An Overview of STRUCTURAL EQUATION MODELS WITH LATENT VARIABLES

Kenneth A. Bollen
Odum Institute for Research in Social Science
Department of Sociology
University of North Carolina
at Chapel Hill

Presented at the Miami University Symposium on Computational Research - March 1-2, 2007, Miami University, Oxford, OH.

I. What are Structural Equation Models? (SEM)

II. Fundamental Hypothesis [$\Sigma = \Sigma(\theta)$]

III. Illustrations

IV. Three Common Types of SEM

V. Overview of Modeling Steps

VI. SEM with Means and Intercepts

VII. Software

VIII. Empirical Example

I. What are SEM?

General Statistical Model

Special Cases:

ANOVA & ANCOVA
Multiple Regression
Econometric Models
Path Analysis
Factor Analysis
Etc.

SEM Allow:

1. Latent & Observed Variables
2. Random & Nonrandom Errors
3. Errors-in-Variables Regressions
4. Multiple Indicators
5. Restrictions on Parameters
6. Test of Model Fit
7. Nonnormal Variables

II. Fundamental Hypothesis

H_o: $\Sigma = \Sigma(\theta)$

Σ = Population Covariance Matrix

θ = Vector of Parameters

$\Sigma(\theta)$ = Model Implied Covariance Matrix

III. Illustrations

A. Simple Regression as a SEM

$$y = \gamma x + \zeta$$

III.A.

$$\Sigma = \begin{bmatrix} \mathbf{VAR\ (y)} & \\ \mathbf{COV\ (x,y)} & \mathbf{VAR\ (x)} \end{bmatrix}$$

$$\theta' = [\gamma \quad \mathbf{VAR(x)} \quad \mathbf{VAR(\zeta)}]$$

$$\Sigma\ (\theta) = \begin{bmatrix} \gamma^2 \mathrm{VAR}(x) + \mathrm{VAR}(\zeta) & \\ \gamma \mathrm{VAR}(x) & \mathrm{VAR}(x) \end{bmatrix}$$

III.A.

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

$$\begin{bmatrix} VAR(y) & \\ COV(x,y) & \mathrm{VAR}(\mathbf{x}) \end{bmatrix} =$$

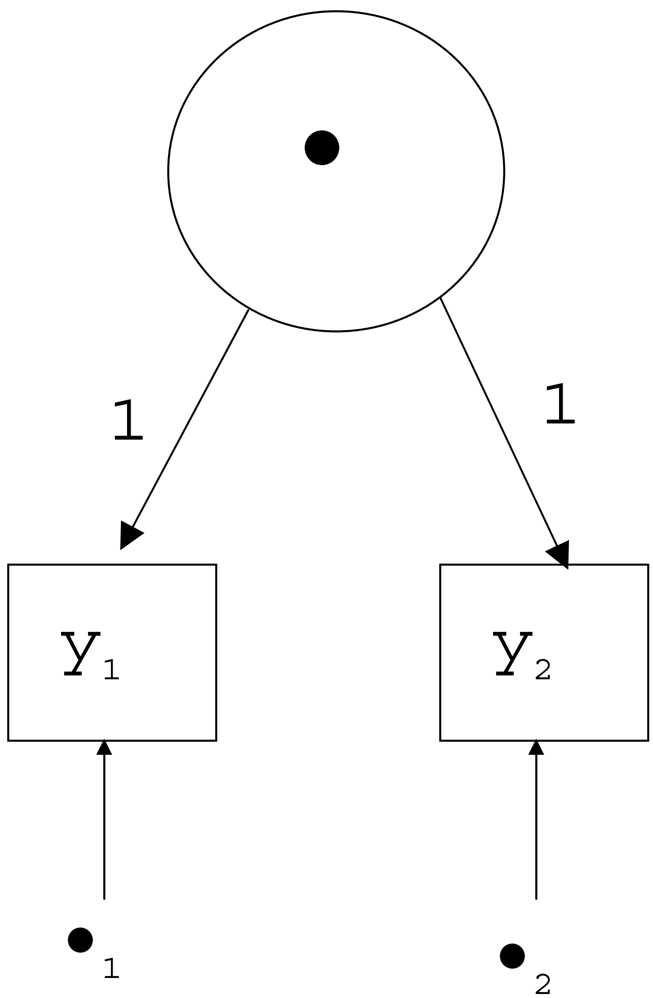
$$\begin{bmatrix} \gamma^2\,\mathrm{VAR}(\mathbf{x}) + \mathrm{VAR}(\zeta) & \\ \gamma\,VAR(x) & \mathrm{VAR}(\mathbf{x}) \end{bmatrix}$$

e.g., $COV(x,y) = COV(x,\ \gamma\mathbf{x} + \zeta)$

$= \gamma COV(x,x) + COV(x,\zeta)$

$= \gamma VAR(x)$

III.B. Simple Factor Analysis



$$y_1 = \eta + \varepsilon_1$$

$$y_2 = \eta + \varepsilon_2$$

III.B.

