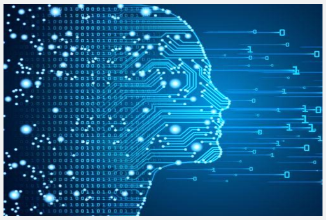


Principles of Causal Inference

Vasant G. Honavar

Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics,
Public Health Sciences and Neuroscience
Center for Artificial Intelligence Foundations and Scientific Applications
Institute for Computational and Data Sciences
Huck Institutes of the Life Sciences
Clinical and Translational Sciences Institute
Northeast Big Data Hub
Pennsylvania State University
vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Synthetic controls

- All previous methods require that we observe both *treated* and *untreated* individuals
- What if we are in a scenario where everyone is treated?

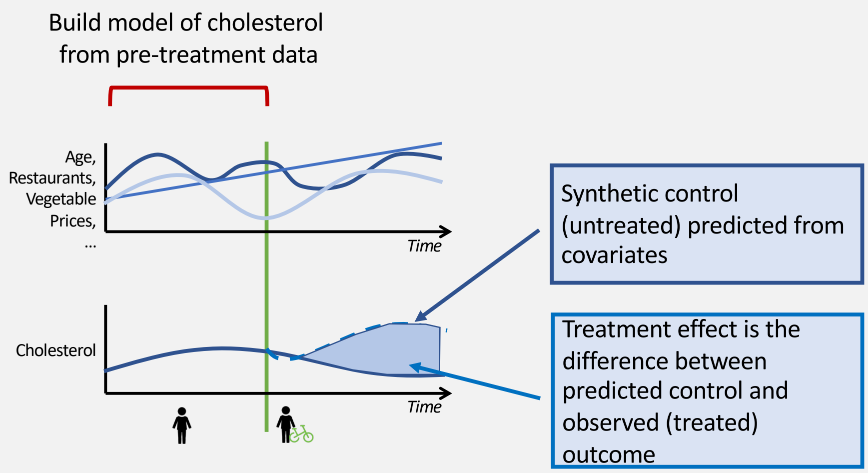
e.g., effect of a large marketing campaign, or a global policy change?

Build synthetic controls that estimate what $\bar{Y}_{T=0}$ would have been for a population were it not for treatment

Synthetic controls: Intuition

1. Decide what the treatment will be
2. Pre-treatment: Observe the world for awhile
   - Record the outcome we care
   - Record covariates that can help us predict our observed outcome, but will not be affected by the treatment. Use domain-knowledge / theory to identify these covariates.
   - Learn a model that predicts outcome based on covariates.
3. Post-treatment:
   - Keep recording outcome. This is now the treated outcome.
   - Predict untreated outcome using learned model and current covariates
   - ATE = Difference between observed outcome and prediction of untreated outcome

Example: policy change to encourage exercise



Synthetic Controls

Definition	Calculate treatment effect by comparing observed outcomes of treated population with synthetic (predicted) outcomes of an untreated population
Intuition	If we can measure covariates that are unaffected by the treatment and predictive of untreated outcomes, then we can build a synthetic control
Example	Predicting effect of global policy change to encourage exercise on population-wide cholesterol
Keep in mind	Ignorability assumption must still hold; Relatedly, be concerned about generalizability/robustness of learned outcome model

Estimating Causal Effects from Observational Data

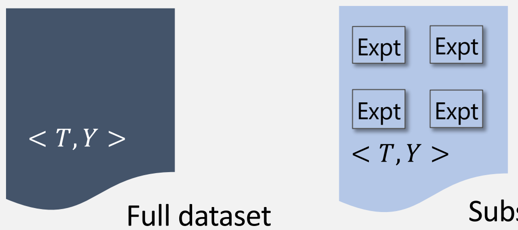
✓Simulating a randomized experiment

- Natural Experiments
  - Simple natural experiments
  - Instrumental variables

Natural experiments

- What can we do without ignorability?
  - Rather than assume ignorability over the entire data set, find data subsets that approximate a “natural” experiment

Finding a natural experiment



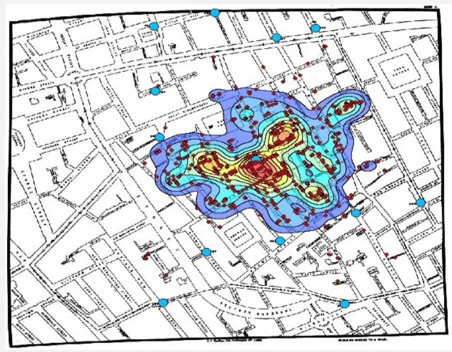
Full dataset
$y = f(t, x)$
$t = g(x)$

Subsets of the data
$y = f(t, u)$
$t = g(r)$
$\boldsymbol{r}: randomized$

How to find such experiments?

