Causality II: How does causal inference fit into public health and what it is the role of statistics?

Statistics for Psychosocial Research II

November 13, 2006

Outline

- Potential Outcomes / Counterfactual
  - Judea Pearl
- Statistics and Causal Inference
  - Aalen and Frigessi
- Causal Inference in Epidemiology
  - Parascandola and Weed

Some history

- 1910's & 1920's: Sewall Wright
  - Introduced path analysis in biology and agricultural sciences
  - Path analysis conceptualized as a method using a sequence of regression models to make inferences about correlations and effects
- 1930's & 1950's: Economists (e.g., Haavelmo) and Bollen
  - SEMs evolve which generalize path models to allow correlated errors.
- 1950's: Turner and Stevens
  - Introduce path analysis and SEMs to statistics literature
- 1960's: Duncan
  - Introduces path analysis to psychological literature
  - Path analysis and SEM used interchangeably
- 1970's: Rubin
  - Introduces concept of potential outcomes. Opens a whole new can of worms….

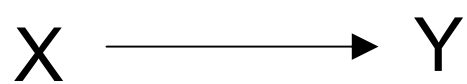
The setting

$$Y = \gamma_1 + \tau X + \varepsilon_1$$

- Bland Statistical Interpretation: A one unit change in X is associated with an expected change in Y of τ units.
- We would really LIKE to say: If I am nature, and I change an X in the population from 0 to 1, then Y will change by τ units on average.

Causal Inference

- Relatively new field
- Uses statistics to make causal inference
- Interplay between science and statistics
  - Science dictates model
  - Statistics measures magnitude of effect
- Hypothesis:



Example: pregnancy smoking and childhood conduct problems

- X = pregnancy smoking (binary)
- Y = childhood conduct (continuous)

- Data Structure:
  - $Y_i(0)$ = childhood conduct for child i if mom does not smoke.
  - $Y_i(1)$ = childhood conduct for child i if mom does smoke.
- Each child has the "potential" to have an outcome under either scenario.
- But, we only observe the outcome under the observed smoking status

Maughan et al, J. Child Psychol. Psychiat. Vol. 42, No. 8, pp. 1021-1028, 2006

Causal Estimation

Observed data:

$$Y = \alpha + \beta_0 X + \varepsilon$$

Observed effect: β_0

"Full" data:

$$E[Y(1)] - E[Y(0)] = \beta_C$$

Causal effect: β_C

Causal Estimation

- When does $\beta_0 = \beta_C$?
- Main problem: Counterfactual
  - E[Y(1) | X = 1] is observed
  - E[Y(1) | X = 0] is "counterfactual"
- How do we identify counterfactual outcomes?

Randomization

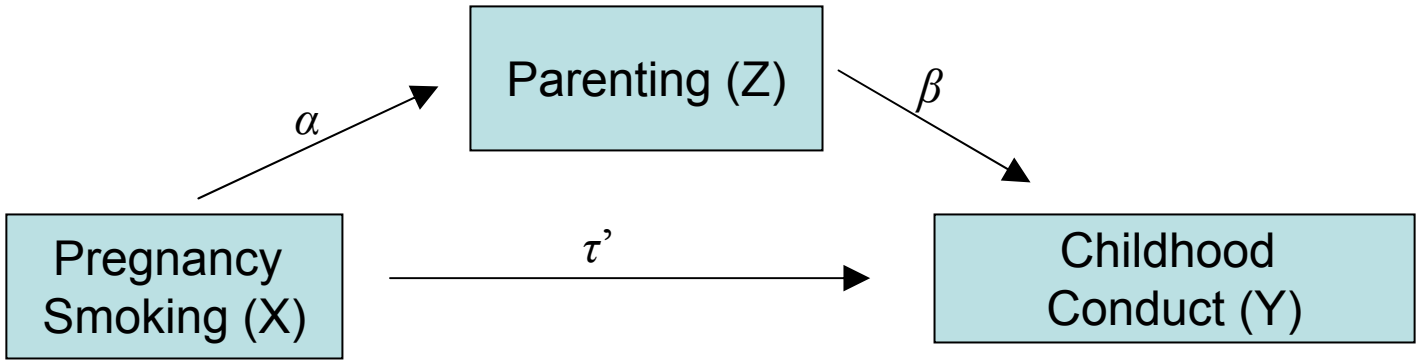
- Under randomization, $Y_i(0)$ and $Y_i(1)$ are INDEPENDENT of X_i
- That is, $E[Y_i(0) \mid X_i = 1] = E[Y_i(0) \mid X_i = 0]$
- In words, counterfactual and observed outcomes are exchangeable

