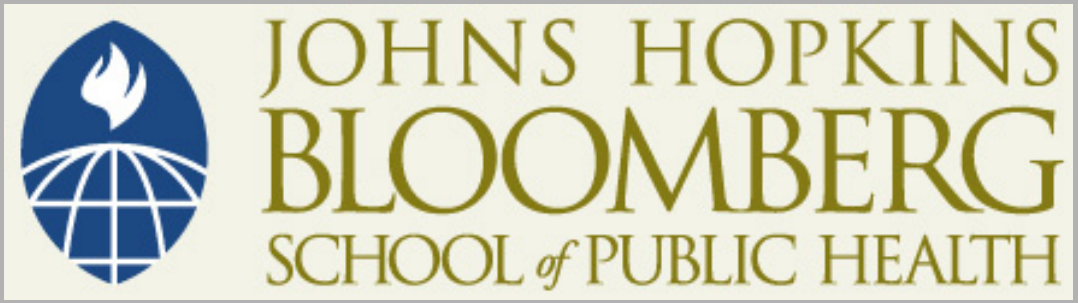


# Introduction to Path Analysis

Statistics for Psychosocial Research II:
Structural Models

Qian-Li Xue

# Outline

- Key components of path analysis
  - Path Diagrams
  - Decomposing covariances and correlations
  - Direct, Indirect, and Total Effects
- Parameter estimation
- Model identification
- Model fit statistics

# The origin of SEM

- Swell Wright, a geneticist in 1920s, attempted to solve simultaneous equations to disentangle genetic influences across generations ("path analysis")
- Gained popularity in 1960, when Blalock, Duncan, and others introduced them to social science (e.g. status attainment processes)
- The development of general linear models by Joreskog and others in 1970s ("LISREL" models, i.e. linear structural relations)

# Difference between path analysis and structural equation modeling (SEM)

- Path analysis is a special case of SEM
- Path analysis contains only observed variables and each variable only has one indicator
- Path analysis assumes that all variables are measured without error
- SEM uses latent variables to account for measurement error
- Path analysis has a more restrictive set of assumptions than SEM (e.g. no correlation between the error terms)
- Most of the models that you will see in the literature are SEM rather than path analyses

# Path Diagram

- Path diagrams: pictorial representations of associations (Sewell Wright, 1920s)
- Key characteristics:
  - As developed by Wright, refer to models that are linear in the parameters (but they can be nonlinear in the variables)
  - Exogenous variables: their causes lie outside the model
  - Endogenous variables: determined by variables within the model
  - May or may not include latent variables
    - for now, we will focus on models with only manifest (observed) variables, and will introduce latent variables in the next lecture.

# Regression Example

Standard equation format for a regression equation:

$$Y_1 = \alpha + \gamma_{11}X_1 + \gamma_{12}X_2 + \gamma_{13}X_3 + \gamma_{14}X_4 + \gamma_{15}X_5 + \gamma_{16}X_6 + \zeta_1$$