Example: Cholera cause estimation in 1850s

London cholera outbreak

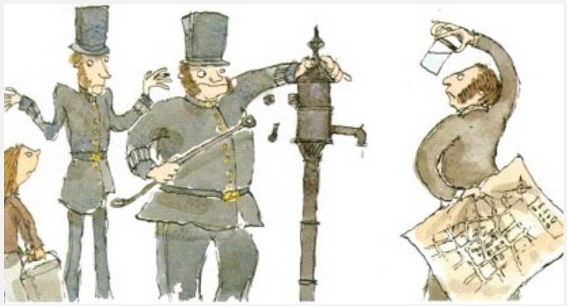


- The first cases of cholera in England were reported in1831, about the time Dr. Snow was finishing up his medical studies at the young age of eighteen.
- Between 1831 and 1854, tens of thousands of people in England died of cholera.

London cholera outbreak

- In the middle 1800s, people didn't have running water or modern toilets in their homes.
- They used town wells and communal pumps to get the water they used for drinking, cooking and washing.
- Septic systems were primitive and most homes and businesses dumped untreated sewage and animal waste directly into the Thames River or into open pits called "cesspools".
- Water companies often bottled water from the Thames and delivered it to pubs, breweries and other businesses.
- Dr. Snow hypothesized that the sewage dumped into the river or into cesspools near town wells could contaminate the water supply, leading to a rapid spread of disease.

London cholera outbreak



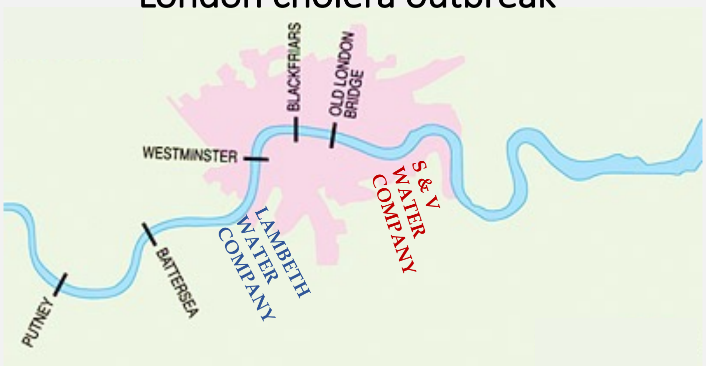
Source: Cricket 31(3), pp. 23-31, Nov. 2003

- Within 250 yards of the spot where Cambridge Street joins Broad Street there were upwards of 500 fatal cholera cases in 10 days
- John Snow noticed higher cholera deaths near a water pump, on Broad Street in London, but that could just be correlation

London cholera outbreak

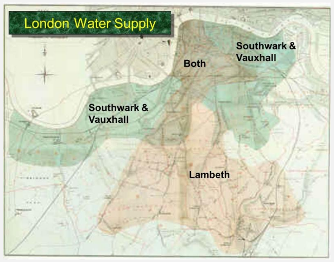
- Snow also investigated groups of people who did not get cholera and tracked down whether they drank pump water.
- A prison near Broad Street had 535 inmates but no cases of cholera. Snow discovered the prison bought water from the Grand Junction Water Works.
- The men who worked in a brewery on Broad Street also escaped cholera. They drank the liquor they made or water from the brewery's own well and not water from the Broad Street pump.
- A factory near the pump, at 37 Broad Street, wasn't so lucky. The factory relied on Broad Street pump for drinking water and 16 of the workers died from cholera.

London cholera outbreak



- Two major water companies supplied water for Londoners
  - Lambeth Water Company
  - S & V Water Company
- One was upstream and one downstream.
- Customers of each company were distributed throughout city

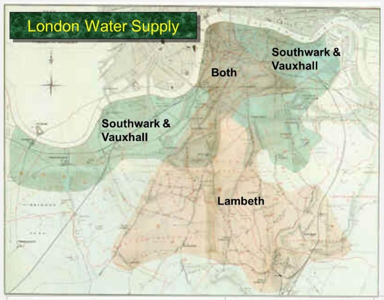
London cholera outbreak



- Southwark and Vauxhall
- Lambeth

- Each company supplies both rich and poor, large houses and small
- There was no difference either in the condition or occupation of the persons receiving the water of the different companies
- There was no difference whatever in the houses or the people receiving the supply of the two water companies; or their surroundings

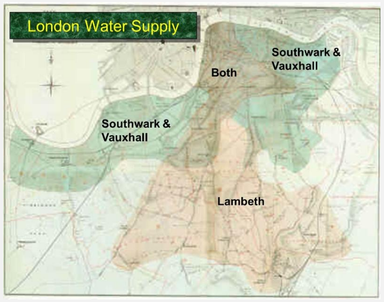
London cholera outbreak



- The death rate from cholera in the S&V houses was almost ten times the rate in the houses supplied by Lambeth
- S & V was downstream from Broad Street Pump

Supply Area	# of houses	# of cholera deaths	# of deaths per 10,000 houses
S&V	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	59

London cholera outbreak



- The numbers pointed accusingly at S&V
- S & V was downstream from Broad Street Pump
- Later it was discovered that a cesspit that was just a few feet away from the well of the Broad Street pump had been leaking into the well
- Thus the pump's water was contaminated by sewage from the houses of cholera victims
- Aha!


