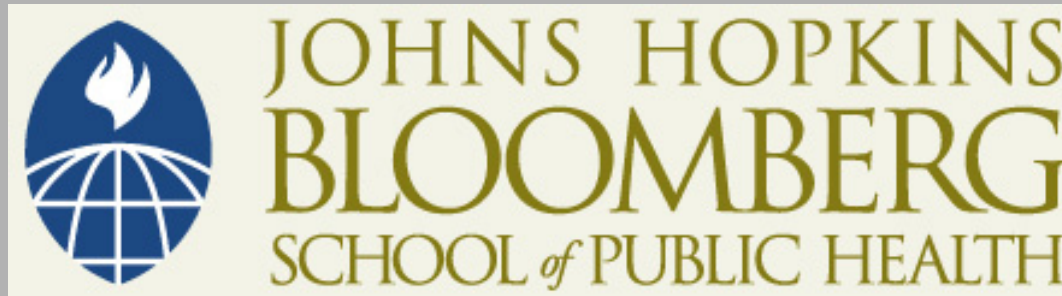


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2007, The Johns Hopkins University and Qian-Li Xue. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

# Concluding Topics

Statistics for Psychosocial Research II:  
Structural Models

Qian-Li Xue

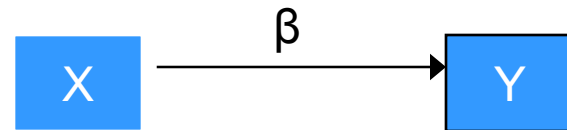
# Outline

- Brief discussion of Standardized Coefficients
- Causal inference
- Design, power, sample size

# Standardized Coefficients

- Make “relative” direct influences clearer
- Standardized coefficient of  $\beta$ :

$$\beta^* = \beta \frac{\sigma_x}{\sigma_y}$$



- The standardized effect is the mean change in standard deviation units of Y for a one standard deviation change in X.
- If X increases by one standard deviation, then we expect that Y will increase by  $\beta^*$  standard deviations
- Standardized can be easier for making inferences from SEMs

# Quick Little Derivation

$$\text{Model: } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$$

Standardizing Variables

$$x_1^* = x_1 / \sigma_{x_1}; \quad x_2^* = x_2 / \sigma_{x_2}; \quad y^* = y / \sigma_y$$

$$\sigma_{x_1} x_1^* = x_1; \quad \sigma_{x_2} x_2^* = x_2; \quad \sigma_y y^* = y$$

Substitute standardized variables into model:

$$\sigma_y y^* = \beta_0 + \beta_1 \sigma_{x_1} x_1^* + \beta_2 \sigma_{x_2} x_2^* + e$$

$$y^* = \frac{\beta_0}{\sigma_y} + \beta_1 \frac{\sigma_{x_1}}{\sigma_y} x_1^* + \beta_2 \frac{\sigma_{x_2}}{\sigma_y} x_2^* + e$$

# Standardized Coefficients

- MPLUS provides different types of “standardized” estimates
  - StdYX –  $\beta^* = \beta \frac{\sigma_x}{\sigma_y}$
  - StdY –  $\beta^* = \beta / \sigma_y$ 
    - ❖ more appropriate for binary variable
    - ❖ interpreted as the change in standard deviation units of y when x changes from zero to one
  - Std – uses the variances of the continuous latent variables for standardization
  - StdYX and Std are the same for parameter estimates involving only latent variables

# Standardized Coefficients

- Caveats:
  - Beware of comparing standardized coefficients for the same variable across groups!
    - ❖ Standardized effects might be different due to different standard deviations
    - ❖ **Look at unstandardized coefficients when comparing across groups**

# Latent Variables in SEM

- Much like path analysis with observed variables
- Some additional considerations

# Constraining Latent Variables

- Recall: 2 options for CFA model
  - set variance of (exogenous) LV equal to 1 (my preference)
  - set path coefficient to one of its indicators equal to 1. This scales the LV in the same units as the indicator.
- More complicated models:
  - endogenous LV
    - ❖ set variance of its error term to 1 or
    - ❖ set path coefficient to one of its indicators to 1
- MUST do one of these or model is not identifiable!