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

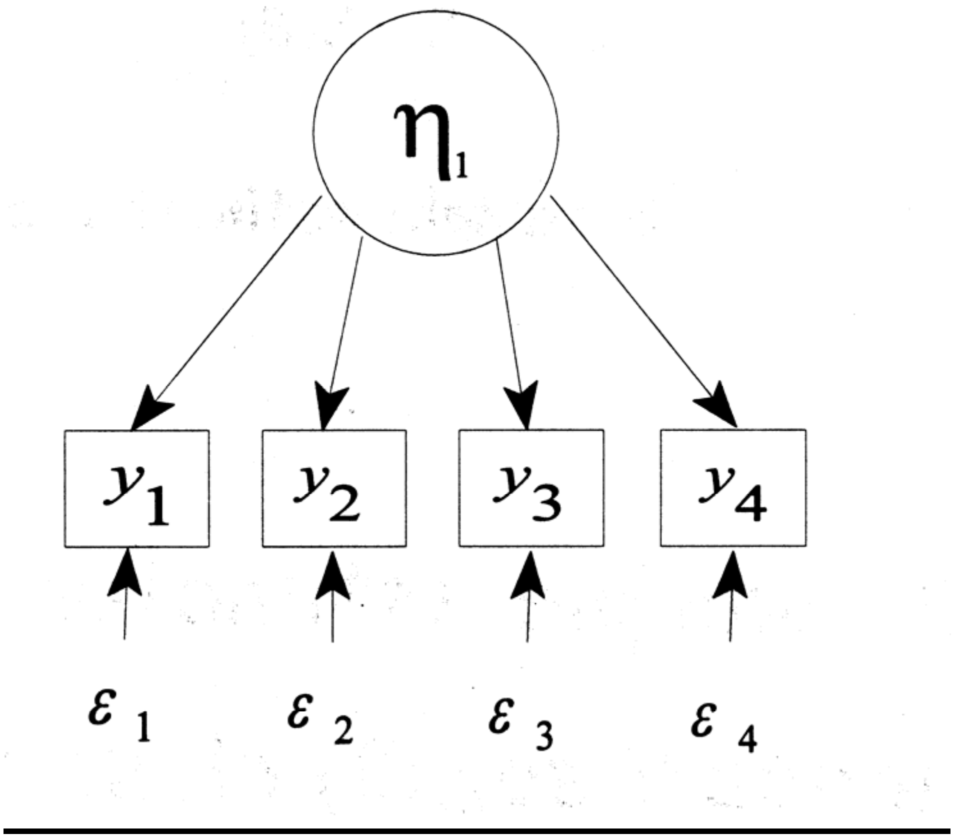
$$\begin{bmatrix} VAR(y_1) & \\ COV(y_1, y_2) & VAR(y_2) \end{bmatrix} = \begin{bmatrix} VAR(\eta)+VAR(\varepsilon_1) & \\ VAR(\eta) & VAR(\eta)+VAR(\varepsilon_2) \end{bmatrix}$$

e.g., $COV(y1,y2)=COV(\eta + \varepsilon_1, \eta + \varepsilon_2)$

$=COV(\eta,\eta)$

$=VAR(\eta)$

Confirmatory Factor Analysis: Air Quality Example



$$y_1 = \eta_1 + \varepsilon_1$$
$$y_2 = \lambda_{21}\eta_1 + \varepsilon_2$$
$$y_3 = \lambda_{31}\eta_1 + \varepsilon_3$$
$$y_4 = \lambda_{41}\eta_1 + \varepsilon_4$$

Latent Variables

Variables of Interest But Not Directly Measured

Common in Sciences:

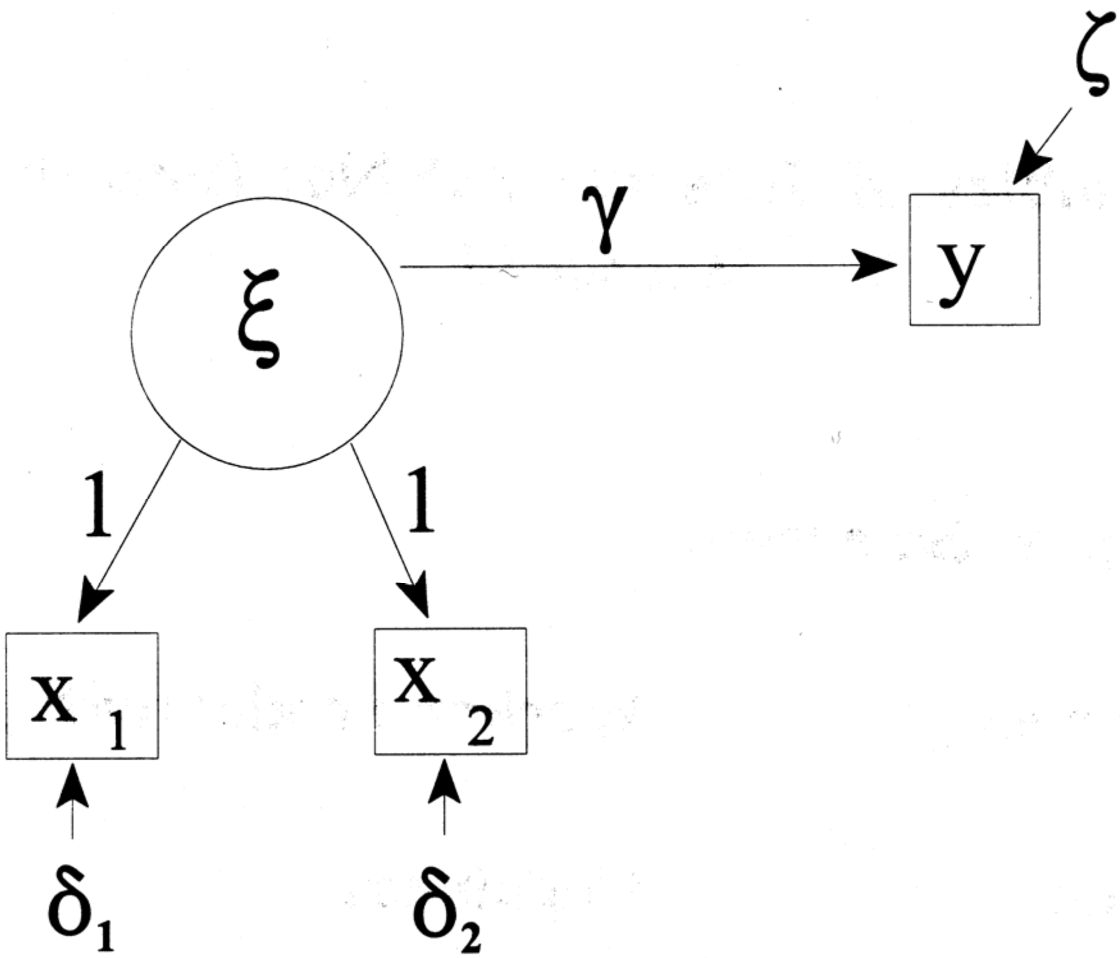
Intelligence, Worker Productivity,

Diseases, Happiness,

Value of House, Carrying Capacity,

"Free" Market, Disturbance Variables

III.C. Simple General Model



$$\mathbf{y} = \gamma\xi + \zeta$$
$$\mathbf{x}_1 = \xi + \delta_1$$
$$\mathbf{x}_2 = \xi + \delta_2$$

III.C.

$$\Sigma = \begin{bmatrix} \mathrm{VAR}(y) & & \\ \mathrm{COV}(x_1, y) & \mathrm{VAR}(x_1) & \\ \mathrm{COV}(x_2, y) & \mathrm{COV}(x_2, x_1) & \mathrm{VAR}(x_2) \end{bmatrix}$$

$$\Sigma(\theta) =$$

$$\begin{bmatrix} \gamma^2\mathrm{VAR}(\xi) + \mathrm{VAR}(\zeta) & & \\ \gamma\mathrm{VAR}(\xi) & \mathrm{VAR}(\xi) + \mathrm{VAR}(\delta_1) & \\ \gamma\mathrm{VAR}(\xi) & \mathrm{VAR}(\xi) & \mathrm{VAR}(\xi) + \mathrm{VAR}(\delta_2) \end{bmatrix}$$

IV. Three Types of SEM

A. Classical Econometric

1. Multiequation System

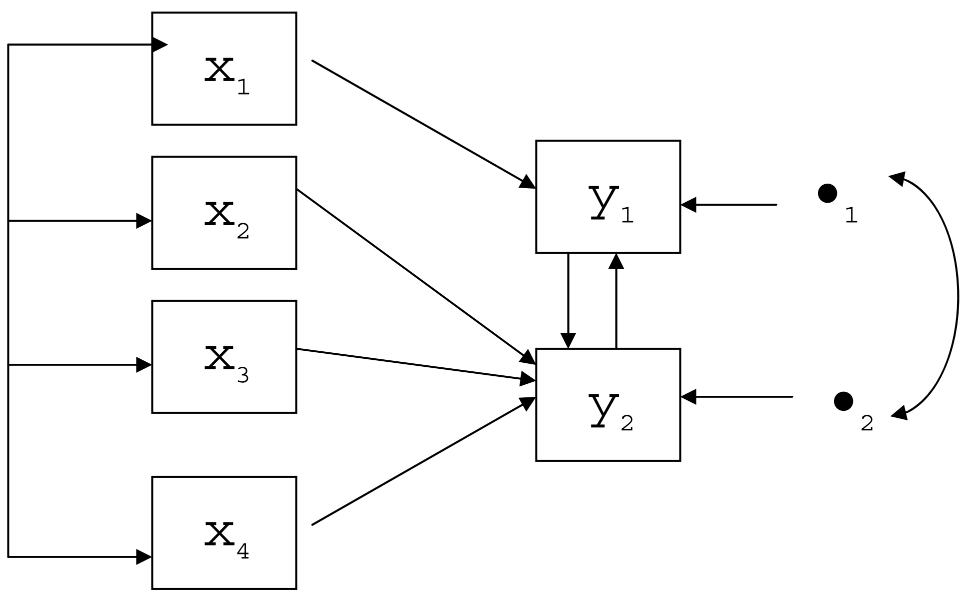
$$y=\beta y + \Gamma x + \zeta$$

2. No Measurement Error

$$y=\eta$$

$$x=\xi$$

Felson and Bohrnstedt (1979) Model of Perceived Attractiveness and Perceived Academic Performance (N = 209)



x_1 = Grade Point Average
x_2 = Deviation of height from mean by grade and sex
x_3 = Weight adjusted for height
x_4 = Physical attractiveness rated by children outside class
y_1 = Perceived academic ability, based on class-mates' ratings
y_2 = Perceived attractiveness, classmates' ratings

$$y_1 = \beta_{12} y_2 + \gamma_{11} x_1 + \zeta_1$$
$$y_2 = \beta_{21} y_1 + \gamma_{22} x_2 + \gamma_{23} x_3 + \gamma_{24} x_4 + \zeta_2$$
$$E(\zeta_i) = 0 \qquad COV(\zeta_1, \zeta_2) \neq 0$$
$$COV(\zeta_i, x_j) = 0 \text{ for } i=1,2;\ j=1,2,3,4$$