SO THE MIASMAL VAPORS...
I FOUND THE SOURCE OF THE CHOLERA OUTBREAK! IT WAS A WATER PUMP!
BUT EVERYONE KNOWS THE MIASMA MADE PEOPLE SICK!
YOU KNOW NOTHING, JOHN SNOW!
VEINITY FAIR

London cholera outbreak: Establishing Causality

- In terms of the terminology we have used in this course
  - People in the S&V houses as the treatment group, and
  - Those in the Lambeth houses at the control group
- A crucial element in Snow's analysis was that the people in the two groups were comparable to each other, apart from the treatment, i.e., they were exchangeable

London cholera outbreak: Establishing Causality

- In order to establish whether it was the water supply that was causing cholera, Snow had to compare two groups that were similar to each other in all but one aspect—their water supply.
- Only then would he be able to ascribe the differences in their outcomes to the water supply.
- Had the two groups been different in some other way as well, it would have been difficult to point the finger at the water supply as the source of the disease.
- For example, if the treatment group consisted of factory workers and the control group did not, then differences between the outcomes in the two groups could have been due to the water supply, or to factory work, or both

London cholera outbreak: Establishing Causality

- Snow's brilliance lay in identifying two groups that would make his comparison clear.
- He had set out to establish a causal relation between contaminated water and cholera infection, and to a great extent he succeeded, even though he was ignored and even ridiculed
  - Snow did not understand the detailed mechanism by which humans contract cholera.
  - That discovery was made in 1883, when the German scientist Robert Koch isolated the *Vibrio cholerae*, the bacterium that can enter the human small intestine through water that is consumed and causes cholera

What Dr. Snow did was to exploit a natural experiment

- A natural experiment permits causal inference by exploiting variation in natural data as if it were the result of randomization
- Because the data is not actively randomized, we have to assume that
  - There are no unobserved confounders – any source of variation between the treated and untreated groups does not directly impact the outcome although it may impact the treatment
  - Check that the treated and control groups are similar with respect to the observed covariates
- Natural experiments offer a powerful means of establishing causality and estimating causal effects from observational data especially in domains, e.g., evolutionary biology, where controlled experiments are, for all practical purposes, impossible

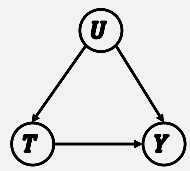
London Cholera Outbreak as a natural experiment

Example:

- London Cholera outbreak 1850s
  - T: Water supply company
  - U: Londoners' Covariates
  - Y: Cholera Status

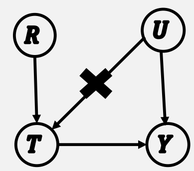
$< T, Y >$

Full dataset
$y = f(t, u)$
$t = g(u)$



Expt Expt
Expt Expt
$< T, Y >$

Subsets of the data
$y = f(t, u)$
$t = g(r)$
$\boldsymbol{r}: randomized$



What we just learned: Simple natural experiment

Definition Exploit "as-if random" assignment of treatments to measure outcome

Intuition When assignment of treatment is unrelated to the measured outcome and their common causes, we can treat it as if it is a randomized experiment to estimate treatment effect.

Example London cholera outbreak

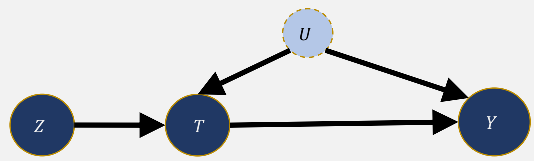
Keep in mind As-if random treatment assignments are hard to find
Estimates sensitive to violation of assumptions

Estimating Causal Effects from Observational Data

✓Simulating a randomized experiment

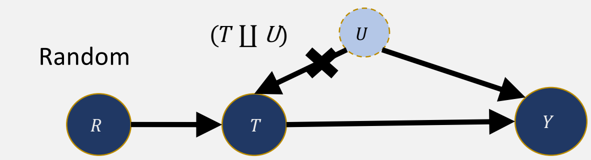
- Natural Experiments
  - Simple natural experiments
  - Instrumental variables

