Causal Inference

Statistics for Psychosocial Research II
Structural Models
November 8, 2006
Many views on this topic!

- Correlation $\neq$ Causation
- But, coupled with other information, correlation can imply causation
- Statistics helps a lot with causal inference
- Statistical models used to draw inferences are distinctly different from those used for showing associational differences
‘Potential’ Cause

• Holland (1986): each ‘unit’ of observation must be able to be ‘exposed’ to the cause

• For causal inference, cause must be subject to “human manipulation.”

• Does
  – Smoking cause lung cancer?
  – Sleet or snow cause traffic accidents?
  – A change in interest rates cause the stock market to fluctuate?
  – Gender or race cause discrimination?
Three examples from Holland

(A) She did well on the exam because she is a woman.
(B) She did well on the exam because she was coached by her teacher.
(C) She did well on the exam because she studied for it.

In (A), is there a ‘cause’? NO
In (B), is there a ‘cause’? YES
In (C), is there a ‘cause’? ?

Why (C)? Studying is voluntary: Can we MAKE someone study? We COULD prevent someone from studying. Debatable....
“Potentially Exposable”

• Every ‘unit’ should be able to be exposed to the cause.

• Good example: randomized clinical trial

• Need to be able to postulate that we could state what WOULD have happened to a patient’s outcome had the cause been “the reverse”.
  – Assume $Y_{ti}$ is the outcome of $Y_i$ if patient $Y_i$ is in the treatment group
  – Assume $Y_{ci}$ is the outcome of $Y_i$ if patient $Y_i$ is in the control group

• We are interested in the causal effect: $Y_{ti} - Y_{ci}$
Fundamental Problem of Causal Inference

“It is impossible to observe the value of $Y_{ti}$ and $Y_{ci}$ on the same patient. Therefore it is impossible to observe the causal effect of treatment on patient $Y_i$.”

• Important point: ‘observe’ is key word.
• But, we can make inferences using statistical reasoning
• Possible exceptions: cross-over designs in some settings.
Statistical Solution

• We use information on a number of different patients to gain knowledge about causal effect.
• We cannot estimate causal effects for individuals.
• We CAN estimate ‘average’ causal effects over a population of patients.
Special Cases of Causal Inference

1. **Temporal Stability**: response does not depend on when exposure occurs.
2. **Causal Translucence**: prior exposure to cause does not affect outcome.
3. **Unit homogeneity**: effect is the same when cause is applied to identical units.
4. **Independence**: This is most relevant to our class.....
4. Independence

- Randomized trial is a special (very special!) case of independence.
- We cannot always assume that “exposure” is assigned randomly.
- Example: smoking and health outcomes
  - Is it reasonable to assume that smoking is randomly assigned?
  - What would be the health outcomes of smokers if, holding all else constant, they were instead non-smokers?
  - Would the health outcomes be the same as the ‘true’ non-smokers?
- ASSUMPTION: The determination of the cause is independent of all other variables, including the outcome of interest.
- Reasonable???
Causality

• Strong assumption of causality in SE models
• Does that mean we cannot include ‘gender’ in our models?
• No: it means that we cannot make ‘causal inferences’ about gender in our models.
• Bollen’s three components for ‘practically’ defining a causal relationship:
  – isolation
  – association
  – direction of influence
ASSOCIATION

• Easier to establish
• Causal variable should have strong association with outcome
• Problems:
  – incorrect standard errors or test statistics (e.g. correlated errors, poor measures)
  – multicollinearity
• Replication/Repetition important (and also helps establish isolation)
Multicollinearity Example

$\eta_1$: morale; $\eta_2$: sense of belonging

\[ x_1 = \gamma_{11}\eta_1 + \zeta_1 \]
\[ : \]
\[ x_4 = \gamma_{14}\eta_1 + \gamma_{24}\eta_2 + \zeta_4 \]
\[ : \]
\[ x_7 = \gamma_{27}\eta_2 + \zeta_7 \]

In truth, $x_4$ is a measure of morale, but we allow it to be related to sense of belonging.
Results? Both $\gamma_{14}$ and $\gamma_{24}$ are insignificant.
Why? Because morale and sense of belonging are highly associated.
Direction of Causation

• Plausibility of association being causal rests on having causal direction correct
• Temporal?
  – x should come before y in time
  – problematic: simultaneous reciprocal causation (feedback) is not possible
  – window of cause and response time
• We often have cross-sectional data.
• Can future event predict past or present event?
Aside: Total, Direct, and Indirect Effects

- $x_1$ is marital status, $y_1$ is income, $y_2$ is depression
- Direct effect: measured by a single arrow between two variables
- Indirect effects: measured by all possible “paths” or “connections” between two variables EXCEPT for the direct path. We multiply the coefficients on path together to get each indirect effect.
- Total effect: the sum of the direct and indirect paths between two variables
ISOLATION

• Isolation: hold everything constant except the cause and the outcome
• Impossible to establish unless x and y occur in a “vacuum”
• Especially difficult in observational studies!
• Without true isolation can never be 100% certain about cause
• Is that ‘weird’ in statistics? NO! We are never 100% certain in statistics!
• **Isolation tends to be the weakest link in determining causality**
“Pseudo-isolation”

\[ y_1 = \gamma_{11} x_1 + \zeta_1 \]

- \( \zeta_1 \) is the unobserved error/disturbance
- \( \zeta_1 \) represents ALL other causes/correlates of \( y_1 \)
- Standard assumption for pseudo-isolation:
  \[ \text{Cov}(x_1, \zeta_1) = 0 \]
- That is, \( x_1 \) is independent of all other causes/correlates of \( y_1 \)
- If the assumption is true, then we can assess causal association of \( x_1 \) and \( y_1 \) “isolated” from all other causes (\( \zeta_1 \)).
“Pseudo-isolation”