Observational Study

- In an observational study, we cannot assume exchangeability

- But, if we can control for confounders, we can regain exchangeability

Mediation Model



$$Y = \tau X + \varepsilon_1$$

$$Z = \alpha X + \varepsilon_2$$

$$Y = \tau' X + \beta Z + \varepsilon_3$$

Total effect of X on Y is τ

Mediation Model

- The indirect effect is the effect of X on Y that is mediated by Z
- One measure of the mediated effect is $\alpha\beta$.
- Under the potential outcomes framework, a mediated effect might be

  $E[\, Y(X = 1, Z = z_1) - Y(X = 1, Z = z_2)]$
- Identification of the potential outcomes is tricky

Mediation Model

- Intuition when Y is binary:

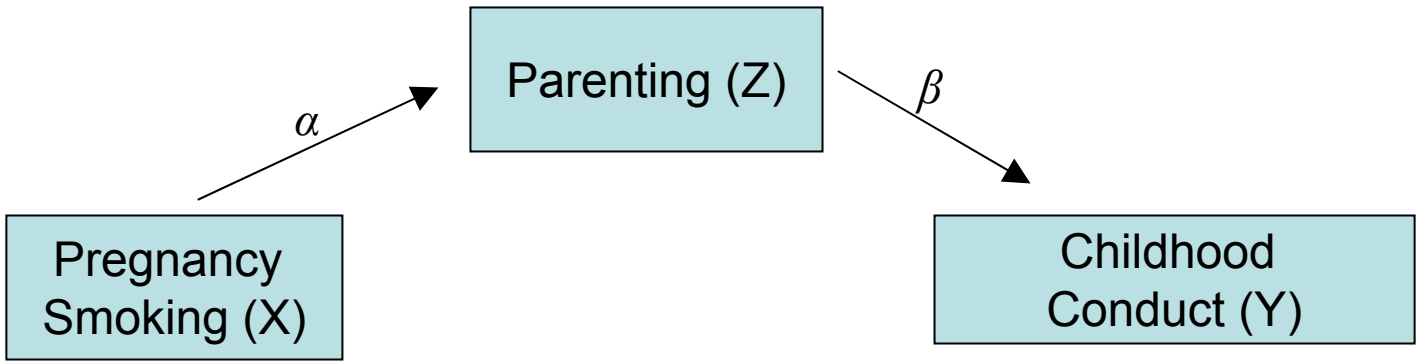
  $E[Y(X = 1, Z = z_1)] = P[Y(X = 1, Z = z_1)]$

- We can talk about odds and odds ratios:

$$OR_C = \frac{oddsP[Y(X = 1, Z = 1)]}{oddsP[Y(X = 1, Z = 0)]}$$

Mediation Model

Complete Mediation

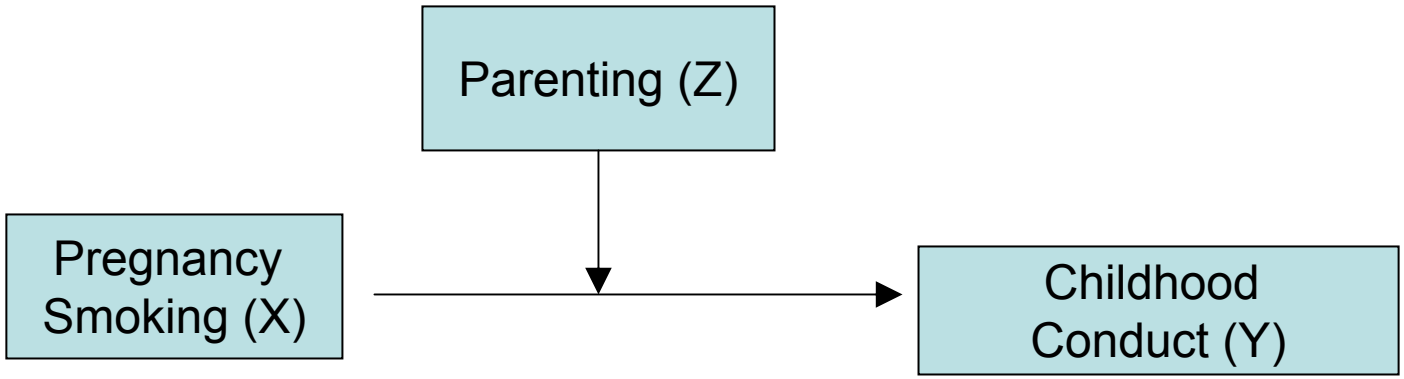


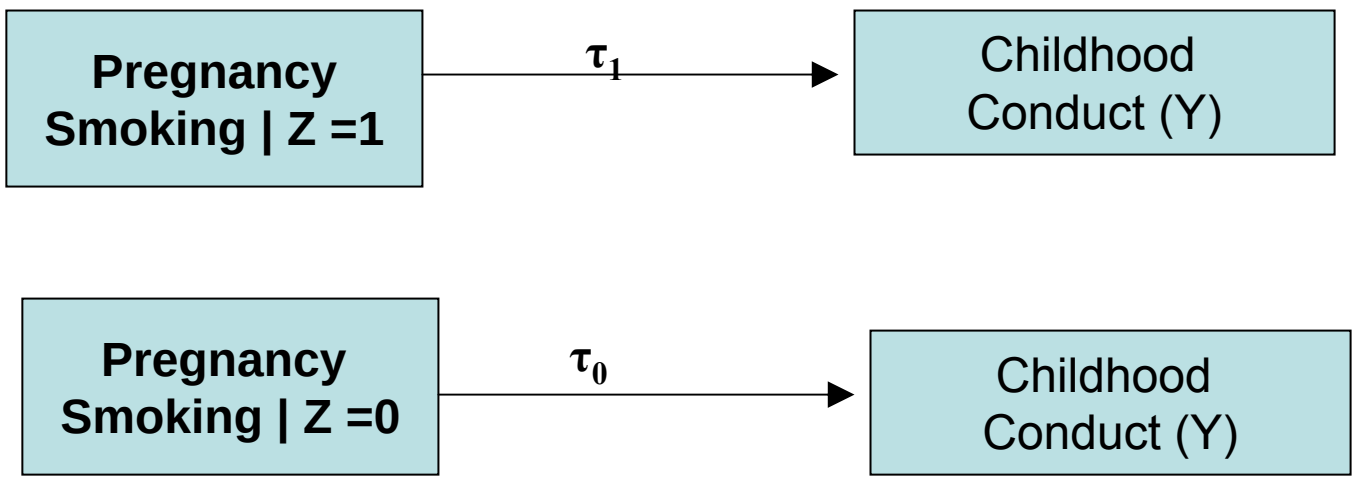
$$P[Y(Z = 1 \mid X = 1)] = P[Y(Z = 1 \mid X = 0)]$$
$$P[Y(X = 1 \mid Z = 1)] \neq P[Y(X = 1 \mid Z = 0)]$$

Mediation Model

- Under complete mediation,
  - $P[Y(X = 1,Z = 1)] = P[Y(X = 0, Z = 1)]$
  - $P[Y(X = 1,Z = 1)] \neq P[Y(X = 1, Z = 0)]$
- In words,
  - For kids with “good” parenting, pregnancy smoking IS NOT associated with conduct disorder
  - For kids whose mom’s smoked during pregnancy, their conduct is still associated with parenting.

Effect Modification?





Historical Problem: Path Analysis

- Traditional statistical formulation of direct and indirect effects
- Weakness: absence of time
- Models of variables versus stochastic processes
- In our mind, we have “time” in mind, but not necessarily in our data

Historical Problem: Path Analysis

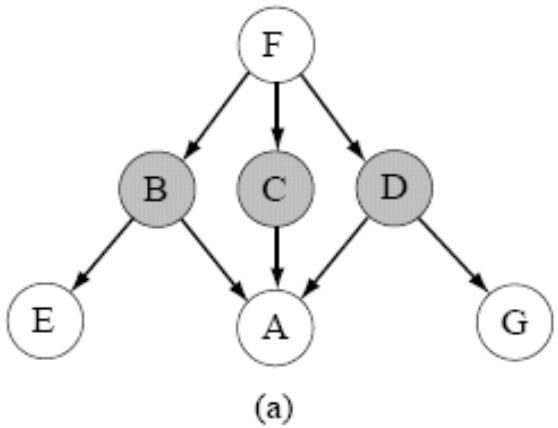
- Associational versus causal relationships
- Distinction has not been made clear
- Statisticians tend to caution people about interpretation
- But, there has tended to be no formal distinction in modeling

Causal Notation

- “do”: forcing a change
- Judea Pearl et al.
- P[Y(X = 1)|do(Z=1)]
- “do” indicates changing the state of nature
- CAUSAL versus OBSERVED conditioning
- Difference is P[Y|X] and P[Y|do(X)]
  - P[Y|do(X)]: what is the change in the expected value of Y if we were to intervene and change the value of X from x to x+1?
  - P[Y|X]: what would be the difference in the expected value of Y is we were to FIND X at level x+1 instead of x?
- Recent work in causal inference “forces” us to explicitly state assumptions
- Failing to do so may lead to incorrect inferences

Causal Notation

- Directed Acyclic Graphs (DAGS)
- Important concepts in understanding causality
  - D-separation
  - Blocking
  - Colliders
  - Non-colliders
  - Descendants



(a)

So far….