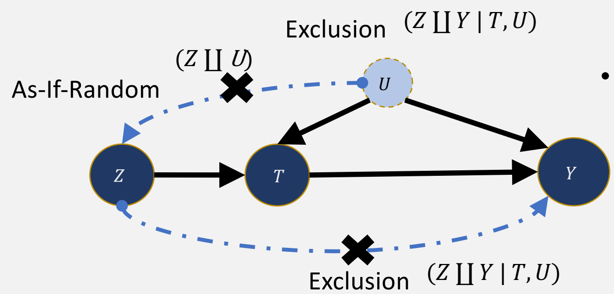
Instrumental variables generalize randomized experiments



- Suppose we want to estimate the causal effect of T on Y
- We need to block the backdoor path $T \leftarrow U \rightarrow Y$
- We can only do so only if U is measured
- If U is not measured, we cannot use stratification or Inverse propensity weighting – or if we use them, we get biased estimates
- Is there another way? – We saw one already – natural experiments

Randomized vs natural experiment

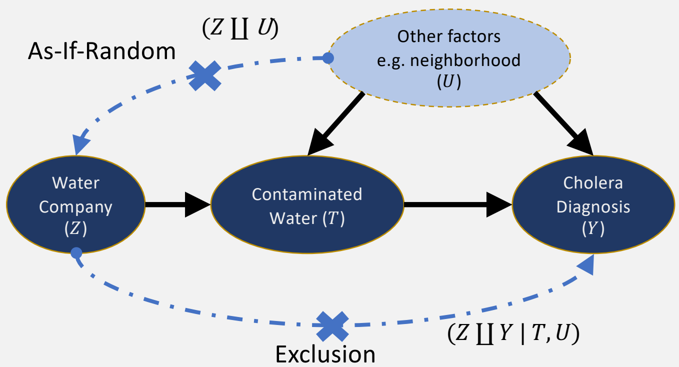




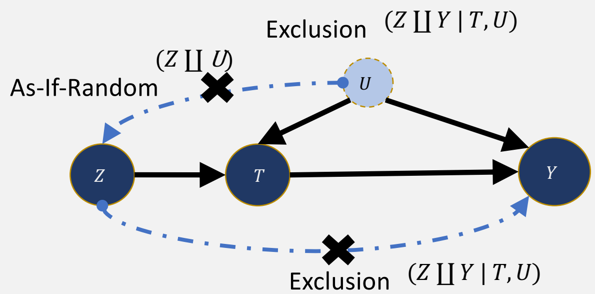
- What if we can find a variable that can take the place of R in the diagram despite not being randomized in an experiment?
- If we can do that, we've clearly got a "natural experiment"
- When we find a variable that can do that, we call it an "instrumental variable" e.g., Z

Instrumental variables

- Instrumental variables generalize simple natural experiments

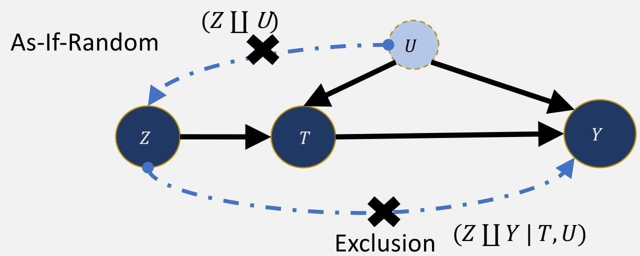


Instrumental variable Z



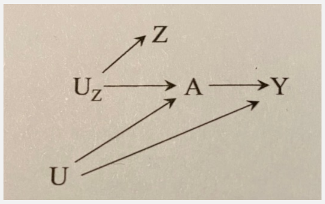
- Z must be related to T (typically $Z \rightarrow T$ but not always)
- There must be no open paths from Z to Y except for ones that go through T
- That is, Z impacts Y only through T (Z and Y do not share causes)

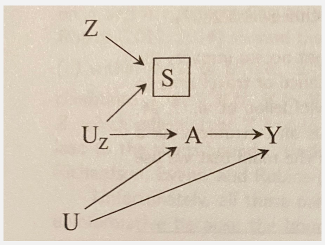
Instrumental variables



- A change in Z can lead to a change in Y *only through* T
- So change in Y is a product of change in $Z \rightarrow T$ and $T \rightarrow Y$ arrows
- Compare the extent by which random assignment affects T versus Y
- Causal effect $(T \rightarrow Y) = \frac{E(Y^{Z=1} - Y^{Z=0})}{E(T^{Z=1} - T^{Z=0})} = \frac{(E(Y \mid Z=1) - E(Y \mid Z=0))}{(E(T \mid Z=1) - E(T \mid Z=0))}$