- Can think of pseudo-isolation as a probabilistic view of causality
- Predictability of $y_1$ lies between two models:

\[ y_1 = \gamma_{11} x_1 \quad \text{All Cause} \]

\[ y_1 = \zeta_1 \quad \text{All Error} \]
Practically Speaking

• Unrealistic to think that $x_1$ is the only cause of $y_1$ and that $\text{Cov}(x_1, \zeta_1) = 0$.
• We need to account for other factors (e.g. cancer, smoking, coffee example).

\[ y_1 = \gamma_{11} x_1 + \gamma_{12} x_2 + \cdots + \gamma_{1m} x_m + \zeta_1 \]

• Latent variable approach? Same…..

\[ y_1 = \gamma_{11} \eta_1 + \gamma_{12} \eta_2 + \cdots + \gamma_{1m} \eta_m + \zeta_1 \]

\[ \text{Cov} (\eta_i, \zeta_1) = 0 \]
Examples of Violations of Isolation

(1) INTERVENING VARIABLES

True Model:

\[ y_1 = \gamma_{11} x_1 + \zeta_1 \]
\[ y_2 = \beta_{21} y_1 + \gamma_{21} x_1 + \zeta_2 \]
\[ \text{Cov}(\zeta_1, \zeta_2) = 0 \]
\[ \text{Cov}(x_1, \zeta_1) = 0 \]
\[ \text{Cov}(x_1, \zeta_2) = 0 \]

(e.g. \( x_1 \) is marital status, \( y_1 \) is household income, \( y_2 \) is depression)
What if we omit $y_1$ (income)?

• Assumed model:
  \[ y_2 = \gamma_{21} x_1 + \zeta_2 \]

• This implies:
  \[ \zeta_2^* = \beta_{21} y_1 + \zeta_2 \]

• And our pseudo-isolation assumption....

\[
\text{Cov}(x_1, \zeta_2^*) = \text{Cov}(x_1, \beta_{21} y_1 + \zeta_2) \\
= \text{Cov}(x_1, \beta_{21} (\gamma_{11} x_1 + \zeta_1) + \zeta_2) \\
= \beta_{21} \gamma_{11} \text{Var}(x_1) + \beta_{21} \text{Cov}(x_1, \zeta_1) + \text{Cov}(x_1, \zeta_2) \\
= \beta_{21} \gamma_{11} \text{Var}(x_1) \\
\neq 0
\]
Effect on Inference?

- $\gamma_{21}^*$ converges to total effect, $\beta_{21} \gamma_{11} + \gamma_{21}$, instead of direct effect, $\gamma_{21}$

- This yields an over or under-estimate of the effect of $x_1$ on $y_2$.

- Can be a really big problem if direct and indirect effects cancel each other out.
  - If $\gamma_{21} = 1$; $\beta_{21} = 0.5$; $\gamma_{11} = -2$
  - Then, $\gamma_{21}^* = -0.5 \times 2 + 1 = 0$
  - We might conclude that there is NO association!
(2) LEFT OUT COMMON CAUSE

Recall True Model

\[ y_1 = \gamma_{11} x_1 + \zeta_1 \]
\[ y_2 = \beta_{21} y_1 + \gamma_{21} x_1 + \zeta_2 \]
\[ \text{Cov}(\zeta_1, \zeta_2) = 0 \]
\[ \text{Cov}(x_1, \zeta_1) = 0 \]
\[ \text{Cov}(x_1, \zeta_2) = 0 \]

What if we omit \( x_1 \) from the model?

Then \( y_2 = \beta_{21}' y_1 + \zeta_2' \)

where \( \zeta_2' = \gamma_{21} x_1 + \zeta_1 \)
• Is pseudo-isolation assumption violated?

\[ \text{Cov}(y_1, \zeta_2^*) = \text{Cov}(\gamma_{11}x_1 + \zeta_1, \gamma_{21}x_1 + \zeta_2) \]
\[ = \gamma_{11}\gamma_{21}\text{Var}(x_1) \]
\[ \neq 0 \]

• What happens to our estimate of \( \beta_{21} \)?

\[ \beta_{21}^* = \beta_{21} + \gamma_{21}\gamma_{11} \]

(again, we get total effect instead of direct)
Effects on Inference

- Worst case scenario: $y_1$ and $y_2$ have little or no association, but both are highly associated with $x_1$.
- Example:
  - $x_1 = \text{age}$
  - $y_1 = \text{proportion of gray hairs}$
  - $y_2 = \text{quality of vision}$
- “Spurious Association”
(3) OMITTED VARIABLE HAS UNSPECIFIED RELATION TO OTHER VARIABLES

True Model:
\[ y_1 = \gamma_{11} x_1 + \gamma_{12} x_2 + \zeta_1 \]

- What if we omit \( x_2 \)?
  - Assumed model is \( y_1 = \gamma_1^* x_1 + \zeta_1^* \)
  - And \( \gamma_1^* = \gamma_{11} + \rho_{12} \gamma_{12} \)
This is an even bigger problem....

• Note that the association between $x_1$ and $x_2$ is unspecified: It could be true that
  – $x_1$ causes $x_2$ and $y_1$ (intervening variable)
  – $x_2$ causes $x_1$ and $y_1$ (common cause)
  – something else

• We can’t infer about the exact consequences of omitting $x_2$ because we don’t know its association to the other variables.
Other Violations

• “feedback” or “reciprocal causation”
• Wrong functional form between 2 variables
• Correlated errors