# Regression Example

$x_1$

$x_2$

$x_3$

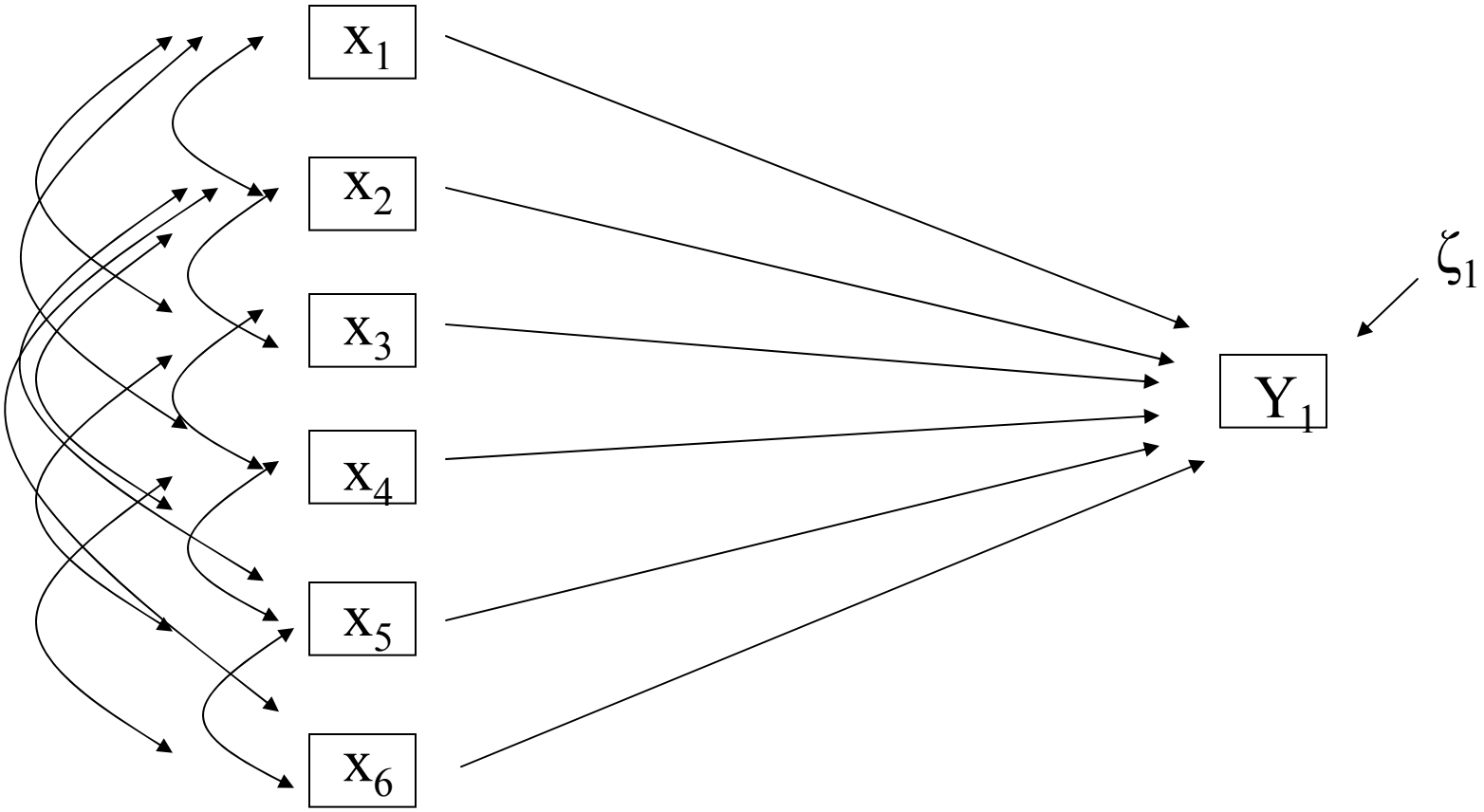
$Y_1$

$x_4$

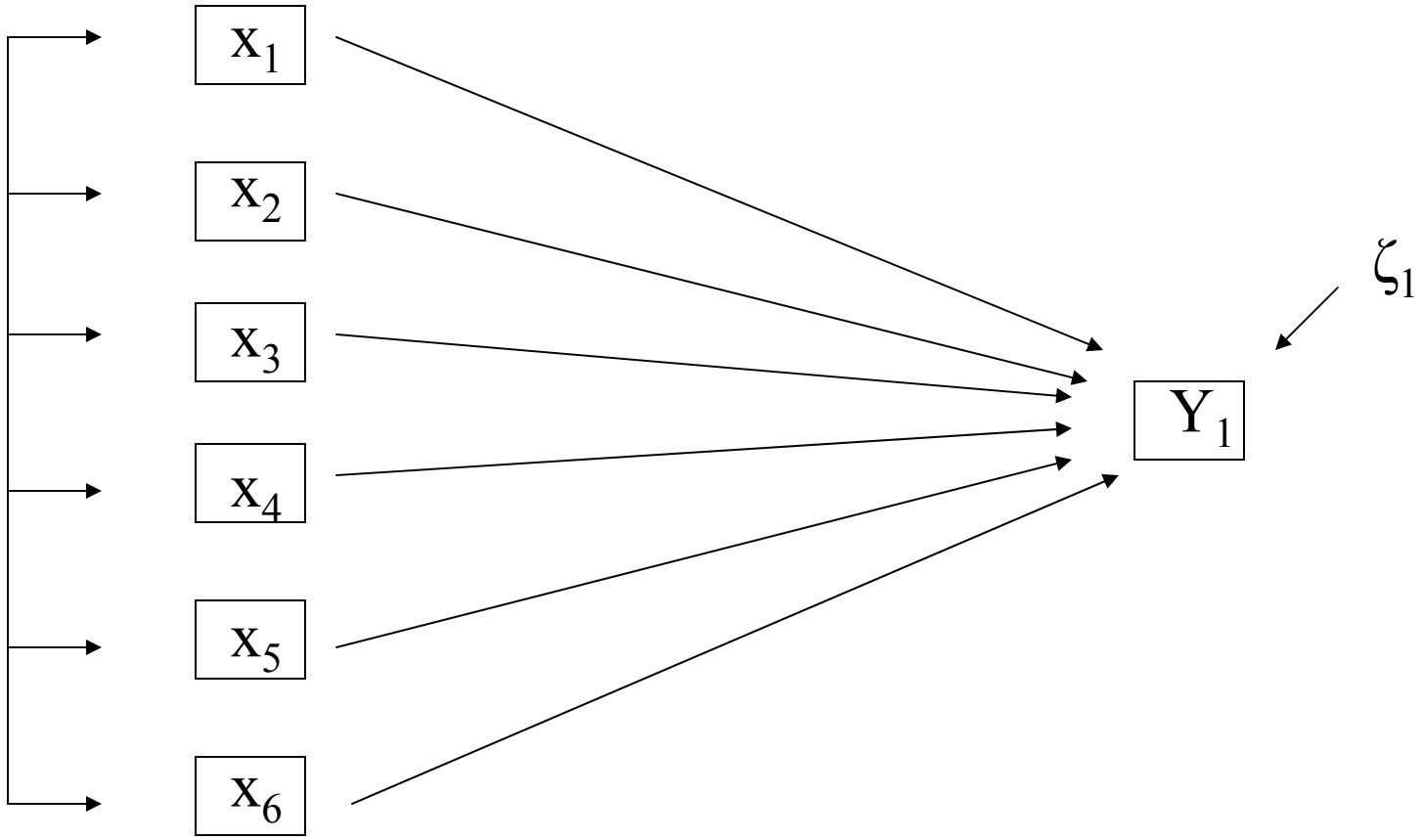
$x_5$

$x_6$

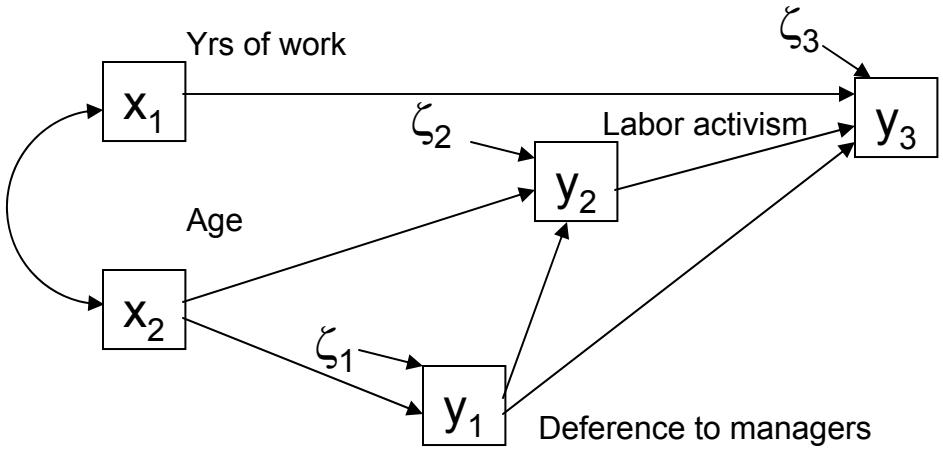
# Regression Example



# Regression Example



# Path Diagram: Common Notation



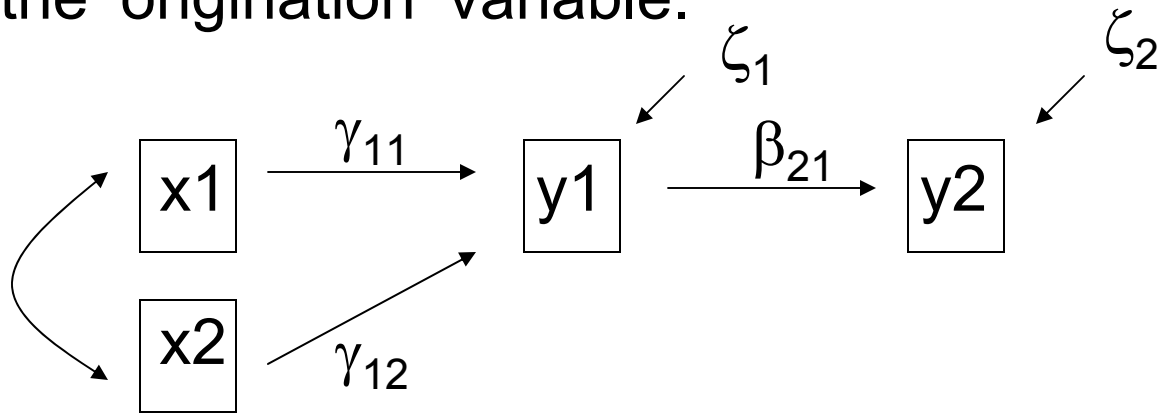

Union Sentiment Example

(McDonald and Clelland, 1984)