Instrumental variables generalize natural experiments





- Z is an instrumental variable that is associated with A as a result of sharing a cause U_Z with A
- U_Z is unmeasured causal instrument, A measured surrogate instrument
- Z is a surrogate instrument in a selected population
- $Z \leftrightarrow A$ association arises from conditioning on a common effect S of the unmeasured causal instrument U_Z and the surrogate instrument Z
- Both causal and surrogate instruments can be used to estimate causal effects from observational data (with some caveats)

Instrumental variables

For Z to be an instrumental variable

1. Z must be associated with A (typically $Z \rightarrow A$ but not always)
2. There must be no open paths from Z to Y except for ones that go through A
3. That is, Z impacts Y only through A (Z and Y do not share causes)

- In observational studies, we cannot prove that Z is an instrument
  - Conditions 2 and 3 cannot be verified from observational data alone
- Causal effect $(A \rightarrow Y) = \frac{E(Y^{Z=1} - Y^{Z=0})}{E(A^{Z=1} - A^{Z=0})} = \frac{(E(Y \mid Z=1) - E(Y|Z=0))}{(E(A \mid Z=1) - E(A|Z=0))}$
- Provided an additional identifiability condition holds

Additional condition for identifiability using instruments

- Constant effect of treatment on outcome across individuals, OR
- (For dichotomous Z and A): Equality of average causal effect of A on Y in both treated and untreated: $E(Y^{a=1} - Y^{a=0} | Z = 1, A = a) = E(Y^{a=1} - Y^{a=0} | Z = 0, A = a)$ for $A = 0,1$ OR
- The average causal effect of A on Y is the same for all levels of unmeasured confounder – usually implausible when unmeasured confounders are also effect modifiers OR
- $Z \leftrightarrow A$ association on the additive scale is constant across all levels of U OR
- Monotonicity
- In practice all of these assumptions are questionable
- IV estimates can be extremely biased when any of the assumptions are violated (denominator in the causal effect estimation formula is small)
- Sensitivity analysis critical

Instrumental variables

$< T, Y >$

Examples of Instrumental Variables

- Lottery
- Weather
- Shocks
- Discontinuities
- Hard-to-find variations

Adapted from: Dunning, T. 2012. Natural Experiments in the Social Sciences, Cambridge University Press

What we just learned: Instrumental Variables

Definition Instrumental variables (IV) introduce “as-if random” noise into treatment assignment, and are used to estimate treatment effect

Intuition Because IVs are not influenced by confounds, IVs’ indirect effect on outcome Y is independent of confounds too. Because IVs do not directly influence outcome, their effect must be due to the effect of the treatment.

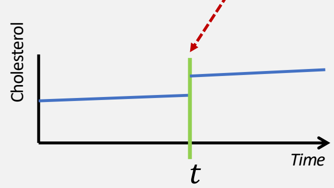
Examples Encouraging people to exercise at random

Keep in Mind Causal Estimate may not generalize to full population. Estimate very sensitive to the violations of IV assumptions.

Discontinuities as instrumental variables (IV)

- Instead of an instrumental variable changing the distribution of treatment over individuals, an arbitrary change deterministically decides the treatment assignment

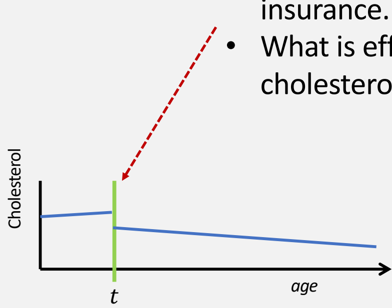
- At time t, a cholesterol drug A is banned, and people switch to another drug B.
- What was the relative effect of drug A over B?



- Due to selection effects, people taking drug A are different from those taking drug B.
- But the patients who switched from A to B at time t can be considered to be similar

Discontinuities as instrumental variables

- Above age threshold t, you get free health insurance.
- What is effect of health insurance on cholesterol?



- Due to selection effects, people with health insurance are different from those without.
- But around the age threshold t, people with or without health insurance are similar

Discontinuities as instrumental variables (IV)

- As-if-random: People near the threshold are similar to each other, as if Nature randomized them on either side of the threshold
- Exclusion: Merely being on one side of the threshold does not affect the outcome in the absence of the arbitrary change
- Very common: Many decisions in organizations, changes in policy, e.g., the rolling out of Obamacare
- Such discontinuities can be thought of as special case of instrumental variables
- We can estimate the treatment effect at the boundary

What we just learned: Discontinuities as IV

Definition Discontinuities identify arbitrary boundaries between treated and untreated populations, measure treatment effect as difference in outcomes at the boundary

