Covariance Structure Analysis: Introduction

Y645

Leslie Rutkowski

Nice Resources

- David Kenny's website
  - http://davidakenny.net/cm/causalm.htm
- Kristopher Preacher's website:
  - https://www.quantpsy.org/medn.htm
- Jason Newsom's website:
  - https://web.pdx.edu/~newsomj/semclass/

- For R
  - http://lavaan.ugent.be/
  - UCLA IDRE (stats.idre.ucla.edu/r/seminars/rsem/)
  - R Google group
  - lavaan Google group

- Supplementary text:
  - Beaujean, A.A. (2014). *Latent variable modeling using* R. New York: Routledge.

SEM – A rose by *many* names

- Structural equation modeling
- Covariance structure analysis
- Structural modeling
- LISREL (linear structural relations) modeling
- Latent variable modeling

- In this class, we'll primarily use SEM

What is SEM?

- Very broadly – it is a modeling method used by social and natural scientists to uncover relationships in data.

- Interesting features:
  – Accounts for measurement error
  – Can build rich models involving *latent constructs*
  – Explicit test of model fit
  – Models are fit to covariance/correlation matrices

- You can cast almost any model in an SEM framework: regression, multilevel models, IRT, etc.

Latent construct

- Typically, a theoretical or hypothetical construct.
- Normally unobservable
- Examples:
  - Proficiency (math, language, athletic)
  - Attitudes (toward learning, crime, spending)
  - Perceptions (of others, of behavior, of laws)
  - *Many* others

What *is* SEM?

- An analytical approach that allows a researcher to build an elegant and parsimonious model of the processes that give rise to the observed data (Little, 2013)

SEM

- We usually focus on cases or individual observations
  - Subjects in a study
  - Students taking a test
- We try to model these observations

- In SEM, our focus is different
  - First, let's briefly review OLS regression

Review of OLS Regression

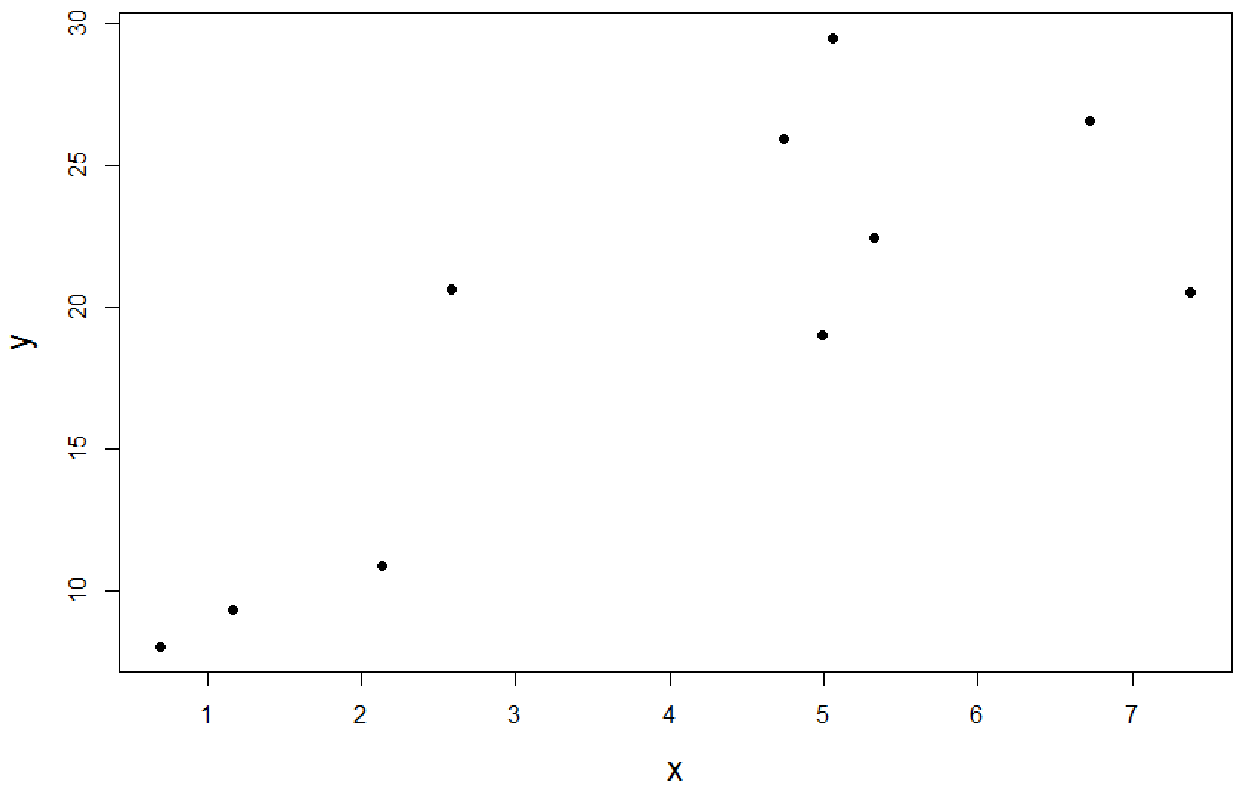
- We hope to describe relationships on a straight line
- We collect observations on many units
- Usually: 1 variable acts as a response/outcome and 1 variable acts as a predictor
- We view outcome (Y) as $f(X)$