- Noncausal associations between exogenous variables indicated by two-headed arrows
- Causal associations represented by unidirectional arrows extended from each determining variable to each variable dependent on it
- Residual variables are represented by unidirectional arrows leading from the residual variable to the dependent variable

# Path Diagram: Common Notation

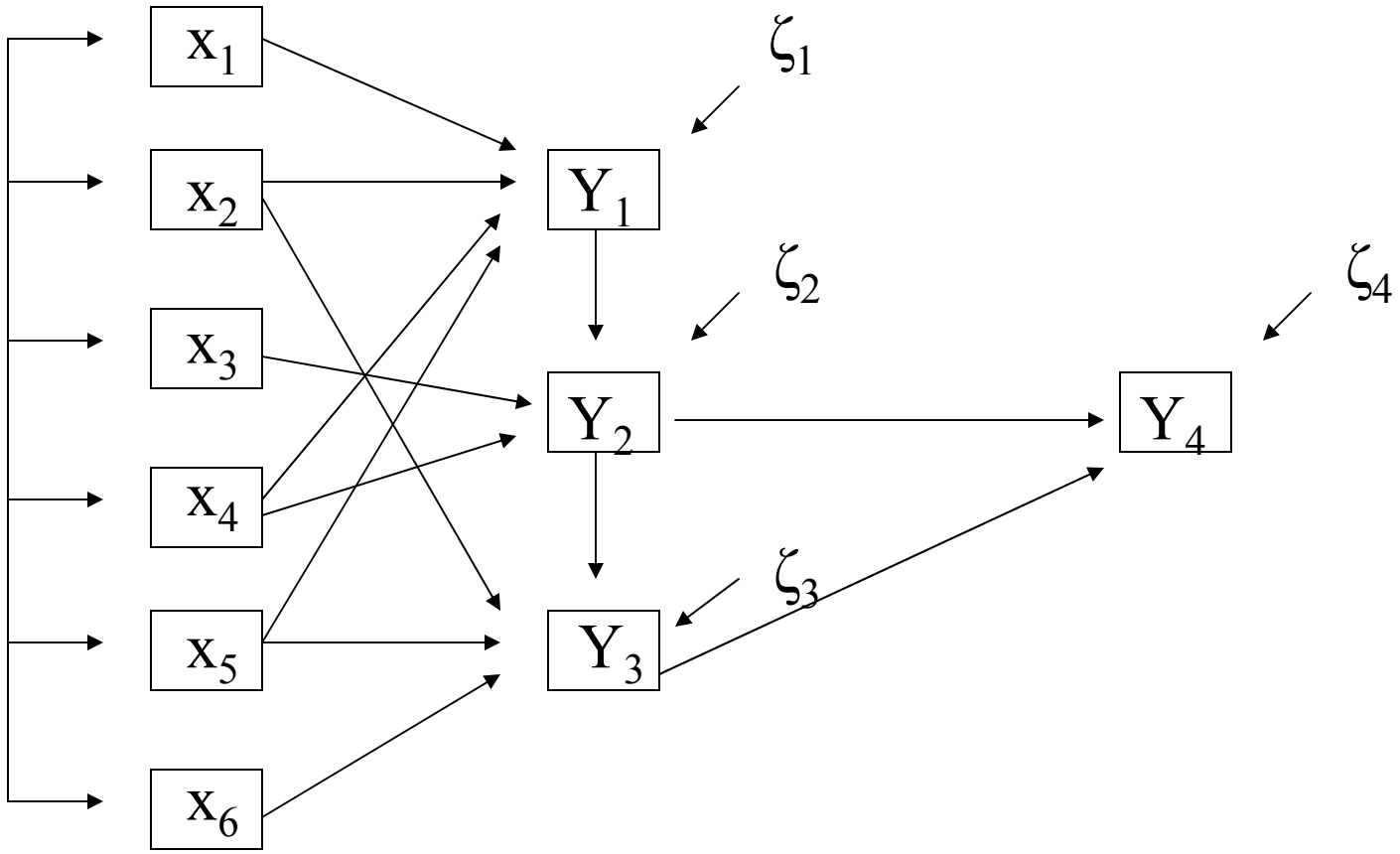
- Gamma (γ): Coefficient of association between an exogenous and endogenous variable
- Beta (β): Coefficient of association between two endogenous variables
- Zeta (ζ): Error term for endogenous variable
- Subscript protocol: first number refers to the 'destination' variable, while second number refers to the 'origination' variable.