IV.B. Confirmatory Factor Analysis

1. Latent Variables

2. Measurement Errors

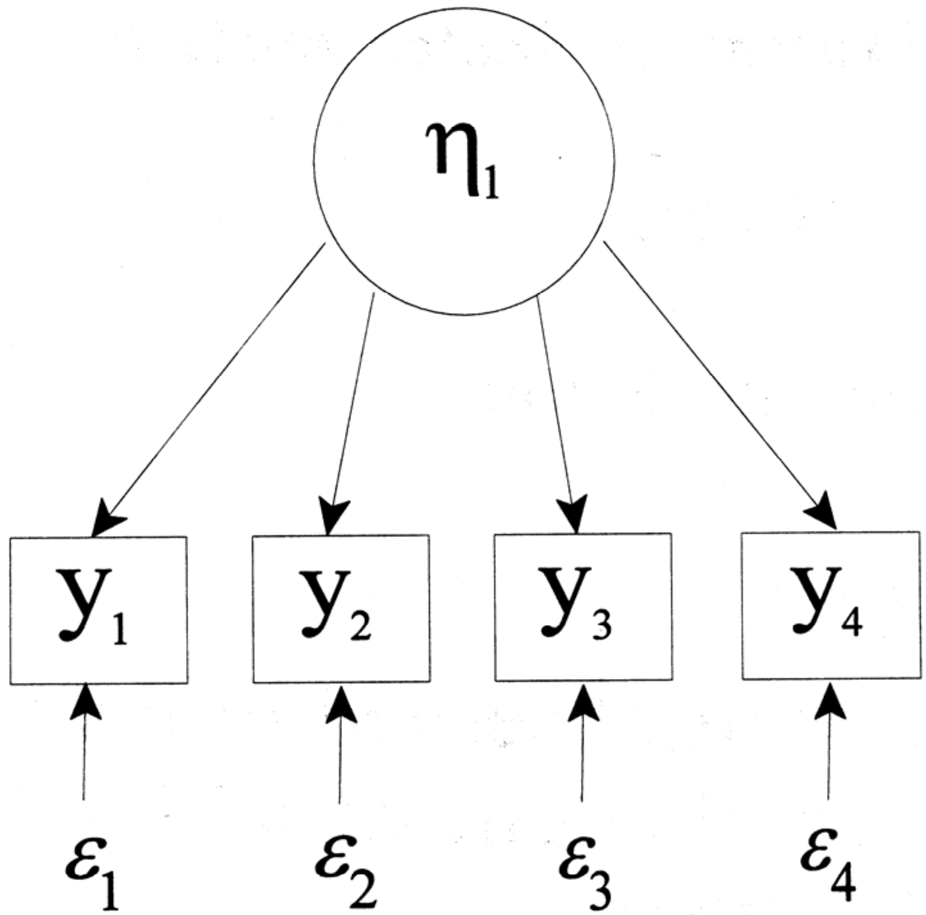
$y = \Lambda_y \eta + \varepsilon$

y = vector of observed indicators

η = vector of latent variables (or "factors")

ε = vector of "measurement errors"

$E(\varepsilon) = 0 \qquad COV(\eta, \varepsilon) = 0$



η_1 = Perceived overall air quality

y_1 = Rating of overall quality

y_2 = Rating of clarity

y_3 = Rating of color

y_4 = Rating of odor

IV.B. Confirmatory Factor Analysis

1. Latent Variable Model

$\eta = B\eta + \Gamma\xi + \zeta$

η= latent endogenous variables

ξ = latent exogenous variables

ζ = disturbance vector

B = coefficient matrix for η or η effects

Γ = coefficient matrix for ξ or η effects

$E(\zeta) = 0$, $COV(\xi, \zeta) = 0$

IV.C. General SEM

2. Measurement Model

$y = \Lambda_y \eta + \varepsilon$ $E(\varepsilon)=0$

$x = \Lambda_x \xi + \delta$ $E(\delta)=0$

y = indicators of η
Λ_y=factor loadings of η or y
ε = errors of measurement for y
x= indicators of ξ
Λ_x= factor loadings of ξ or x
δ = errors of measurement for x
δ, ε, ζ,ξ are uncorrelated