- In particular

$$f(x_i) = \beta_0 + \beta_1 x_i + e_i$$

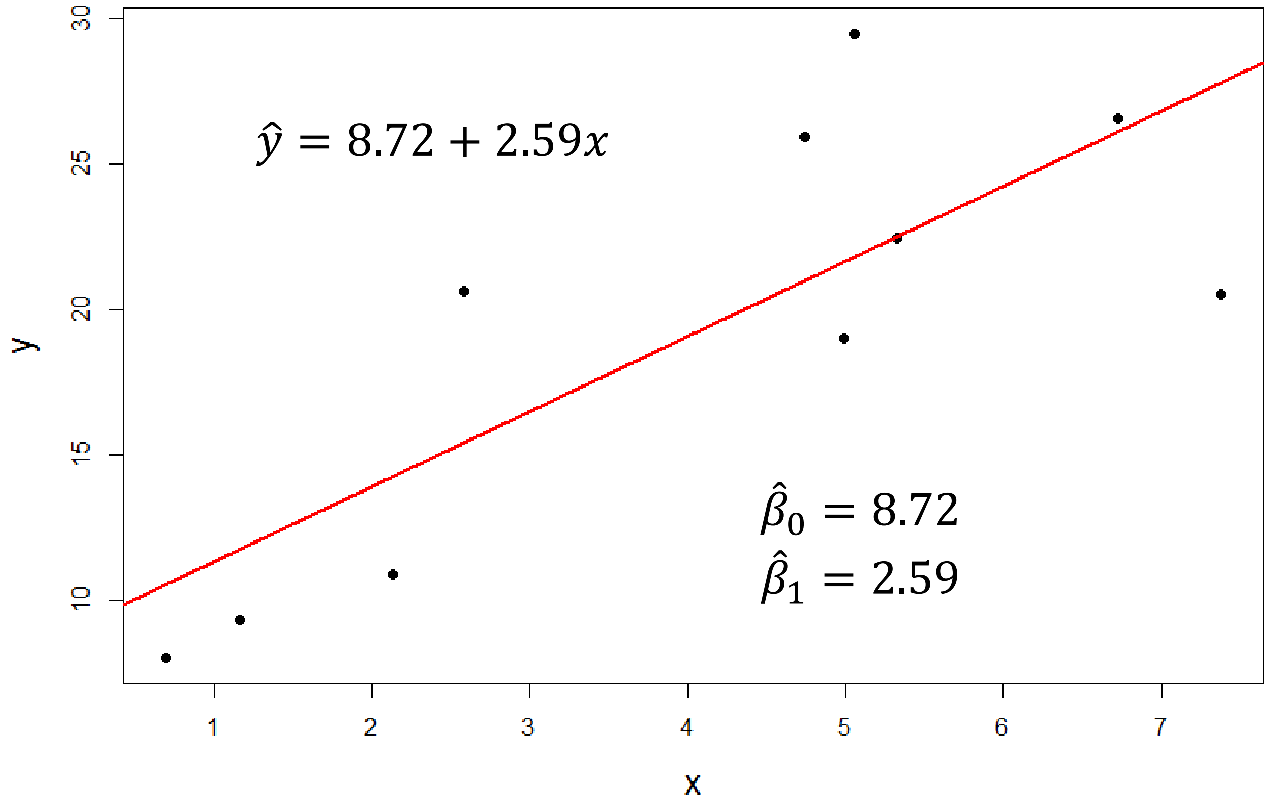
- The hypothesized model is such that the predictor, x_i, is set or known by data collector and Y_i is a $f(xi)$.
- The hypothesized model specifies, except for some unknown parameters, the response behavior for given x values
- Model also characterizes failure to provide exact fit through error terms.
- Then we use the data to estimate unknown parameters

- We observe:



- Notation:
  - X, Y are random variables
  - x_i, yi are observed values for the i^{th} case
- Equation of a line: $Y = \beta_0 + \beta_1 X$
  - Parameters
    - β_0 – intercept or value of Y when X is 0.
    - β_1 – rate of change in Y for 1 unit change in X.