# Path Diagram: Rules and Assumptions

- Endogenous variables are never connected by curved arrows
- Residual arrows point at endogenous variables, but not exogenous variables
- Causes are unitary (except for residuals)
- Causal relationships are linear

# Path Model: Mediation



# Path Diagram

Can you depict a path diagram for this system of equations?

$$y_1 = \gamma_{11}x_1 + \gamma_{12}x_2 + \beta_{12}y_2 + \zeta_1$$

$$y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \beta_{21}y_1 + \zeta_2$$

# Path Diagram

How about this one?

$$y_1 = \gamma_{12}x_2 + \beta_{12}y_2 + \zeta_1$$

$$y_2 = \gamma_{21}x_1 + \gamma_{22}x_2 + \zeta_2$$

$$y_3 = \gamma_{31}x_1 + \gamma_{32}x_2 + \beta_{31}y_1 + \beta_{32}y_2 + \zeta_3$$

# Wright's Rules for Calculating Total Association

- Total association: simple correlation between x and y
- For a proper path diagram, the correlation between any two variables = sum of the compound paths connecting the two variables
- Wright's Rule for a compound path:

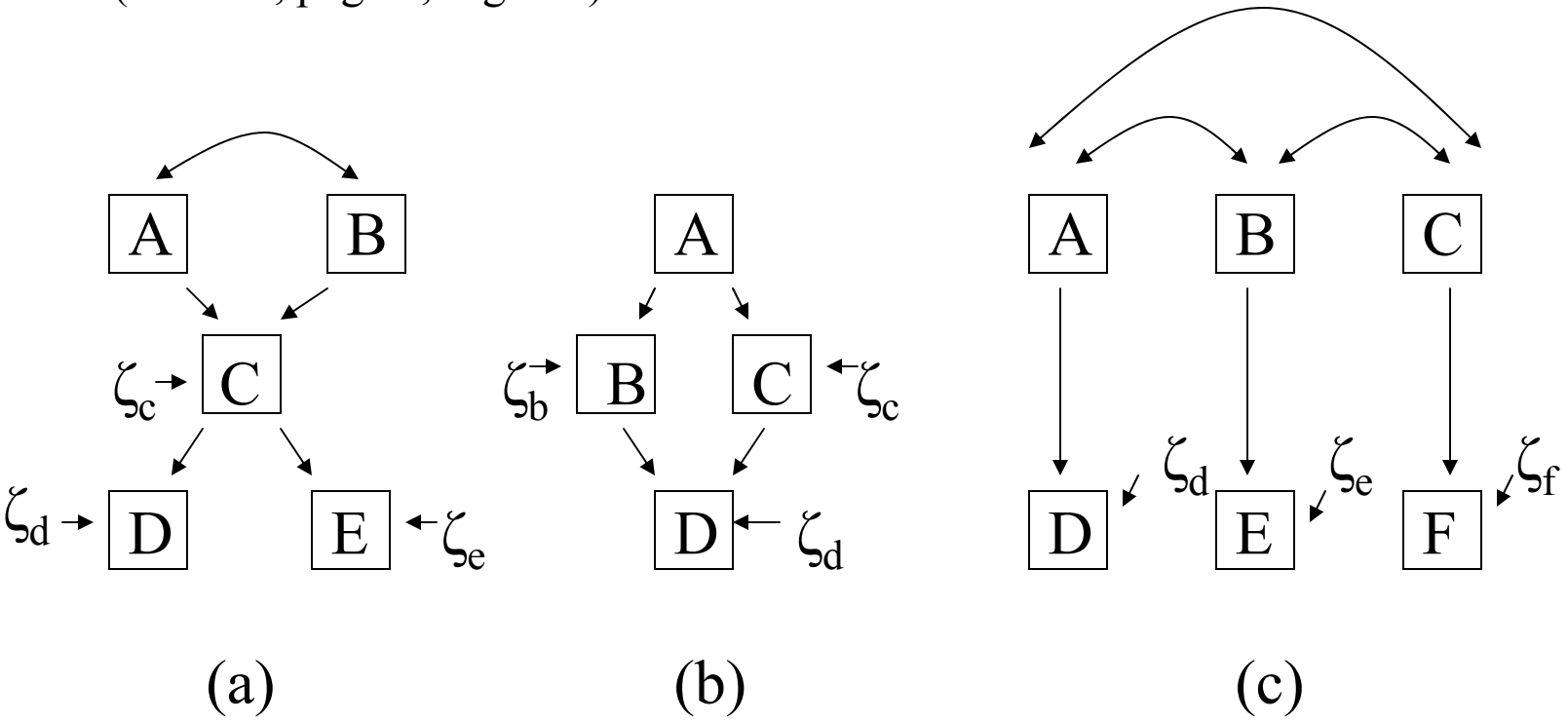
  1) no loops

  2) no going forward then backward

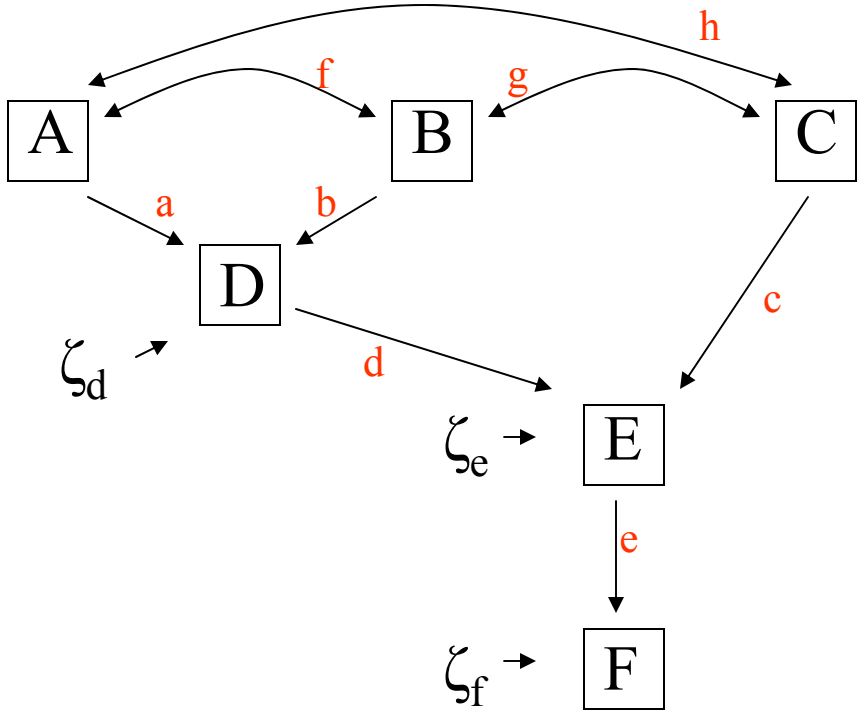
  3) a maximum of one curved arrow per path

# Example: compound path

(Loehlin, page 9, Fig. 1.6)