IV.C. General SEM

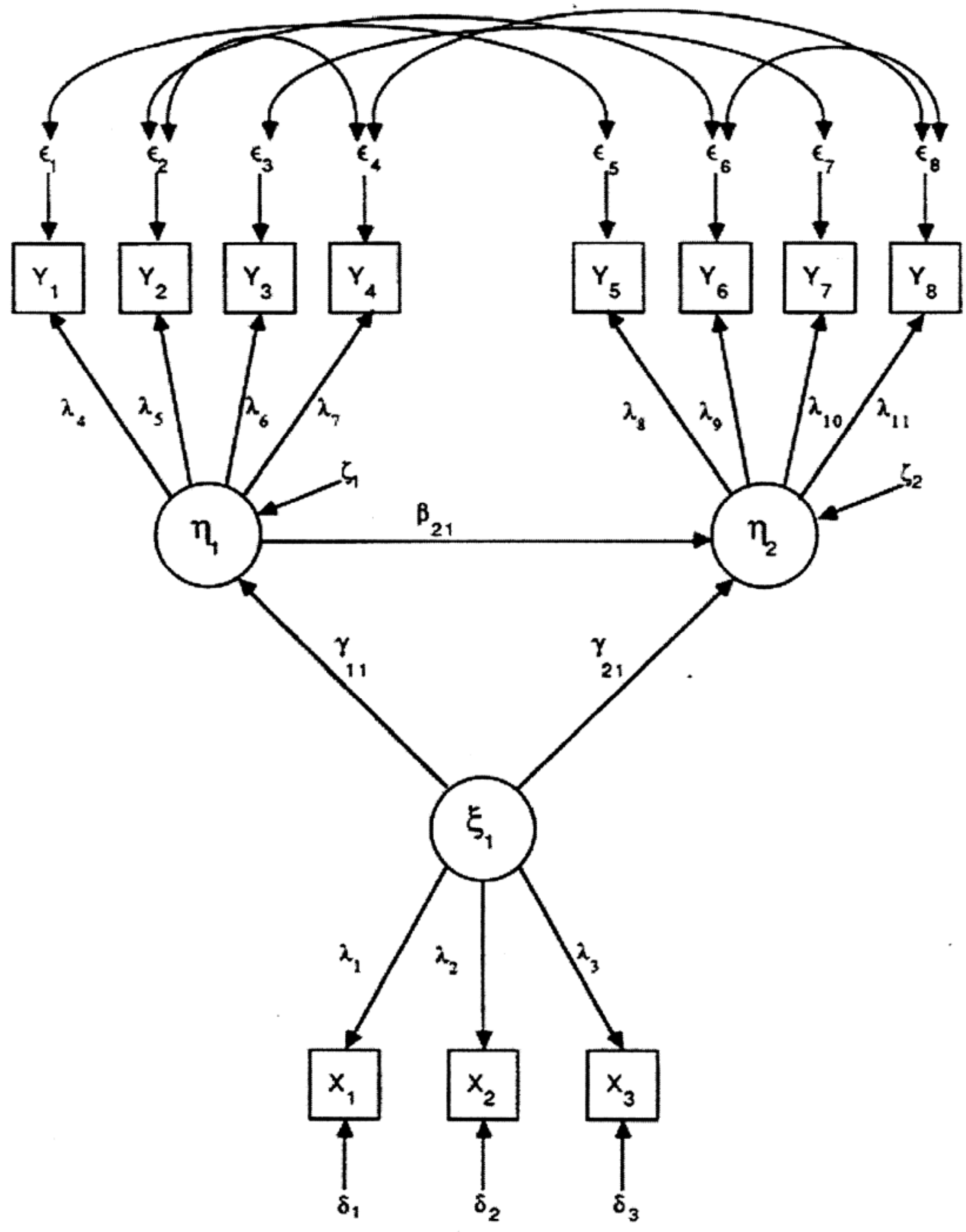


Figure 2.6 Path Diagram of Industrialization and Political Democracy Model

IV.C.

ξ_1 = Industrialization 1960

x_i = Indicators of industrialization

η_1 = Democratic political structure 1960

η_2 = Democratic political structure 1965

y_i = Indicators of political democracy

V. Overview of Steps in Modeling

A. Specification

B. Implied Covariance Matrix

C. Identification

D. Estimation

E. Testing and Diagnostics

F. Respecification

V.A. Specification

1. What latent variables?

2. Relation between latent variables?

3. What measures?

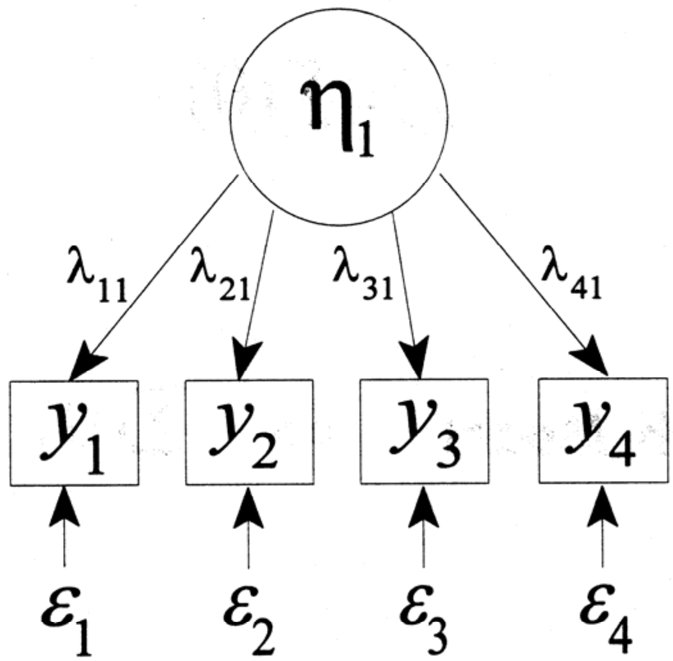
4. Relation between measures and latent variables?

V.B. Implied Covariance Matrix

H_o: $\Sigma = \Sigma(\theta)$

Each Model => $\Sigma(\theta)$

e.g.,



$$\Sigma(\theta)=$$

$$\begin{bmatrix} \lambda_{11}^2\Psi + VAR(\varepsilon_1) & & & \\ \lambda_{21}\lambda_{11}\Psi & \lambda_{21}^2\Psi + VAR(\varepsilon_2) & & \\ \lambda_{31}\lambda_{11}\Psi & \lambda_{31}\lambda_{21}\Psi & \lambda_{31}^2\Psi + VAR(\varepsilon_3) & \\ \lambda_{41}\lambda_{11}\Psi & \lambda_{41}\lambda_{21}\Psi & \lambda_{41}\lambda_{31}\Psi & \lambda_{41}^2\Psi + VAR(\varepsilon_4) \end{bmatrix}$$

V.C. Identification

Unique values for parameters?