Then we have…


$\hat{y} = 8.72 + 2.59x$
$\hat{\beta}_0 = 8.72$
$\hat{\beta}_1 = 2.59$
y
x
10
15
20
25
30
1
2
3
4
5
6
7

OLS Regression: Errors

- Real data don't fall on a straight line
- Instead, there are statistical errors
- Random errors
  - Measurement error (in Y)
  - Omitted variables
  - Natural variability
- e_i – statistical error for i^{th} case ($i = 1, \ldots, n$)

OLS Assumptions

- $\mathrm{E}(e_i) = 0$ for $i = 1, \ldots, n$
- $\mathrm{cov}(e_i, e_j) = 0$ for all $i \neq j$
  - Errors are mutually independent and uncorrelated
- $\mathrm{var}(e_i) = \sigma^2$
  - Common though usually unknown variance
- Then $e_1 \sim N(0, \sigma^2), i = 1, \ldots, n$

Simple regression model

- The model is given by:

$$y_i = \beta_0 + \beta_1 x_i + e_i, i = 1, \ldots, n$$

$$\mathrm{E}(e_i) = 0$$

$$\mathrm{var}(e_i) = \sigma^2$$

$$\mathrm{cov}(e_i, e_j) = 0 \quad i \neq j$$

- How many unknowns?
  - 3: $\beta_0, \beta_1, \& \ \sigma^2$
  - e_i are unobservable quantities to account for failure of observed values to fall on a straight line. Only x_i and y_i are observed

Notation

- Parameters: $\alpha, \beta, \gamma, \sigma$
- Estimators: $\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\sigma}$

Fitting Errors vs. Statistical Errors

- While e_is are not parameters, we use $\hat{e}_i$ to describe “observed fitting errors" or residuals:

$$\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad i = 1, 2, \ldots, n$$

- Statistical error (not observable):

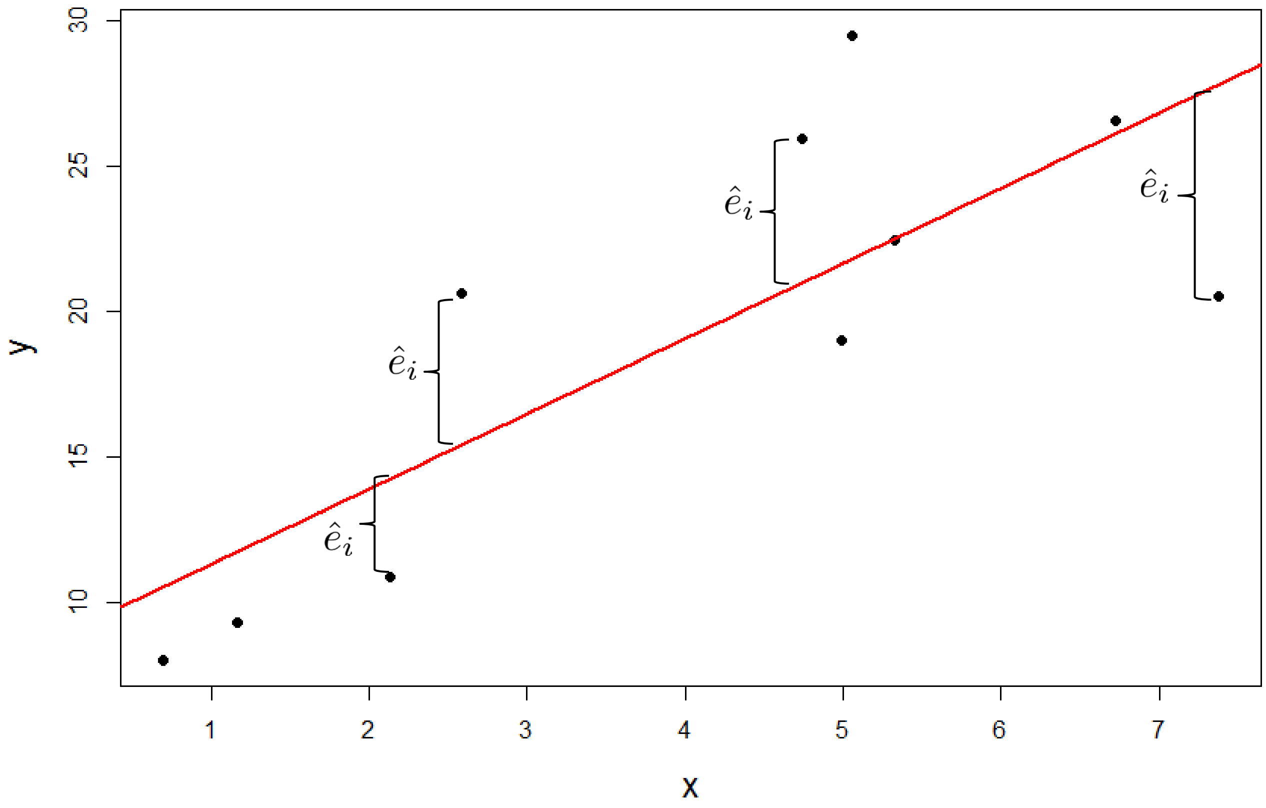
$$e_i = y_i - (\beta_0 + \beta_1 x_i) \quad i = 1, 2, \ldots, n$$