# Causal Inference

# 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 causal 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?

# “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.
- Possible exceptions: cross-over designs in some settings.
- However, we CAN estimate ‘average’ causal effects over a population of patients
- Topic of the new course offered in 3<sup>rd</sup> and 4<sup>th</sup> term

# Causality

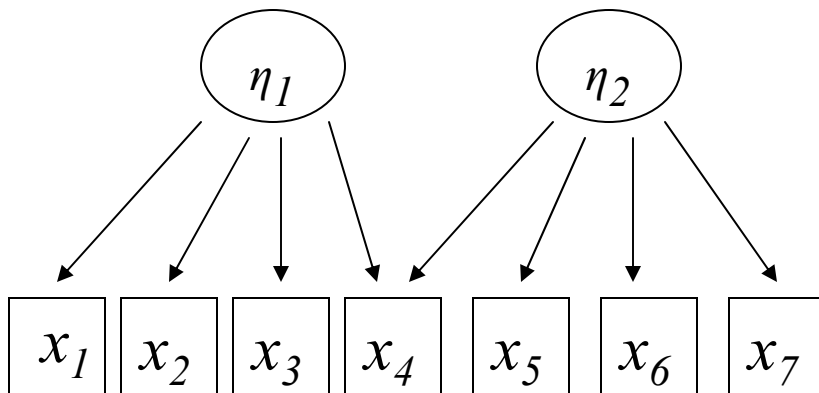
- Strong assumption of causality in SE models
- Bollen's three components for 'practically' defining a causal relationship:
  - Association
  - Direction of influence
  - Isolation

# 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



$$x_1 = \gamma_{11}\eta_1 + \zeta_1$$

$\vdots$

$$x_4 = \gamma_{14}\eta_1 + \gamma_{24}\eta_2 + \zeta_4$$

$\vdots$

$$x_7 = \gamma_{27}\eta_2 + \zeta_7$$

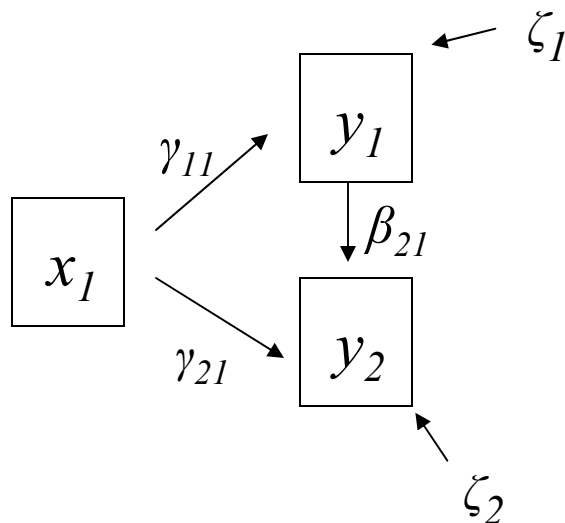
- In truth,  $x_4$  is a measure of morale, but we allow it to be related to sense of belonging
- Results? Both  $\gamma_{14}$  and  $\gamma_{24}$  are insignificant
- Why? Because morale and sense of belonging are highly associated

# 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?

# Recap: 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



Direct effect of  $x_1$  on  $y_2$ :

Indirect effect(s) of  $x_1$  on  $y_2$ :

Total effect of  $x_1$  on  $y_2$ :

---

Direct effect of  $y_1$  on  $y_2$ :

Indirect effect(s) of  $y_1$  on  $y_2$ :

Total effect of  $y_1$  on  $y_2$ :

# 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
- Isolation tends to be the weakest link in determining causality!