If $\Sigma(\theta_1) = \Sigma(\theta_2)$, then $\theta_1 = \theta_2$

Identification

$$\mathrm{VAR}(y) = \theta_1 + \theta_2$$

θ_1	θ_2	VAR(y)
5	5	10
7	3	10
9	1	10

Establishing Identification

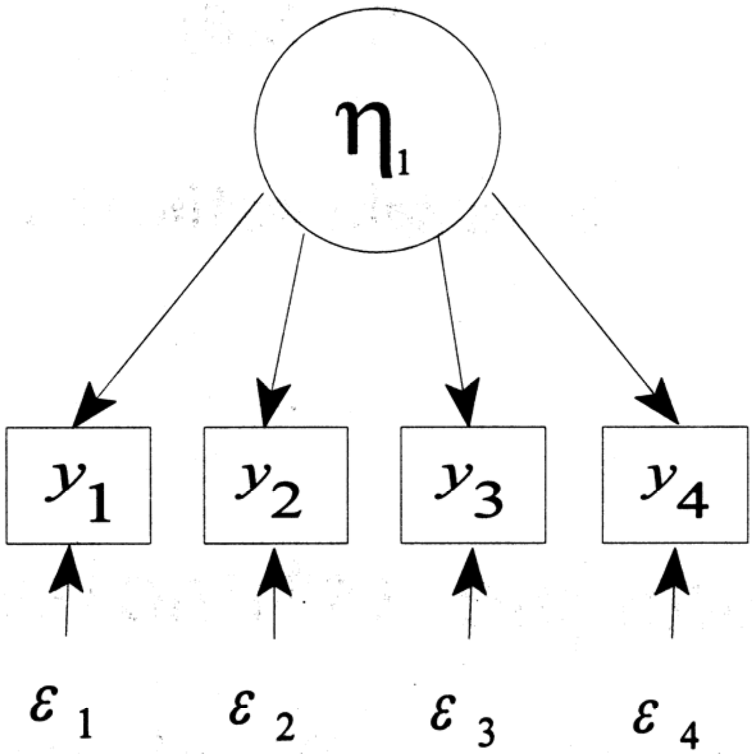
1. Algebraic Means

   $\Sigma = \Sigma(\theta)$ solve for θ

2. Identification Rules

3. Empirical Tests

e.g.,



Identified? Yes,

Three Indicator Rule

V.D. Estimation

$$H_o: \ \Sigma = \Sigma(\theta)$$

S sample estimator of Σ

$\Sigma(\hat{\theta})$ sample estimator of $\Sigma(\theta)$

Choose $\hat{\theta}$ so $\Sigma(\hat{\theta})$ close to S

e.g., $y_1 = x_1 + \zeta_1$

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

$$\begin{bmatrix} VAR(y_1) & \\ COV(x_1 y_2) & VAR(x_2) \end{bmatrix} = \begin{bmatrix} \phi_{11} + \psi_{11} & \\ \phi_{11} & \phi_{11} \\ & \end{bmatrix}$$

Goal to find θ

$$S = \begin{bmatrix} 10 & 6 \\ 6 & 4 \end{bmatrix} \qquad \Sigma(\hat{\theta}) = \begin{bmatrix} \hat{\phi}_{11} + \hat{\psi}_{11} & \hat{\phi}_{11} \\ \hat{\phi}_{11} & \hat{\phi}_{11} \end{bmatrix}$$

Find $\hat{\phi}_{11}$ and $\hat{\psi}_{11}$

Make $\Sigma(\hat{\theta})$ close to S

Say $\hat{\phi}_{11}$ = 7 , $\hat{\psi}_{11}$ = 3

$$\Sigma(\hat{\theta}) = \begin{bmatrix} 10 & 7 \\ 7 & 7 \end{bmatrix} \qquad S - \Sigma(\hat{\theta}) = \begin{bmatrix} 0 & -1 \\ -1 & -3 \end{bmatrix}$$

ESTIMATORS

A. Full Information

1. Maximum Likelihood (ML)

2. Generalized Least Squares (GLS)

3. Unweighted Least Squares (ULS)

4. Weighted Least Squares (WLS)

   [Arbitrary Distribution Function (ADF)]

B. Limited Information

1. Two-Stage Least Squares (2SLS)

V.E. Testing and Diagnostics

H_o: $\Sigma = \Sigma(\theta)$

χ^2 Test

T_m = (N-1)* Fit Function Min.

