}{n}\sum_{i=1}^{n}[Y_{1,i} - Y_{0,i}]$$

$$ATE = \frac{1}{n}\sum_{i=1}^{n}Y_{1,i} - \frac{1}{n}\sum_{i=1}^{n}Y_{0,i}$$

HOOD COLLEGE

- For entire groups (of size n), the average treatment effect (ATE) is:

$$ATE = \frac{1}{n}\sum_{i=1}^{n}[Y_{1,i} - Y_{0,i}]$$

$$ATE = \frac{1}{n}\sum_{i=1}^{n}Y_{1,i} - \frac{1}{n}\sum_{i=1}^{n}Y_{0,i}$$

- Taking all of the individual treatment effects and averaging them over n individuals

Example: Thinking About the Data II

- For entire groups (of size n), the average treatment effect (ATE) is:

$$ATE = \frac{1}{n}\sum_{i=1}^{n}[Y_{1,i} - Y_{0,i}]$$

$$ATE = \frac{1}{n}\sum_{i=1}^{n} Y_{1,i} - \frac{1}{n}\sum_{i=1}^{n} Y_{0,i}$$

- Taking all of the individual treatment effects and averaging them over n individuals
  - i.e. $Y_{1,J} - Y_{0,J}$ and $Y_{1,M} - Y_{0,M}$

Example: Thinking About the Data III

- Basic comparisons (such as from the table) tell us *something* about outcomes, but *not* ATE!

$$\text{Difference in group outcomes} = \text{ATE} + \text{Selection bias}$$

- Basic comparisons (such as from the table) tell us *something* about outcomes, but *not* ATE!
    - We only see insured's $Y_{1,i}$ and uninsured's $Y_{0,i}$ outcomes

$$\text{Difference in group outcomes} = \text{ATE} + \text{Selection bias}$$

Example: Thinking About the Data III

- Basic comparisons (such as from the table) tell us *something* about outcomes, but *not* ATE!
  - We only see insured's $Y_{1,i}$ and uninsured's $Y_{0,i}$ outcomes
- Difference in average health outcomes by treatment (insurance) status is:

$$\text{Difference in group outcomes} = \text{ATE} + \text{Selection bias}$$

- Basic comparisons (such as from the table) tell us *something* about outcomes, but *not* ATE!
  - We only see insured's $Y_{1,i}$ and uninsured's $Y_{0,i}$ outcomes
- Difference in average health outcomes by treatment (insurance) status is:

$$\text{Difference in group outcomes} = \text{ATE} + \text{Selection bias}$$

- **Selection bias**: difference in average $Y_{0,i}$ between groups *before* any treatments

Example: Thinking About the Data III

- Basic comparisons (such as from the table) tell us *something* about outcomes, but *not* ATE!
  - We only see insured's $Y_{1,i}$ and uninsured's $Y_{0,i}$ outcomes
- Difference in average health outcomes by treatment (insurance) status is:

$$\text{Difference in group outcomes} = \text{ATE} + \text{Selection bias}$$

- **Selection bias**: difference in average $Y_{0,i}$ between groups *before* any treatments
  - $Y_{0,i}$ includes *everything* about person i related to health but *not* treatment (insurance) status

Example: Thinking About the Data III

- Basic comparisons (such as from the table) tell us *something* about outcomes, but *not* ATE!
  - We only see insured's $Y_{1,i}$ and uninsured's $Y_{0,i}$ outcomes
- Difference in average health outcomes by treatment (insurance) status is:

$$\text{Difference in group outcomes} = \text{ATE} + \text{Selection bias}$$

- **Selection bias**: difference in average $Y_{0,i}$ between groups *before* any treatments
  - $Y_{0,i}$ includes *everything* about person i related to health but *not* treatment (insurance) status
  - Imagine a world where *nobody* has insurance, who would be healthiest?

HOOD COLLEGE

- If treatment is randomly assigned for a large sample, it eliminates selection bias!

- If treatment is randomly assigned for a large sample, it eliminates selection bias!
    - Treatment & control groups differ *only* by members being treated or not treated

- If treatment is randomly assigned for a large sample, it eliminates selection bias!
  - Treatment & control groups differ *only* by members being treated or not treated
  - Creates *ceterus paribus* conditions in economics: groups are otherwise identical (same education, family size, age, etc...) *on average*

Treatment Group

Control Group

Natural Experiments

- Society is not our laboratory (probably a good thing)

- Society is not our laboratory (probably a good thing)
- We rarely can collect experimental data

- Instead, we often rely on observational data

- Instead, we often rely on observational data
    - This data is *not random*!

The Quest for Causal Effects II

- Instead, we often rely on observational data
  - This data is *not random*!
  - Must take extra care in forming an **identification strategy**

- Instead, we often rely on observational data
    - This data is *not random*!
    - Must take extra care in forming an **identification strategy**
- To make good claims about causation in society, we must get clever!