# “Pseudo-isolation”

$$y_1 = \gamma_{11}x_1 + \zeta_1$$

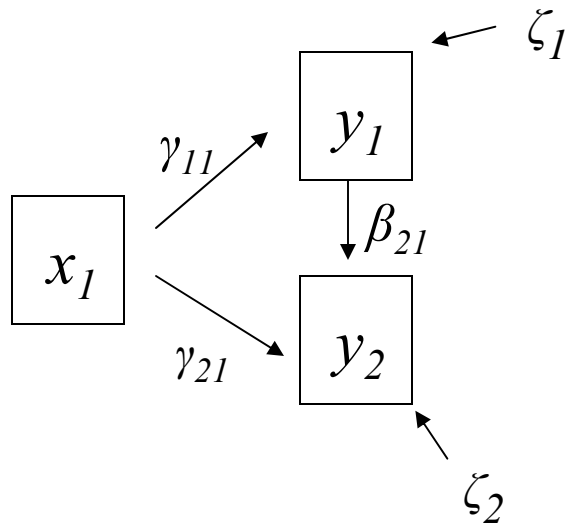
$\zeta_1$  is the unobserved error/disturbance

$\zeta_1$  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 ( $\zeta_1$ ).

# 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)

(Bollen, p.46)

# 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....

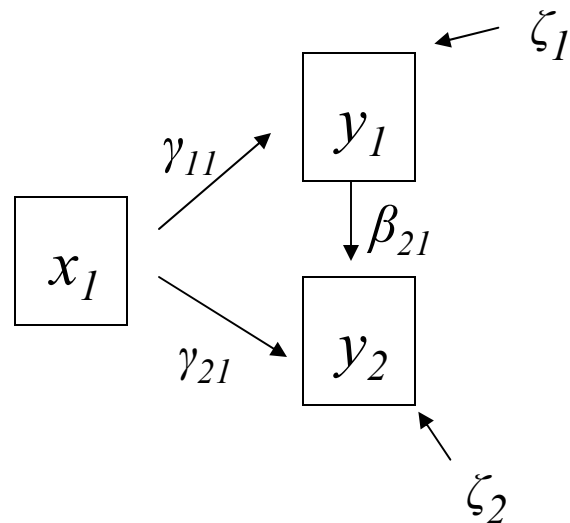
$$\begin{aligned} \text{Cov}(x_1, \zeta_2^*) &= \text{Cov}(x_1, \beta_{21} y_1 + \zeta_2) \\ &= \text{Cov}(x_1, \beta_{21} (\gamma_{11} x_1 + \zeta_1) + \zeta_2) \\ &= \beta_{21} \gamma_{11} \text{Var}(x_1) + \beta_{21} \text{Cov}(x_1, \zeta_1) + \text{Cov}(x_1, \zeta_2) \\ &= \beta_{21} \gamma_{11} \text{Var}(x_1) \\ &\neq 0 \end{aligned}$$

# 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,  $\gamma_{21}^* = -0.5 \cdot 2 + 1 = 0$
  - We might conclude that there is NO association!

# Examples of Violations of Isolation

## (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?

$$\text{Then } y_2 = \beta_{21}^* y_1 + \zeta_2^*$$

$$\text{where } \zeta_2^* = \gamma_{21}x_1 + \zeta_2$$

# Examples of Violations of Isolation

- Is pseudo-isolation assumption violated?

$$\begin{aligned}Cov(y_1, \zeta_2^*) &= Cov(\gamma_{11}x_1 + \zeta_1, \gamma_{21}x_1 + \zeta_2) \\ &= \gamma_{11}\gamma_{21}Var(x_1) \\ &\neq 0\end{aligned}$$

- What happens to our estimate of  $\beta_{21}$  ?

$$\beta_{21}^* = \beta_{21} + \gamma_{21}b_{11}$$

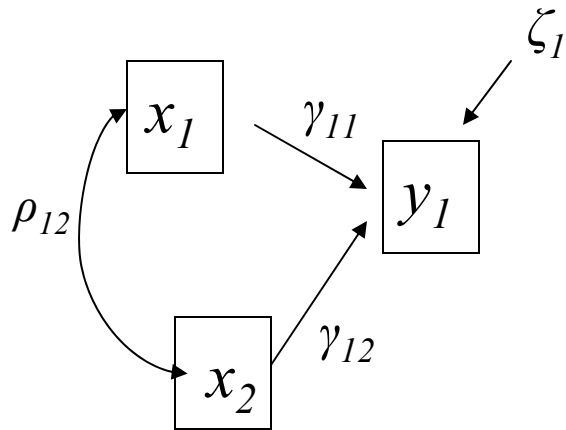
Where  $b_{11}$  is the regression coefficient from an “auxiliary” regression of  $x_1$  on  $y_1$