Fitted Values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad i = 1, 2, \ldots, n$$

- Notice: $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$
  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- Then: $\hat{e}_i = y_i - \hat{y}_i$

Errors



LS Estimators

- Those values: $\hat{\beta}_0$ of β_0 and $\hat{\beta}_1$ of β_1
- That minimize: $RSS(\beta_0, \beta_1) = \Sigma[y_i - (\beta_0 + \beta_1 x_i)]^2$

$$\hat{\beta}_1 = \frac{SXY}{SXX} = r_{xy}\frac{SD_y}{SD_x} = r_{xy}\left(\frac{SYY}{SXX}\right)^{\frac{1}{2}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

and

$$\hat{\sigma}^2 = \frac{RSS}{n-2} \quad \text{where} \quad RSS = SYY - \frac{(SXY)^2}{SXX} = SYY - \hat{\beta}_1{}^2 SXX$$

Matrix Notation

$\boldsymbol{x}, \boldsymbol{e}, \boldsymbol{\beta}$ ← Vectors and matrices

x_{ij}, e_i, β_j ← Elements of a vector or matrix

Multiple Predictors & Matrix Notation

$$\boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$\boldsymbol{Y}$ and $\boldsymbol{e}$ are n x 1 vectors

$\boldsymbol{\beta}$ is $(p+1) \times 1$

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

and $p' =$ no. columns in $\boldsymbol{X}$

$\boldsymbol{X}$ is $n \times (p+1)$:

OLS Regression

- Then we have

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times p'}{\boldsymbol{X}}\ \underset{p'\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{e}}$$

$$\boldsymbol{e} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$$

where **0** is $n \times 1$ and I is an $n \times n$ identity matrix

$$E(\boldsymbol{e}) = 0$$
$$\text{var}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I} n$$

Then the estimator $\hat{\beta} = \big(\underset{p' \times n}{\boldsymbol{X}^T} \underset{n \times p'}{\boldsymbol{X}}\big)^{-1} \underset{p' \times n}{\boldsymbol{X}^T} \underset{n \times 1}{\boldsymbol{Y}}$

minimizes $RSS(\boldsymbol{\beta}) = \sum (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$

The important thing to note here is that the estimator deals with ***cases / observations***

A change in focus: Covariance instead of cases

- Instead of minimizing $RSS(\boldsymbol{\beta})$ we want to minimize difference between sample covariance and model predicted covariance

$$\text{obs covar} - \widehat{\text{covar}} = \text{residuals}$$

- H_0: Observed covariance is a function of a set of parameters *if* model is correct and parameters are known, then residuals = **0**

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta}) - \text{ Fundamental Equation}$$

$\boldsymbol{\Sigma}$ - population covariance matrix of observed variables

$\boldsymbol{\theta}$ - vector of model parameters

$\boldsymbol{\Sigma}(\boldsymbol{\theta})$ - population covariance matrix as a function of model parameters

I cannot stress enough how important this equation is.
We will return to it again and again and again.