# Total Association: Calculation




Loehlin, page 10, Fig. 1.7

What is correlation between A and D? a+fb

What is correlation between A and E? ad+fbd+hc

What is correlation between A and F? ade+fbde+hce

What is correlation between B and F? gce+fade+bde

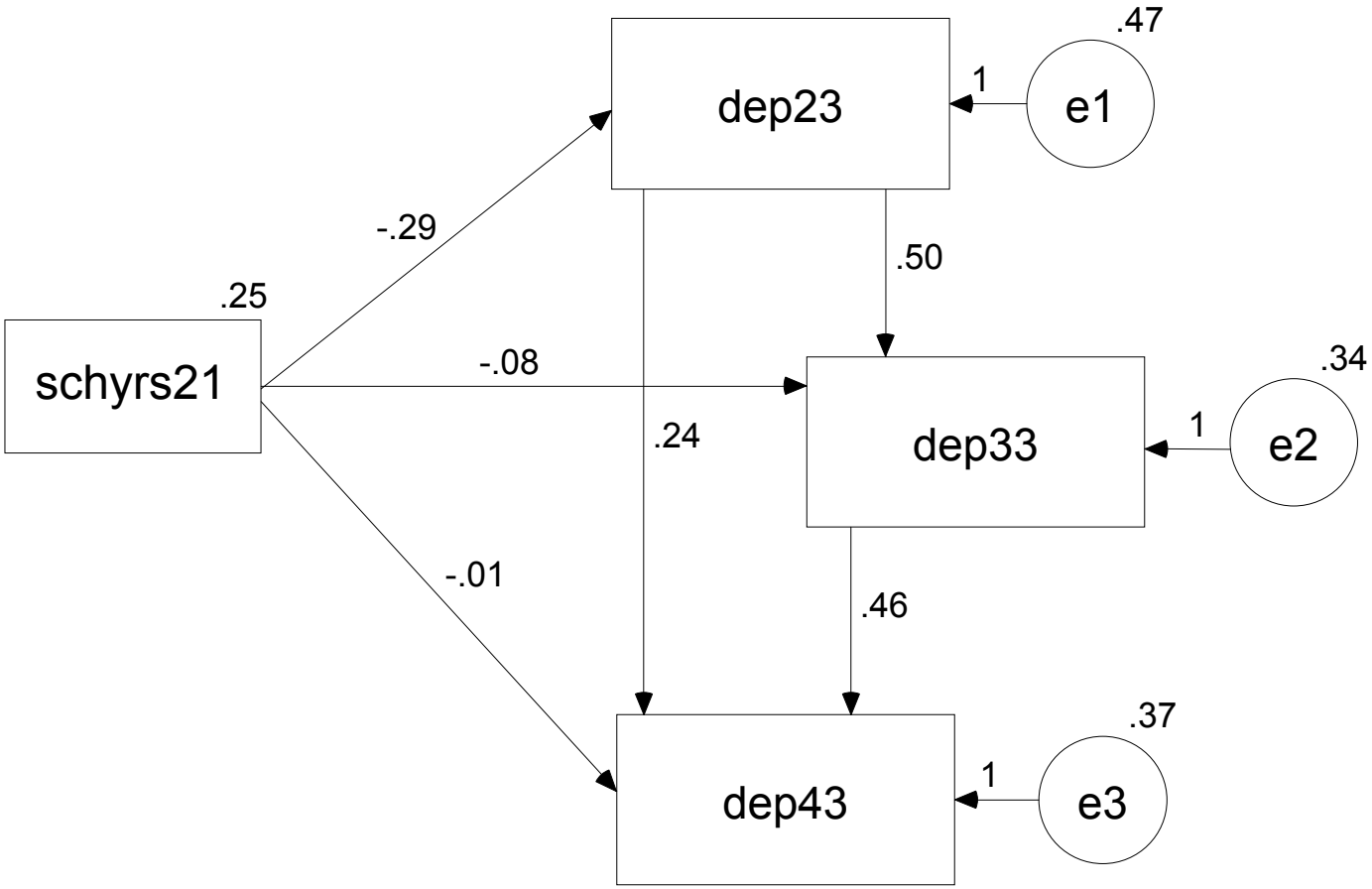
# Direct, Indirect, and Total Effects

- Path analysis distinguishes three types of effects:
- **Direct effects:** association of one variable with another net of the indirect paths specified in the model
- **Indirect effects:**
  - association of one variable with another mediated through other variables in the model
  - computed as the product of paths linking variables
- **Total effect:** direct effect plus indirect effect(s)
- Note: the decomposition of effects always is model-dependent!!!

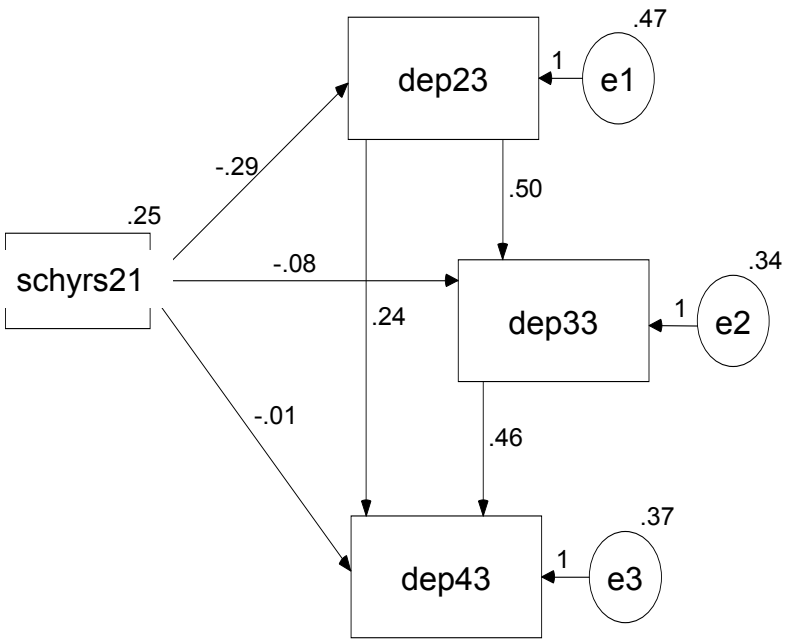
# Direct, Indirect, and Total Effects

- Example:
  - Does association of education with adult depression represent the influence of contemporaneous social stressors?
    - E.g., unemployment, divorce, etc.
    - Every longitudinal study shows that education affects depression, but not vice-versa
  - Use data from the National Child Development Survey, which assessed a birth cohort of about 10,000 individuals for depression at age 23, 33, and 43.

# Direct, Indirect, and Total Effects: Example



# Direct, Indirect, and Total Effects: Example




"schyrs21" on "dep33":

Direct effect = -0.08

Indirect effect = -0.29*0.50 =-0.145

Total effect = -0.08+(-0.29*0.50 )

What is the indirect effects of "schyrs21" on "dep43"?

# Path Analysis

- Key assumptions of path analysis:
- $E(\zeta i)=0$: mean value of disturbance term is 0
- $cov(\zeta_i, \zeta_j)=0$: no autocorrelation between the disturbance terms
- $var(\zeta_i|X_i)= \sigma^2$
- $cov(\zeta_i,X_i)=0$

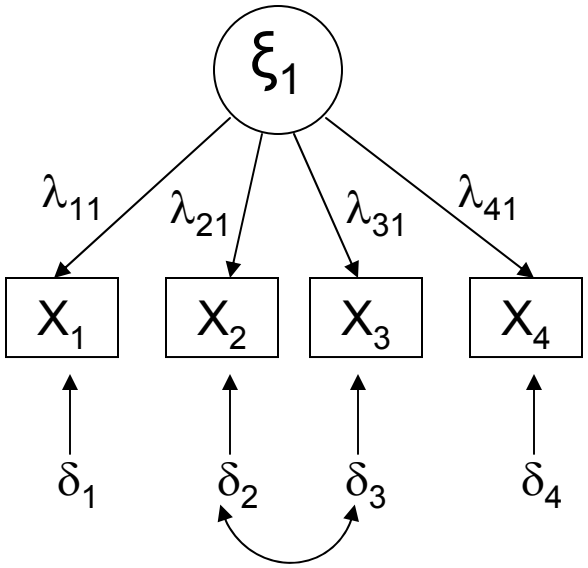
# The Fundamental hypothesis for SEM

- "The covariance matrix of the observed variables is a function of a set of parameters (of the model)" (Bollen)
- If the model is correct and parameters are known:

$$\Sigma=\Sigma(\theta)$$

where $\Sigma$ is the population covariance matrix of observed variables; $\Sigma(\theta)$ is the model-based covariance matrix written as a function of $\theta$