# 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:
  - $x_1$  = age
  - $y_1$  = proportion of gray hairs
  - $y_2$  = quality of vision
- “Spurious Association”

# Examples of Violations of Isolation

## (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  $x_2$ ?

$$y_1 = \gamma_{11}^*x_1 + \zeta_1^*$$

$$\gamma_{11}^* = \gamma_{11} + \rho_{12}\gamma_{12}$$

(Bollen, p.54)

# 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
- Bottom line:
  - SEM CAN NOT be used to PROVE causation!
  - Instead, SEM can be used to
    - ❖ “provide estimates of the strength of all the hypothesized relationships between variables in a theoretical model”
    - ❖ Assess consistency between data and model
    - ❖ “Distinguish between or among alternative perspectives”

# Sample Size and Power

- Factor analysis
  - General rule: need larger sample size (~300-500) for lower communalities, more factors, and fewer indicators per factor (MacCallum, 1999)
- SEM
  - $> 50 + 8 \times$  number of variables in the model
  - 10 to 20  $\times$  number of variables (Mitchell, 1993)
  - $>15$  cases per measured variable or indicator (Stevens, 1996)
  - $>5$  cases per parameter estimate (including error terms as well as path coefficients) is model assumptions are met (Bentler and Chou, 1987)
  - Larger N when data are non-normal or incomplete
- Monte Carlo Simulation studies

# Monte Carlo Simulation Study for a CFA

TITLE: this is an example of a Monte Carlo simulation study for a CFA with covariates (MIMIC) with continuous factor indicators and patterns of missing data

Simulation of a hypothetical population

MONTECARLO:

Numer of samples drawn

NAMES ARE y1-y4 x1 x2;

NOBSERVATIONS = 500;

NREPS = 500;

SEED = 4533;

!GENERATE=

CUTPOINTS = x2(1);

PATMISS = y1(.1) y2(.2) y3(.3) y4(1) |  
y1(1) y2(.1) y3(.2) y4(.3);

PATPROBS = .4 | .6;

generate different types of dependent variables (default: continuous)

Generate binary covariate based on Normal(0,1) using threshold=1

Specify patterns of missing data, here we have two missing patterns

MODEL POPULATION:

[x1-x2@0];

x1-x2@1;

f BY y1@1 y2-y4\*1;

f\*.5;

y1-y4\*.5;

f ON x1\*1 x2\*.3;

Provide true population parameter values for data simulation

MODEL:

f BY y1@1 y2-y4\*1;

f\*.5;

y1-y4\*.5;

f ON x1\*1 x2\*.3;

OUTPUT: TECH9;

# Monte Carlo Study Output

## MODEL RESULTS

		ESTIMATES			S. E.	M. S. E.	95%	% Sig
Population		Average	Std. Dev.	Average		Cover	Coeff	
F	BY							
	Y1	1.000	1.0000	0.0000	0.0000	1.000	0.000	
	Y2	1.000	1.0028	0.0663	0.0622	0.940	1.000	
	Y3	1.000	1.0005	0.0674	0.0633	0.930	1.000	
	Y4	1.000	1.0020	0.0806	0.0743	0.932	1.000	
F	ON							
	X1	1.000	1.0023	0.0671	0.0620	0.938	1.000	
	X2	0.300	0.2991	0.1065	0.1052	0.958	0.810	
Intercepts								
	Y1	0.000	-0.0018	0.0718	0.0688	0.950	0.050	
	Y2	0.000	0.0022	0.0515	0.0500	0.946	0.054	
	Y3	0.000	0.0015	0.0535	0.0519	0.944	0.056	
	Y4	0.000	0.0005	0.0603	0.0644	0.962	0.038	
Residual Variances								
	Y1	0.500	0.4887	0.0776	0.0779	0.944	1.000	
	Y2	0.500	0.4982	0.0502	0.0513	0.950	1.000	
	Y3	0.500	0.4966	0.0558	0.0532	0.926	1.000	
	Y4	0.500	0.4905	0.0755	0.0697	0.920	1.000	
	F	0.500	0.5019	0.0715	0.0679	0.930	1.000	