Simple regression:

$y = \gamma x + \zeta$ where y, x, and ζ are random variables and $E(\zeta)=0$

Then: $\qquad \mathbf{\Sigma} \qquad = \qquad \mathbf{\Sigma}(\theta)$

$$\begin{bmatrix} \mathrm{VAR}(y) & \\ \mathrm{COV}(x,y) & \mathrm{VAR}(x) \end{bmatrix} = \begin{bmatrix} \gamma^2\mathrm{VAR}(x) + \mathrm{VAR}(\zeta) & \\ \gamma\mathrm{VAR}(x) & \mathrm{VAR}(x) \end{bmatrix}$$

where: $\mathrm{VAR}(y) = \gamma^2\mathrm{VAR}(x) + \mathrm{VAR}(\zeta)$ because $\mathrm{VAR}(\gamma x) = \gamma^2\mathrm{VAR}(x)$

and: $\mathrm{COV}(x,y) = \gamma\mathrm{VAR}(x)$ because $\gamma = \frac{SXY}{SXX}$ and $\mathrm{VAR}(x) = \frac{SXX}{n-1}$

Then: $\frac{SXY}{SXX} * \frac{SXX}{n-1} = \frac{SXY}{n-1} = \mathrm{COV}(x,y)$

Here, θ has 3 parameters: γ , VAR(x), and VAR(ζ)

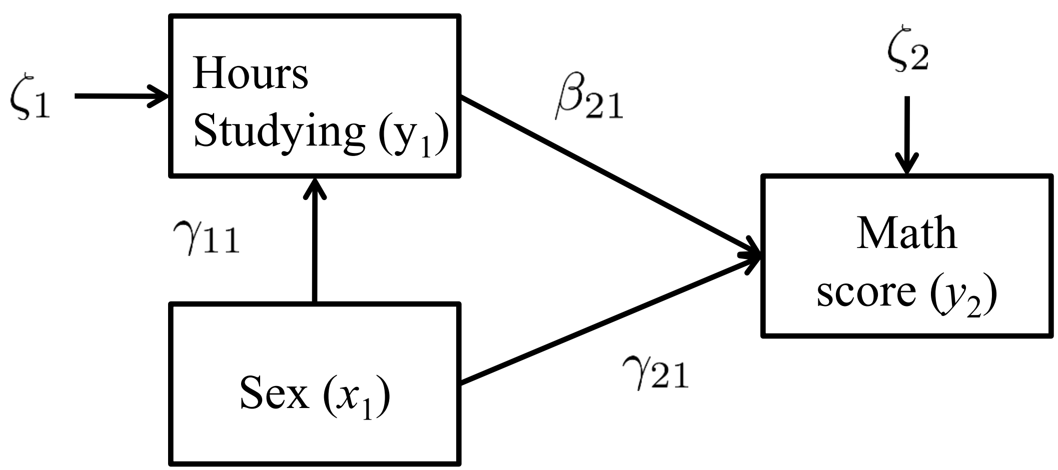
Types of SEMs

1. Path analysis models
   – Usually involves only observed variables

Path analysis

(Wright, 1918, 1921, 1934, 1960)

1) Path Diagram



<u>One way</u> "causal" influences from variable at base to point

Path Analysis cont.