- “Potential Outcomes” (i.e., counterfactual) framework
- Changes the way statisticians think about structural modeling
- Application? Used appropriately in medicine and clinical trials.
- What about observational studies? Does it still fit?

What can statistics contribute to causal understanding?

- Many problems in statistics applied to causal inference
- Pearl, in response to avoidance of causality in statistics literature : “This position of caution and avoidance has paralyzed many fields that look to statistics for guidance, especially economics and social science.”

What can statistics contribute to causal understanding?

- Various types of causal thinking
- Experience based
  - Statistics specializes in this
  - If you take the medicine, you will be cured
  - Why? Doesn't necessarily matter
  - Analysis of randomized clinical trial does not need to understand mechanism for treatment effect
  - "black box" causality
- Mechanistic based
  - Looks into the "black box" to understand mechanism
  - Validity of mechanism varies substantially in medical research (e.g., heart function versus cancer versus psychiatric disorders)
- Can think of a "hierarchy" between these (more later).

New Developments in Statistics (Pearl, 2000)

- Precise definitions are given of what one should mean by a causal effect (e.g., counterfactual)
- This has clarified causal “thinking”
- New methods for approximating counterfactual comparison (e.g. marginal structural model (Robins, 1986))
- Sensitivity studies can be made to see whether confounders or other factors can explain differences

Mechanistic Causality

- These developments apply to experience based causality
- Mechanistic insights driving science play little role in analysis
- Randomized trial ignores mechanism
- Counterfactual causality typically related to action being taken (e.g. pregnancy smoking)
- Mechanistic causality aims at understanding mechanisms or processes.
- May be 2ndary to understand whether or not mechanisms can be influenced

Mechanistic Causality

- Statistics is generally most helpful when mechanism is very poorly understood
- Mechanisms unfold over time.
- Need sequence of events
- “Granger causality”
  - Measurements taken over time
  - How they influence each other
  - Present and past influencing future

Levels of Mechanistic Understanding

- Can be studied at many levels
- Example: genetic studies
  - Can derive genetic versus environment component
  - Can say “genetic cause”
  - But, wouldn’t it be more detailed to know which genes were the cause?
- Statistics: causality tends to be thought of absolute
  - Better to think in less definite terms
  - A study can make a step towards mechanistic understanding
  - But, understanding may still be superficial
- These concepts depend on the level of detail
- A direct effect may become an indirect effect if new intermediate variables are observable

Causation in Epidemiology

- Essential in epidemiology
- But, no agreed upon definition
- Five categories can be delineated
  - Production
  - Necessary and sufficient
  - Sufficient-component
  - Counterfactual
  - Probabilistic

“Production”

- A cause is something that produces or creates an effect
- Definition of production and creation are not well-defined
- Rejected due to this ambiguity

“Necessary & Sufficient Causes”

- Necessary: must be present for effect to occur
- Sufficient: in its presence effect must occur
- 4 combinations
- Few epidemiologists believe that “cause” should be limited to necessary conditions
- Support is based on scientific *determinism* and “one cause” model
- Requires one-to-one correspondence.
- No role for chance.
- Too many “neither necessary nor sufficient” to make this practical.

Sufficient-Component

- Rothman: widely cited.
- Causes can be neither sufficient nor necessary
- Made up of a number of components, no one of which is sufficient on its own.
- Still assumes determinism: no variation or chance allowed.
- Must assume existence of countless hidden effects: big assumption
- Rejected because unwieldy.

Probabilistic Causation

- A cause increases the probability that an effect will occur
- Offers alternative to determinism
- Makes fewer biological assumptions
- Little discussion of their strengths and weaknesses in epi literature
- Fails to explain, for example, why some smokers develop cancer and others do not.
- Unclear about what it means to “increase” the probability
- Cox and Holland object to this idea: how can causal and non-causal associations be differentiated?
- On its own, probabilistic causation is not enough

Counterfactuals

- Compares outcomes under different conditions
- Can be either deterministic or probabilistic
- Counterfactuals are not *inconsistent* with 4 previous definitions
- They articulate additional attribute by strengthening distinction between cause and correlation
- Counterfactual alone is not sufficient for causation

Probabilistic + Counterfactual

- This combination is suggested as the best option for epidemiology
- Consistent with both deterministic and probabilistic models.
- Makes few assumptions about unobservables
- Probabilistic is implicit in practical reasoning:
  - What does physician mean when she tells her patient that he can reduce risk of lung cancer by giving up smoking?
  - Does she mean he MIGHT be an individual for whom smoking 'tips the balance'?
  - Implication: if other "component causes" are known, he might not need to give up smoking.
  - Trivializes the nature of public health advice
  - By quitting smoking, the probability of lung cancer is lowered.
  - Deterministic account does not allow this approach.