df = ½ (p+q) (p+q+1) -# of parameters

Bollen

2. Overall Model Fit

T_b = **Chi-square test statistic for baseline model**

T_m= Chi-square test statistic for hypothesized model

df_b= degrees of freedom of baseline model

df_m= degrees of freedom of hypothesized model

$$IFI = \frac{T_b - T_m}{T_b - df_m}$$

$$TLI = \frac{T_b/df_b - T_m/df_m}{T_b/df_b - 1}$$

$$RMSEA = \sqrt{\frac{T_m - df_m}{(N-1)df_m}}$$

$$BIC = T_m - df \ln(N)$$

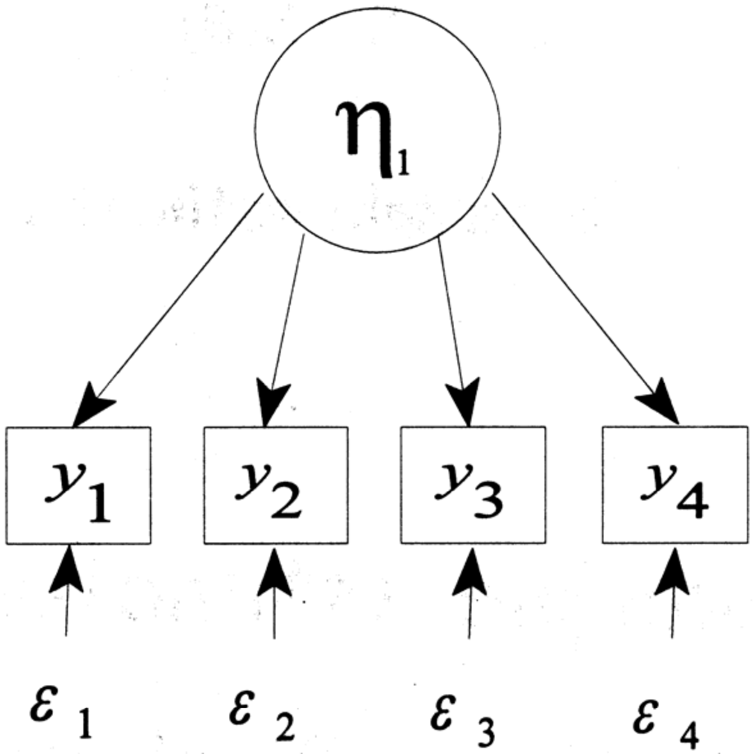
V.E. Testing and Diagnostics

3. **Residuals** $(S-\Sigma(\hat{\theta}))$

4. **Component Fit**

5. **Statistical Power**

e.g.,



χ^2 = **16.0** **df = 2** **p = .0003**

VI.E. Respecification

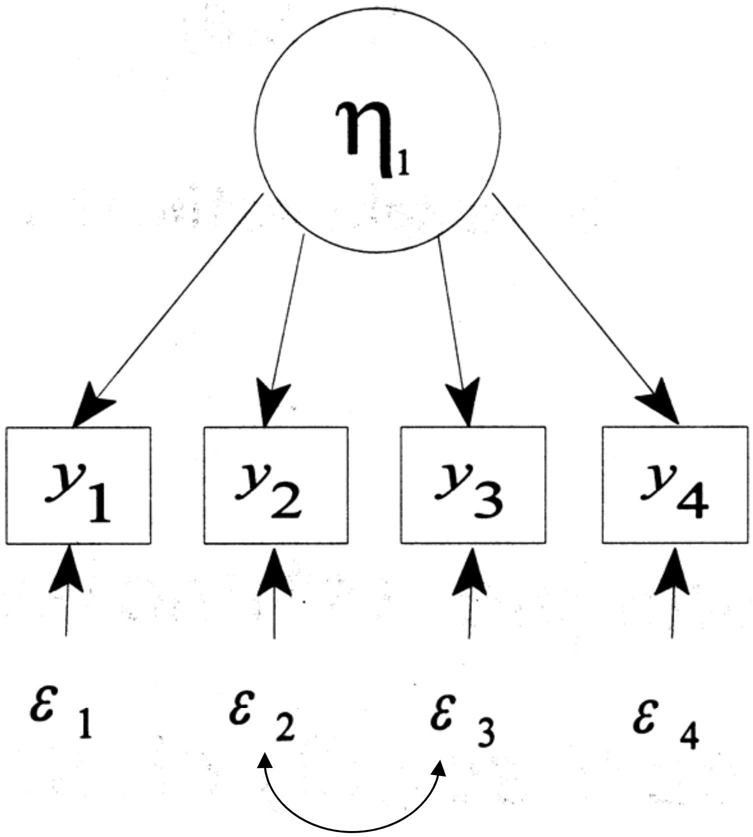
1. Substantive-Based Revisions

2. Lagrangian Multiplier

3. WALD

4. Residuals $(S - \Sigma(\hat{\theta}))$

e.g.,



$\chi^2 = 4.2$ df = 1 p = .04

Steps in Modeling

1. Specification

2. Implied Covariance Matrix

3. Identification

4. Estimation

5. Testing

6. Respecification

VI. SEM with Means and Intercepts

A. Specification

1. Latent Variable Model

$$\boldsymbol{\eta} = \alpha_{\eta} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

intercept

e.g, $\eta_1 = \alpha_{\eta 1} + B_{12}\eta_2 + \gamma_{11}\xi_1 + \zeta_1$

$\eta_2 = \alpha_{\eta 2} + B_{23}\eta_3 + \gamma_{22}\xi_2 + \zeta_2$

VI. SEM with Means and Intercepts

A. Specification

2. Measurement Model

$$\mathbf{y} = \alpha_y + \Lambda_y \eta + \varepsilon$$

$$\mathbf{x} = \alpha_x + \Lambda_x \xi + \delta$$

intercept

e.g., $y_1 = \alpha_{y1} + \lambda_{y11}\, \eta_1 + \varepsilon_1$

$y_2 = \alpha_{y2} + \lambda_{y21}\, \eta_1 + \varepsilon_2$

$x_1 = \alpha_{x1} + \lambda_{x11}\, \xi_1 + \delta_1$

VI. SEM with Means and Intercepts

B. Implied Moments

1. Implied Covariance Matrix

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

2. Implied Mean Vextor

$$\mu = \mu(\theta)$$

Mean of observed variables=
Model implied means

VI. C. Identification

Still have Σ, covariance matrix of observed variables.