2) Equations

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

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

Path Analysis continued

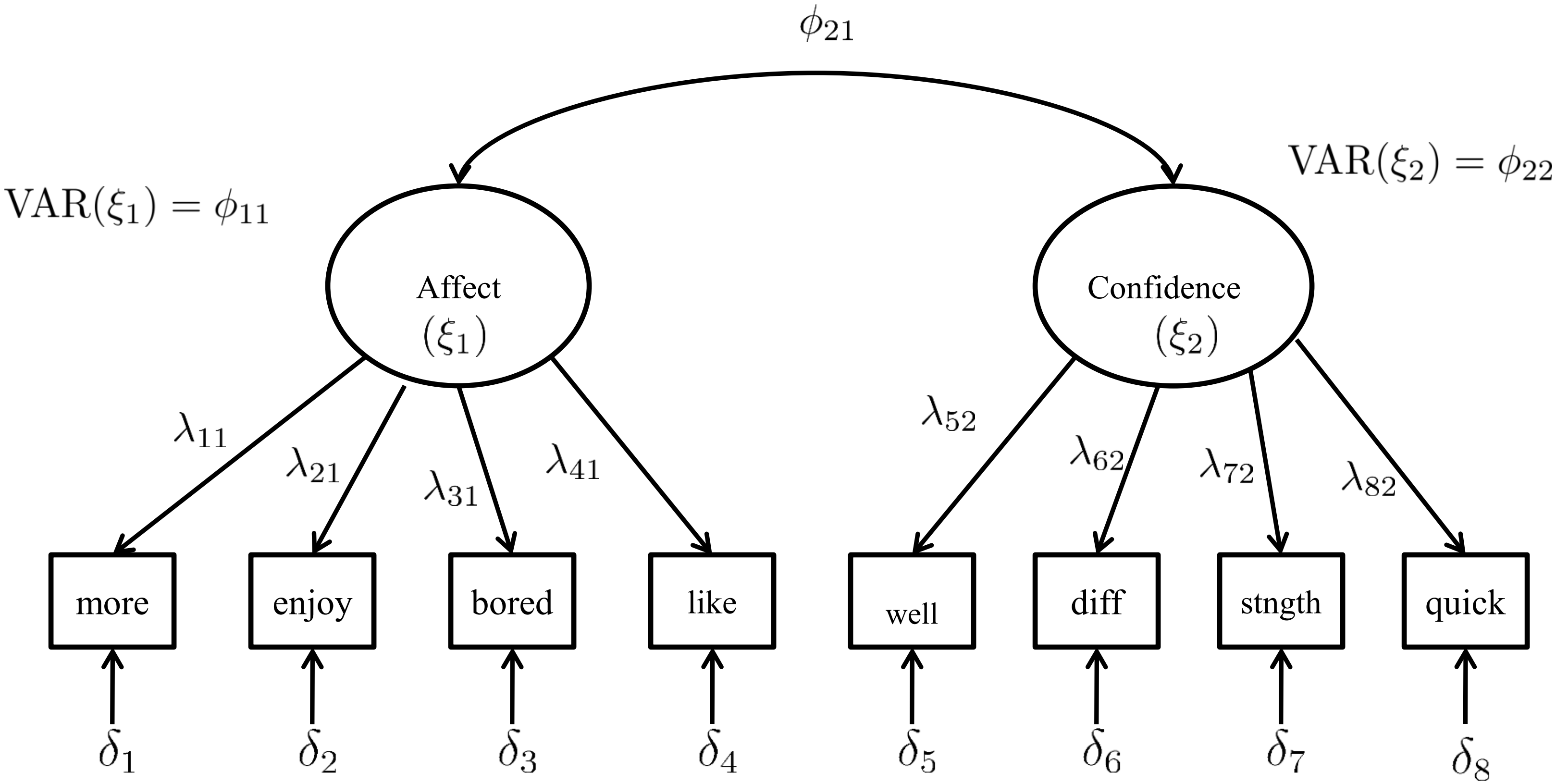
3) Can distinguish direct, indirect, total } effects

Types of SEMS

2. Confirmatory factor analysis / measurement model
    1. Involves one or more latent variables and the way in which they relate to manifest/observed variables
    2. No specified relationship between latent variables
    3. Useful for scale development

Confirmatory factor analysis

- Two factors:



Confirmatory Factor Analysis

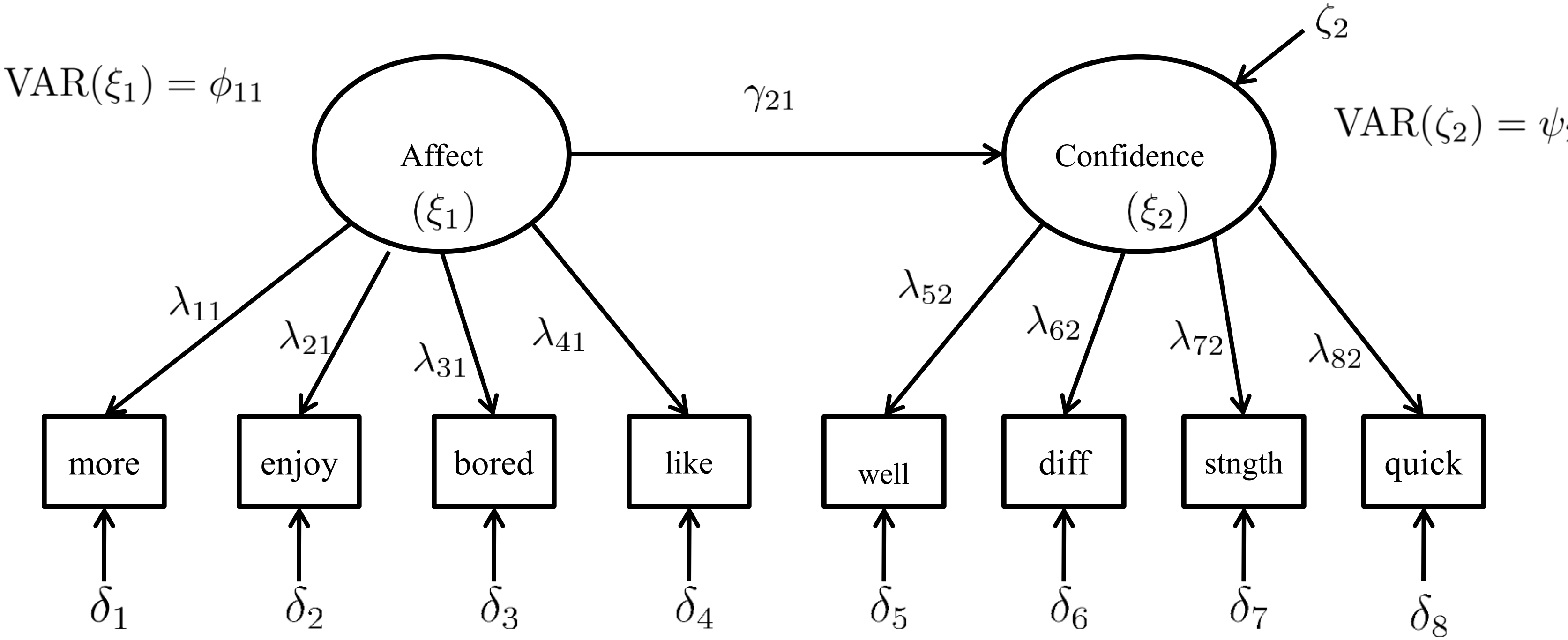
- Describes structure of data in terms of relationships with latent variables
- Lots of assumptions underlying – we’ll discuss in a few weeks.

Types of SEMS

3. Structural models
   – Somewhat similar to measurement models except relationships between latent variables are also postulated.
   – Particularly well suited for testing hypotheses about relationships among latent variables.

Structural model

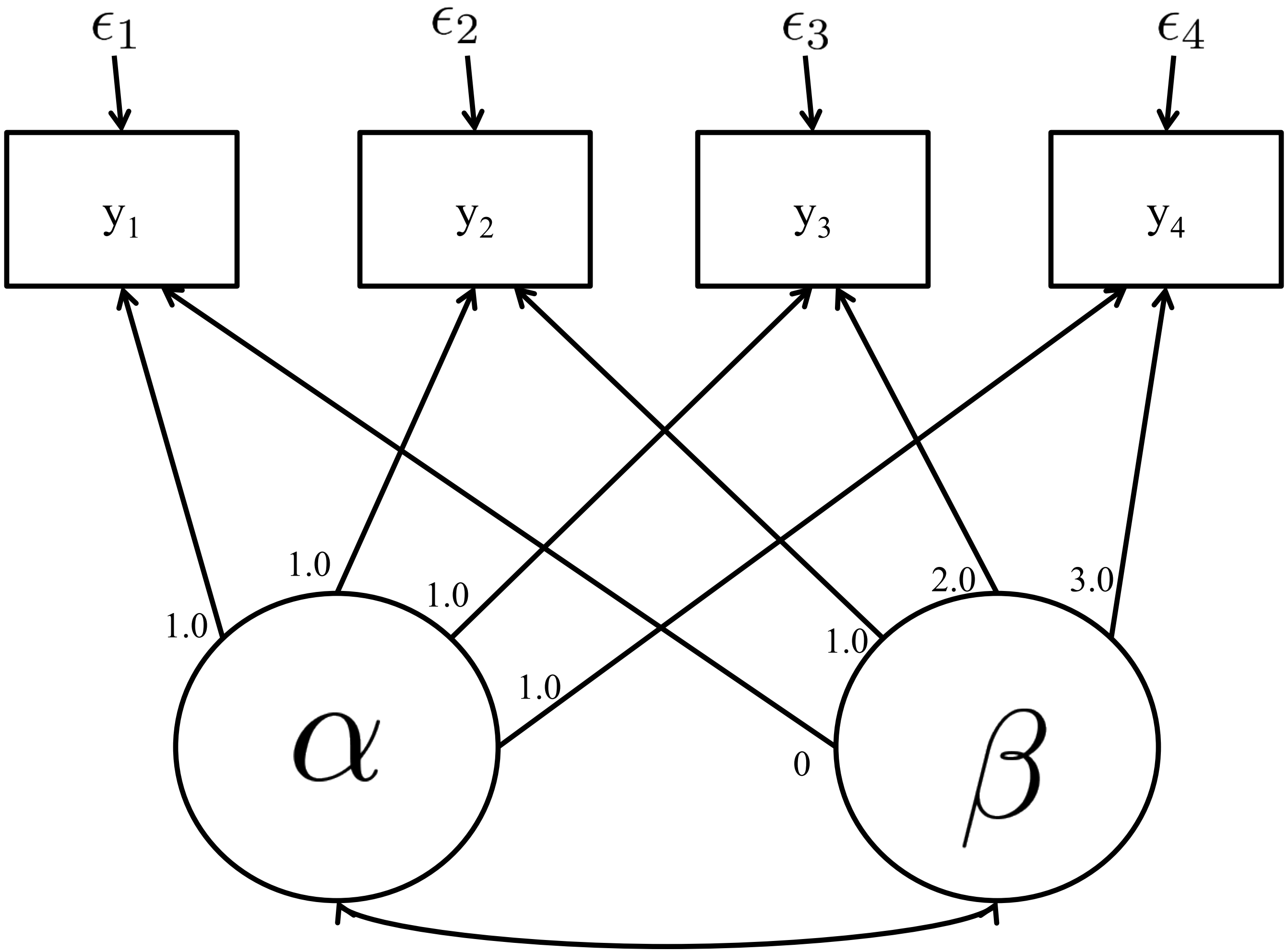
- Two factors: *Affect* predicts *Confidence*