Intuition Regression discontinuities approximate randomized experiments as long as no substantial differences between people just on one side or the other. That is, at the boundary, $\mathrm{T} \perp\!\!\!\perp X, U$

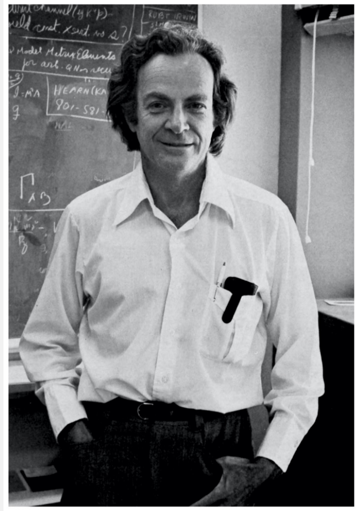
Example Policy decisions based on income or time; exogenous shocks; and are all common sources of regression discontinuities

Keep in mind Only estimates treatment effect at the boundary. Effect may vary elsewhere!

Checking your assumptions

"Science is a way of trying not to fool yourself. The first principle is that you must not fool yourself — and you are the easiest person to fool."

— Richard Feynman

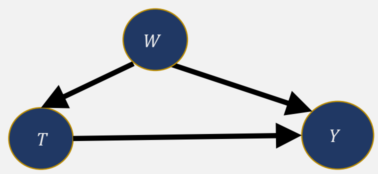


Not fooling yourself

- Recall that without causal assumptions, there are no causal conclusions
  - Identifiability assumptions
  - The assumptions allow us to specify the estimand
- The data are used simply for estimating the estimand
- Hence, it is critical to verify your assumptions – How?

Not fooling yourself

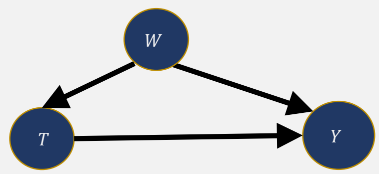
- Understand the difference between identification and estimation
- Why do observational studies yield incorrect conclusions?
  - Failure to identify the causal effect of interest
  - Estimation is a (relatively) easy statistical inference problem



- Identification: $E[Y|T,W]$
- Estimation: One of several methods for estimating $E[Y|T,W]$

Not fooling yourself

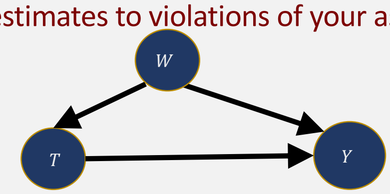
- Explicitly state your identifying and estimating assumptions



- Identifying assumptions: All of the missing covariates and arrows in the causal graph
- Estimating assumptions: Overlap between treated and untreated population – can be addressed by collecting more data

Not fooling yourself

- Refute your assumptions, and analyze the sensitivity of your causal effect estimates to violations of your assumptions



- Identifying assumption: All the arrows missing in the causal graphical model
  - e.g. No other common cause exists
  - Untestable in general
- Sensitivity analysis
  - What happens when another common cause exists?
  - What happens when treatment is placebo?

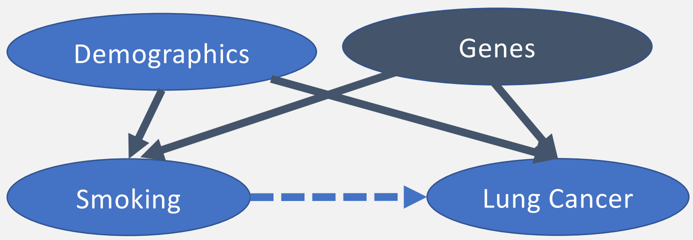
Not fooling yourself

- Add random covariates to your causal model
  - Does your causal effect estimate change?
  - If it does, your causal model is almost certainly wrong
- Replace the treatment by a placebo – one that should not have a causal effect or randomize or permute the treatment
  - Does your causal effect disappear?
  - If it does not, your causal model is almost certainly wrong
- Cross-validation: Divide your data into subsets
  - Does your causal effect estimate vary significantly across the subsets?
  - If it does, your estimate probably cannot be trusted
- Test balance of covariates – matching, weighting, discontinuity-based methods depend on covariate balance which is testable

Not fooling yourself

- Question: How sensitive is your estimate to minor violations of assumptions?
  - E.g. How big should the effect of a confounder be so that your estimated causal effect changes its direction (sign)?
- Use simulation to add effect of unknown confounders
- Compare the estimates with and without unknown confounders

Not fooling yourself – Sensitivity analysis



Cornwell (1959) showed that the effect of Genes had to be 8 times that of any known confounder for the causal effect of smoking on lung cancer to disappear