# Decomposition of Covariances and Correlations




$$X_1 = \lambda_{11}\xi_1 + \delta_1$$
$$X_2 = \lambda_{21}\xi_1 + \delta_2$$
$$X_3 = \lambda_{31}\xi_1 + \delta_3$$
$$X_4 = \lambda_{41}\xi_1 + \delta_4$$

or in matrix notation :

$$X = \Lambda_x \xi + \delta$$

$$\mathrm{E}(\delta) = 0, \mathrm{Var}(\xi) = \Phi, \mathrm{Var}(\delta) = \Theta_\delta,$$

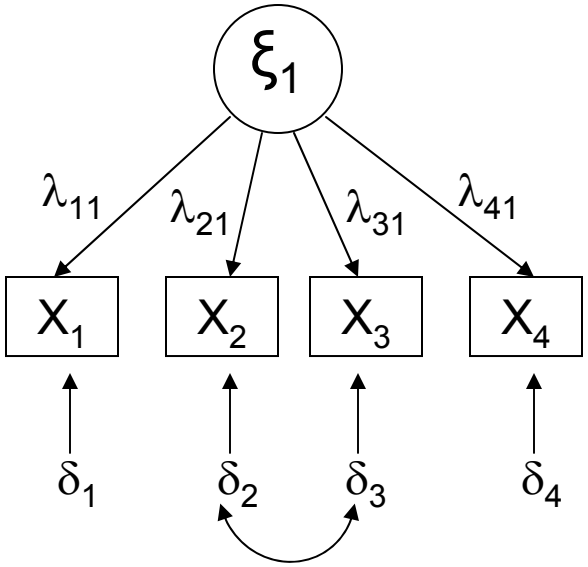
and $\mathrm{Cov}(\xi, \delta) = 0$

$$\mathrm{cov}(X_1, X_4) = \mathrm{cov}(\lambda_{11}\xi_1 + \delta_1, \lambda_{41}\xi_1 + \delta_4)$$

$$= \lambda_{11}\lambda_{41}\varphi_{11},$$

where $\varphi_{11} = \mathrm{var}(\xi_1)$

# Decomposition of Covariances and Correlations




$$X = \Lambda_x \xi + \delta$$

$$Cov(X) = \Sigma = E(XX')$$

$$XX' = (\Lambda_x \xi + \delta)(\Lambda_x \xi + \delta)'$$

$$= (\Lambda_x \xi + \delta)(\xi' \Lambda_x' + \delta')$$

$$= \Lambda_x \xi \xi' \Lambda_x' + \Lambda_x \xi \delta' + \delta \xi' \Lambda_x' + \delta \delta'$$

$$\Sigma = E(XX') = \Lambda_x \Phi \Lambda_x' + \Theta_\delta$$

Covariance matrix of ξ

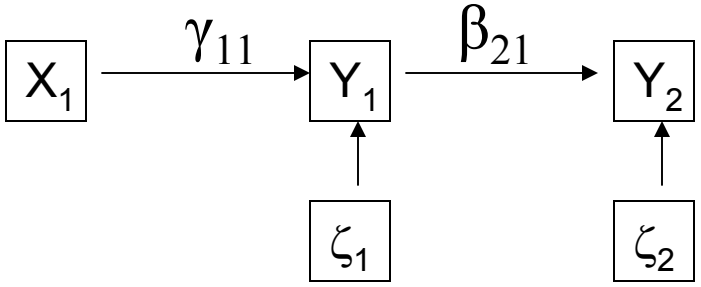
Covariance matrix of δ

# Path Model: Estimation

- Model hypothesis: $\Sigma = \Sigma(\theta)$
- But, we don't observe $\Sigma$, instead, we have sample covariance matrix of the observed variables: S
- Estimation of Path Models:
  - Choose $\hat{\theta}$ so that $\Sigma(\hat{\theta})$ is close to S

# Path Model: Estimation

1) Solve system of equations



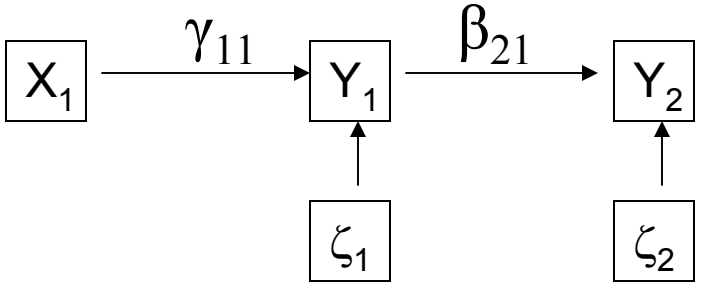

$$Y1 = \gamma 11 X1 + \zeta 1$$
$$Y2 = \beta 21 Y1 + \zeta 2$$

a) Using covariance algebra

$$\underbrace{\begin{bmatrix} var(Y_1) & & \\ cov(Y_2,Y_1) & var(Y_2) & \\ cov(X_1,Y_1) & cov(X_1,Y_2) & var(X_1) \end{bmatrix}}_{\Sigma} = \underbrace{\begin{bmatrix} \gamma_{11}^2 var(X_1) + var(\zeta_1) & & \\ \beta_{21}(\gamma_{11}^2 var(X_1) + var(\zeta_1)) & \beta_{21}^2(\gamma_{11}^2 var(X_1) + var(\zeta_1)) + var(\zeta_2) & \\ \gamma_{11} var(X_1) & \beta_{21}\gamma_{11} var(X_1) & var(X_1) \end{bmatrix}}_{\Sigma(\theta)}$$