Types of SEMs

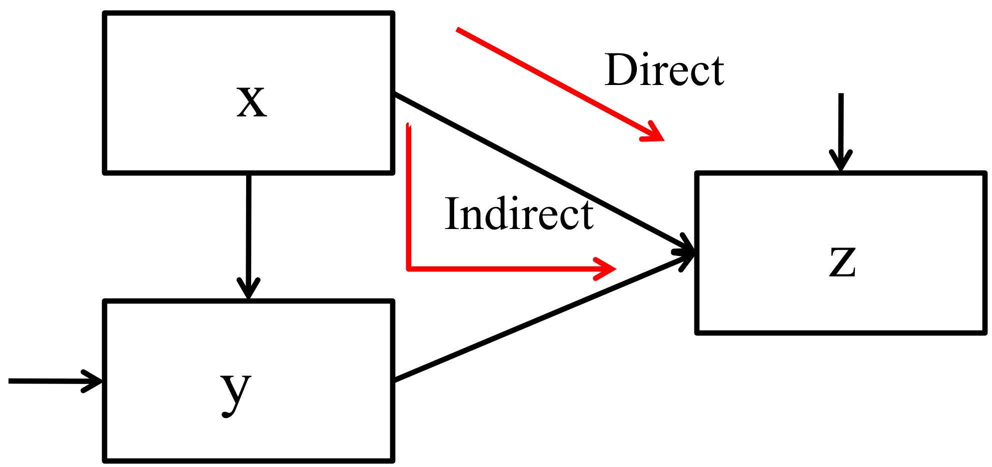
4. Latent growth/change models
   – Intended to study change over time

Growth model



When are SEMs useful?

- To test theories about phenomena
- To validate measures
- Theory development (exploratory)
- For studying direct & indirect effects



Systems of Linear Equations

- Linear in the observed, latent, and disturbance variables
- Not necessarily linear in the covariance structure equations

Evolution

- From path analysis to LISREL model
  - LISREL: linear structural relations (Jöreskog)

- We use "LISREL" notation
- There are others:
  - Bentler & Weeks (1980)
  - McArdle & McDonald (1984)

On your own time

- Review of matrix algebra. Be sure to have a look at:
    - Bollen Appendix A
    - Your old Y604 notes on linear algebra
- For a general orientation:
    - MacCallum, R. & Austin, J. (2000). Applications of structural equation modeling in psychological research. *Annual review of Psychology, 51*(1), 201-227.
    - Weston, R. & Gore, P. (2006). A brief introduction to structural equation modeling. *The Counseling Psychologist, 34*(5), 719-751.

1st Exercise

- Matrix/linear algebra exercise
  - Details are on Canvas
  - You are free to work with **one** partner, but please only turn in one joint submission on Canvas.

Next: Notation & Path Analysis

- Model notation
  - Get friendly with the Greek alphabet!
  - There is a Greek alphabet saved in the Class 2 folder on Canvas.
- See the syllabus for upcoming readings