Useful tool

- DoWhy – a Python library for causal inference
- https://microsoft.github.io/dowhy/index.html
  - Modeling assumptions explicitly using a causal graph
  - Identifying a causal effect using do calculus
  - Estimating causal effects using some basic estimation methods
  - Refuting causal assumptions (by verifying their testable implications, whenever possible) or analyzing sensitivity of violations

Using DoWhy

```
from dowhy.do_why import CausalModel

# Create a causal model from the data and given graph.
model=CausalModel(
        data = df,
        treatment=data["treatment_name"],
        outcome=data["outcome_name"],
        graph=data["dot_graph"],
        )

# Identify causal effect and return target estimands
identified_estimand = model.identify_effect()

# Estimate the target estimand using a statistical method.
estimate = model.estimate_effect(identified_estimand,
        method_name="backdoor.propensity_score_matching")

# Refute the obtained estimate using multiple robustness checks.
refute_results=model.refute_estimate(identified_estimand, estimate,
        method_names=["random_common_cause", "placebo_treatment_refuter",
                      "data_subset_refuter"])
```

Learning causal models from observational data

- Constraint-based
- Bayesian
- Other

The Constraint-Based Learning of Causal Models

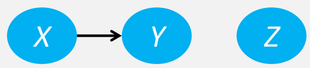
- Determine constraints that hold among the nodes (e.g., independence conditions based on statistical tests)
- Use the patterns of constraints to narrow the causal possibilities

Constraint-Based Search for a Causal Model: Example

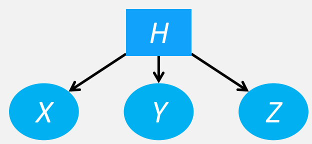
- Three binary variables *X*, *Y*, *Z*
- Suppose time ordering is known (we can relax this condition):
  X occurs before *Y* which occurs before *Z*
- For instance
  - *X*: economic circumstances
  - *Y*: environmental risk
  - *Z*: disease
- Question: Does *Y* cause *Z*?

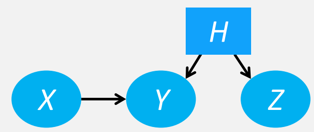
Constraint-Based Search for a Causal Model: Example

- Suppose statistical testing yields the following constraints
  dep(X, Y), dep(Y, Z), dep(X, Z), ind(X, Z | Y)
- Consider the consistency of these constraints with respect to the following causal models:

X 

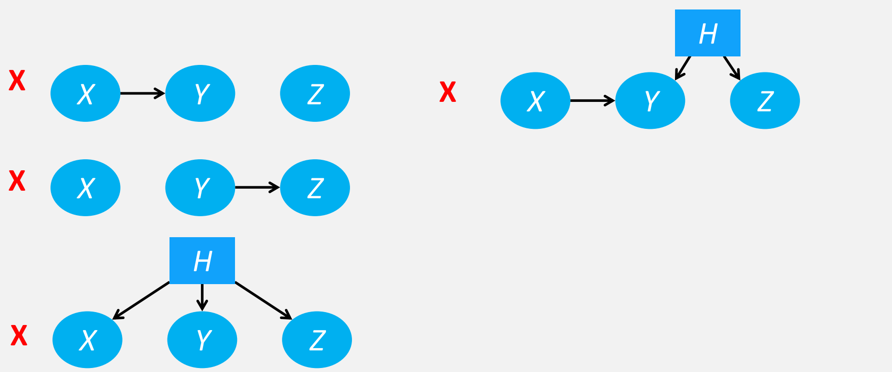
X 

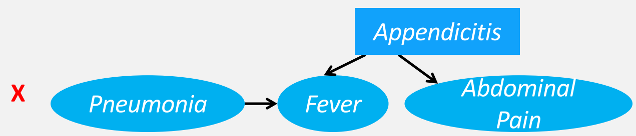
X 



Constraint-Based Search for a Causal Model: Example

- Suppose statistical testing yields the following constraints
  dep(*X*, *Y*), dep(*Y*, *Z*), dep(*X*, *Z*), ind(*X*, *Z* | *Y*)
- Consider the consistency of these constraints with respect to the following causal models:



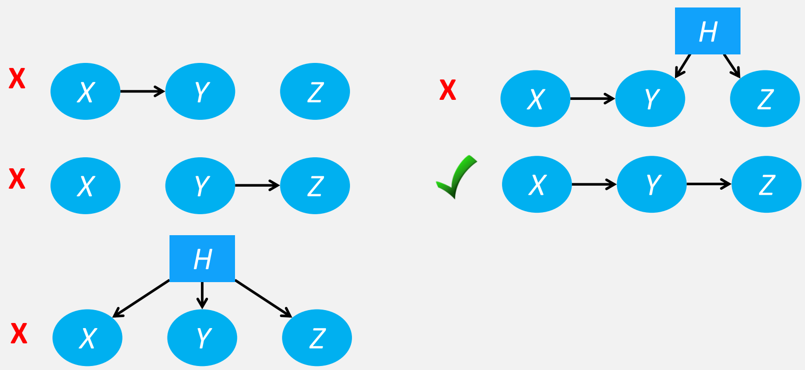


X Y Z

- Given Fever = present, if Pneumonia = present then Appendicitis Is unlikely and therefore Abdominal Pain is unlikely.
- Given Fever = present, if Pnemonia = absent then Appendicitis is likely and therefore Abdominal Pain is likely.
- Thus, when fever is present, Pneumonia and Abdominal Pain have an inverse statistical relationship
- This causal model is not consistent with the known constraint $ind(X, Z \mid Y)$.
- The pneumonia story is more complicated because pneumonia does lead to toxemia which leads to abdominal pain

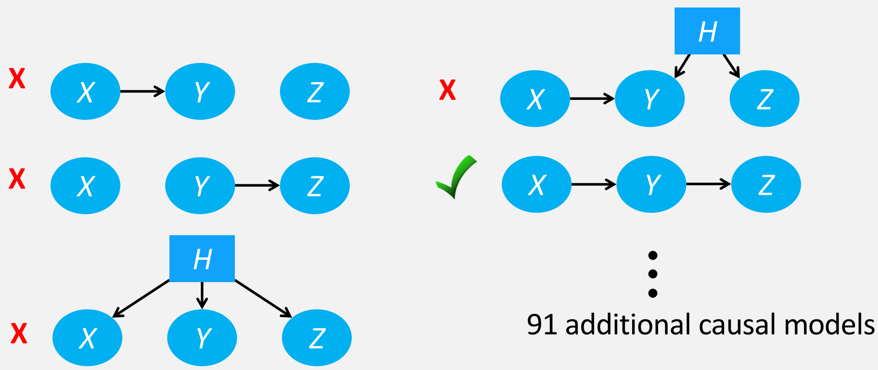
Constraint-Based Search for a Causal Model: Example

- Suppose statistical testing yields the following constraints
  $dep(X,Y), dep(Y,Z), dep(X,Z), ind(X,Z \mid Y)$
- Consider the consistency of these constraints with respect to the following causal models:



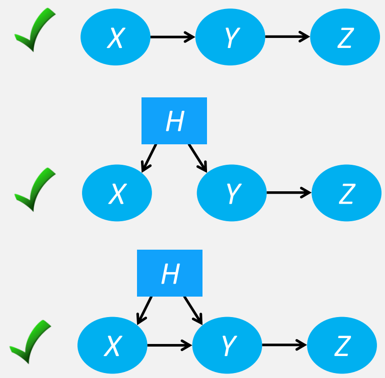
Constraint-Based Search for a Causal Model: Example

- Suppose statistical testing yields the following constraints
  dep(*X*, *Y*), dep(*Y*, *Z*), dep(*X*, *Z*), ind(*X*, *Z* | *Y*)
- Consider the consistency of these constraints with respect to the following causal models:



Constraint-Based Search for a Causal Model: Example

- The following models are the only ones consistent with the constraints:



- In all of these models, *Y* causes *Z* and there is no confounding of *Y* and *Z*.

Constraint-Based Search for a Causal Model: Example

- Reduce the large number of causal model possibilities to just those models consistent with the constraints obtained from the data
- Look for causal relationships that are invariant across those models (e.g., Y → Z).

Constraint-Based Causal Discovery Algorithms

- They find general patterns of statistical dependency among the measured variables that are consistent with the causal models that they output
- They make the following assumptions:
    - Causal Markov Condition: Causality is local.

Constraint-Based Causal Discovery Algorithms

- They find general patterns of statistical dependency among the measured variables that are consistent with the causal networks that they output
- They make the following assumptions:
  - Causal Markov Condition: A node is independent of its non-effects given it direct causes. $A \rightarrow B \rightarrow C$
  - Causal Faithfulness Condition: The only independence among nodes is due to the Causal Markov Condition.
  - Test accuracy: The tests of statistical independence are correct.

Evaluation of Causal Models

- Evaluating classifiers trained using machine learning is relatively straightforward (ground truth is known)
- Evaluating causal models is not as straightforward
  - Need independently discovered causal relationships
  - Need new experiments
  - Additional background knowledge
  - Open area of research

Learning causal models

- Learning causal models from observational data
  - Constraint-based
  - Bayesian
  - Other
- Learning causal models from observational and experimental data

Challenges of Bayesian Learning of Causal Models from Observational Data

- Major challenges
  - Large search spaces
  - Hidden variables
  - Feedback
  - Assessing parameter and structure priors
  - Modeling complicated distributions