# Path Model: Estimation

1) Solve system of equations




$$Y1 = \gamma 11 X1 + \zeta 1$$
$$Y2 = \beta 21 Y1 + \zeta 2$$

b) Using Wright's rules based on correlation matrix

$$\underbrace{\begin{bmatrix} 1 & & \\ \text{cor}(Y_2,Y_1) & 1 & \\ \text{cor}(X_1,Y_1) & \text{cor}(X_1,Y_2) & 1 \end{bmatrix}}_{\Sigma} = \underbrace{\begin{bmatrix} 1 & & \\ \beta_{21} & 1 & \\ \gamma_{11} & \beta_{21}\,\gamma_{11} & 1 \end{bmatrix}}_{\Sigma(\theta)}$$

# Path Model: Estimation

2) Write a computer program to estimate every single possible combination of parameters possible, and see which fits best (i.e. minimize a discrepancy function of $S-\Sigma(\theta)$ evaluated at $\hat{\theta}$)

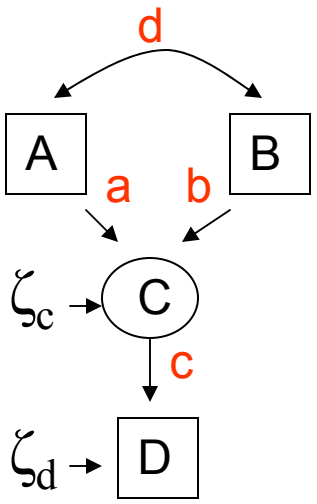
3) Use an iterative procedure (See Figure 2.2 on page 38 of Loehlin's Latent Variable Models)

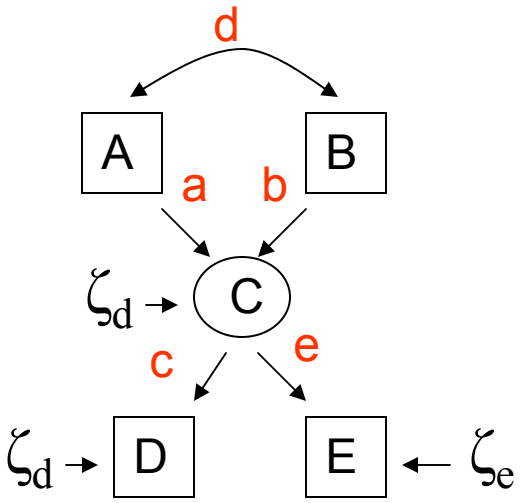
# Model Identification

- Identification: A model is identified if is theoretically possible to estimate one and only one set of parameters. Three helpful rules for path diagrams:
- "t-rule"
  - necessary, but not sufficient rule
  - the number of unknown parameters to be solved for cannot exceed the number of observed variances and covariances to be fitted
- analogous to needing an equation for each parameter
- number of parameters to be estimated: variances of exogenous variables, variances of errors for endogenous variables, direct effects, and double-headed arrows
- number of variances and covariances, computed as: n(n+1)/2, where n is the number of observed exogenous and endogenous variables

# Model Identification

- Just identified: # equations = # unknowns
- Under-identified: # equations < # unknowns
- Over-identified: # equations > # unknowns



# Model Fit Statistics

- **Goodness-of-fit tests based on predicted vs. observed covariances:**

1. **$\chi^2$ tests**
   - d.f.=(# non-redundant components in S) – (# unknown parameter in the model)
   - Null hypothesis: lack of significant difference between $\Sigma(\hat{\theta})$ and S
   - Sensitive to sample size
   - Sensitive to the assumption of multivariate normality
   - $\chi^2$ tests for difference between **NESTED** models

2. **Root Mean Square Error of Approximation** (RMSEA)
   - A population index, insensitive to sample size
   - No specification of baseline model is needed
   - Test a null hypothesis of poor fit
   - Availability of confidence interval
   - <0.10 “good”, <0.05 “very good” (Steiger, 1989, p.81)

3. **Standardized Root Mean Residual** (SRMR)
   - Squared root of the mean of the squared standardized residuals
   - SRMR = 0 indicates “perfect” fit, < .05 “good” fit, < .08 adequate fit

# Model Fit Statistics

- Goodness-of-fit tests comparing the given model with an alternative model

1. **Comparative Fit Index** (CFI; Bentler 1989)
   - compares the existing model fit with a null model which assumes uncorrelated variables in the model (i.e. the "independence model")
   - Interpretation: % of the covariation in the data can be explained by the given model
   - CFI ranges from 0 to 1, with 1 indicating a very good fit; acceptable fit if CFI>0.9
2. **The Tucker-Lewis Index** (TLI) or **Non-Normed Fit Index** (NNFI)
   - Relatively independent of sample size (Marsh et al. 1988, 1996)
   - NNFI >= .95 indicates a good model fit, <0.9 poor fit

- More about these later