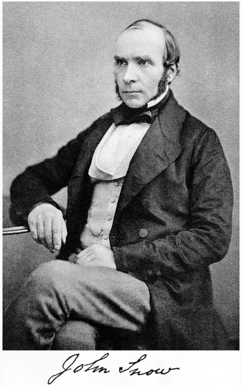
- Economists often resort to searching for natural experiments

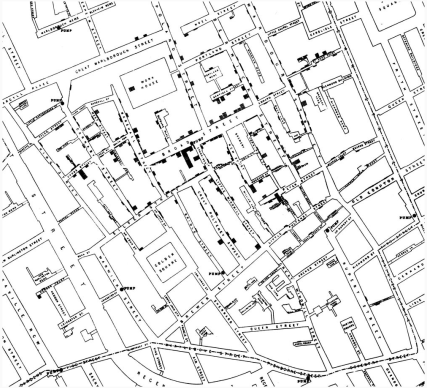
- Economists often resort to searching for natural experiments
  - Some events beyond our control occur that separate *otherwise similar* entities into a "treatment" group and a "control" group that we can compare

- Economists often resort to searching for natural experiments
  - Some events beyond our control occur that separate *otherwise similar* entities into a "treatment" group and a "control" group that we can compare
    - e.g. natural disasters, U.S. State laws, military draft

Jon Snow utilized the first famous natural experiment to establish the foundations of epidemiology and the germ theory of disease

The 'First' Natural Experiment II





*John Snow (1855) utilized the first famous natural experiment to establish the foundations of epidemiology and the germ theory of disease: water pumps *downstream* of a sewage dump in the Thames river spread cholera while water pumps *upstream* did not - Read More

- **Oregon Health Insurance Experiment**: Oregon used lottery to grant Medicare access to 10,000 people, showing access to Medicaid increased use of health services, lowered debt, etc. relative to those not on Medicaid

Famous Natural Experiments

- **Oregon Health Insurance Experiment**: Oregon used lottery to grant Medicare access to 10,000 people, showing access to Medicaid increased use of health services, lowered debt, etc. relative to those not on Medicaid
- **Angrist (1990)** finds that lifetime earnings of (random) drafted Vietnam veterans is 15% lower than non-veterans

Famous Natural Experiments

- **Oregon Health Insurance Experiment**: Oregon used lottery to grant Medicare access to 10,000 people, showing access to Medicaid increased use of health services, lowered debt, etc. relative to those not on Medicaid
- **Angrist (1990)** finds that lifetime earnings of (random) drafted Vietnam veterans is 15% lower than non-veterans
- **Card & Kreuger (1994)** find that minimum wage hike in fast-food restaurants on NJ side of border had no disemployment effects relative to restaurants on PA side of border during the same period

HOOD COLLEGE

- **Oregon Health Insurance Experiment**: Oregon used lottery to grant Medicare access to 10,000 people, showing access to Medicaid increased use of health services, lowered debt, etc. relative to those not on Medicaid
- **Angrist (1990)** finds that lifetime earnings of (random) drafted Vietnam veterans is 15% lower than non-veterans
- **Card & Kreuger (1994)** find that minimum wage hike in fast-food restaurants on NJ side of border had no disemployment effects relative to restaurants on PA side of border during the same period
- **Acemoglu, Johnson, and Robinson (2001)** find that inclusive institutions lead to higher economic development than extractive institutions, determined by a colony's disease environment in 1500

Famous Natural Experiments

- **Oregon Health Insurance Experiment**: Oregon used lottery to grant Medicare access to 10,000 people, showing access to Medicaid increased use of health services, lowered debt, etc. relative to those not on Medicaid
- **Angrist (1990)** finds that lifetime earnings of (random) drafted Vietnam veterans is 15% lower than non-veterans
- **Card & Kreuger (1994)** find that minimum wage hike in fast-food restaurants on NJ side of border had no disemployment effects relative to restaurants on PA side of border during the same period
- **Acemoglu, Johnson, and Robinson (2001)** find that inclusive institutions lead to higher economic development than extractive institutions, determined by a colony's disease environment in 1500
- We will look at some of these in greater detail throughout the course

- **Oregon Health Insurance Experiment**: Oregon used lottery to grant Medicare access to 10,000 people, showing access to Medicaid increased use of health services, lowered debt, etc. relative to those not on Medicaid
- **Angrist (1990)** finds that lifetime earnings of (random) drafted Vietnam veterans is 15% lower than non-veterans
- **Card & Kreuger (1994)** find that minimum wage hike in fast-food restaurants on NJ side of border had no disemployment effects relative to restaurants on PA side of border during the same period
- **Acemoglu, Johnson, and Robinson (2001)** find that inclusive institutions lead to higher economic development than extractive institutions, determined by a colony's disease environment in 1500
- We will look at some of these in greater detail throughout the course
- A great list, with explanations is here