New parameters:

intercepts ($\alpha_\eta, \alpha_y, \alpha_x$)   &

means of latent ξ (k)

New observed variable information:

Means of observed variables (μ_y μ_x)

Can we use Σ, μ_y, and μ_x to find unique values of all model parameters?   Yes $\longrightarrow$ identified

VI. D. Estimation

Same as before (e.g., ML, WLS, 2SLS)

E. Testing

Same as before (e.g., chi-square, IFI, etc.)

F. Respecification

Same as before

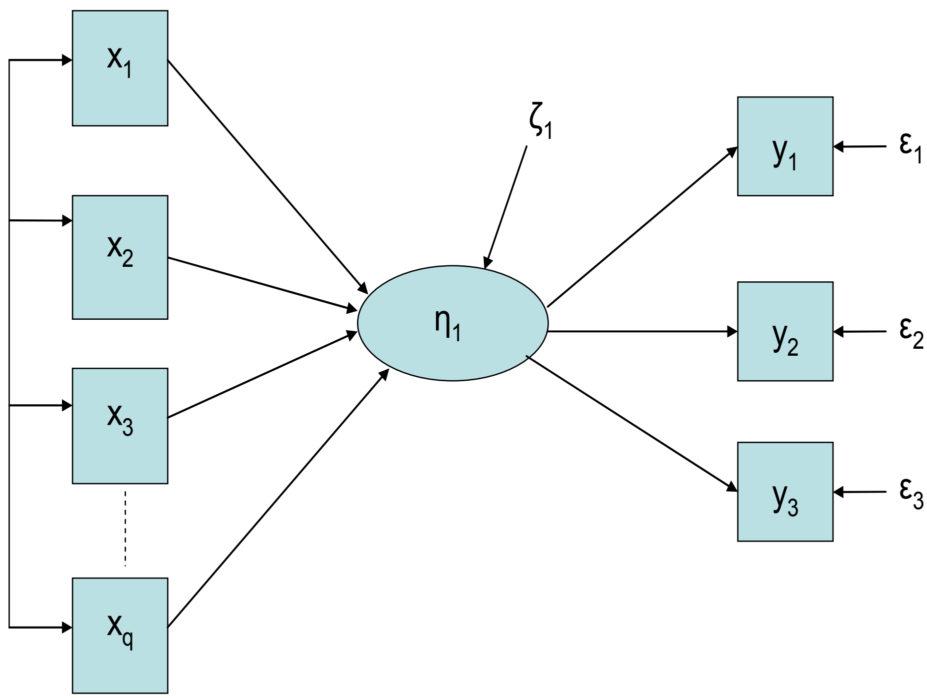
VII. Software

- LISREL [PRELIS, SIMPLIS]
- AMOS
- EQS
- Mplus
- Proc Calis in SAS
- Mx
- EzPath
- AUFIT
- LINCS

VIII. Empirical Example

Robins, P. K. and R. W. West. 1977. Measurement Errors in the Estimation of Home Value. *Journal of the American Statistical Association 72*, 290-94.

Figure Robins and West (1977, JASA) Value of Home (η_1) with Causal Indicators of lot size (x_1), square footage (x_2), number of rooms (x_3), etc. (x_q), and Effect Indicators of Appraised Value (y_1), Owner Estimate (y_2), & Assessed Value (y_3), and Disturbances (ζ_1, ε_1 to ε_3)



Journal of the American Statistical Association, June 1977

2. Maximum Likelihood Estimates ($N = 138$)

a. Home value measurement equations[a]

Source	γ_0	$\hat{\gamma}$	$\hat{\theta}$	R^2_I
Appraised value	0	1.000 (—)	2.415 (.158)	.553
Owner-estimate	1.257	.973 (.119)	2.771 (.177)	.470
Assessed value	−5.909	1.339 (.120)	1.612 (.169)	.832

b. Casual equation[b]

Variable	$\hat{\alpha}$	Estimated asymptotic standard error of $\hat{\alpha}$
Construction grade	1.077	.201
1 if attached garage	.134	.555
1 if detached garage	.693	.248
1 if basement garage	1.537	.433
Finished area (hundreds of sq. ft.)	.223	.045
1 if substandard storage space	−1.391	.679
1 if quality is below neighborhood standards	−1.144	.646
Number of stories	.493	.366
Number of built-ins	.461	.172
Effective age	−.073	.010
Number of rooms	.287	.129
Lot size (hundreds of sq. ft.)	.014	.006
Constant term ($\hat{\alpha}_0$)	5.704	(—)

c. Derived statistics

Statistic	Estimate
R^2_C	.907
R^2_{CI}	.982
$\boldsymbol{\alpha}'\boldsymbol{\phi}\boldsymbol{\alpha}$	6.530
$\boldsymbol{\gamma}'\boldsymbol{\Omega}^{-1}\boldsymbol{\gamma}$	.544
$\hat{V}(y^*)$	7.201

[a] Estimated asymptotic standard errors in parentheses.
[b] $\hat{\sigma} = .819$, and standard error of $\hat{\sigma} = .163$.

VIII. Empirical Example

Bollen, K.A. 1993. Liberal Democracy: Validity and Method Factors in Cross-National Measures. *American Journal of Political Science*, 37:1207-1230.

Figure 2. Path Diagram of Model for Eight Indicators of Political Liberties and Democratic Rule

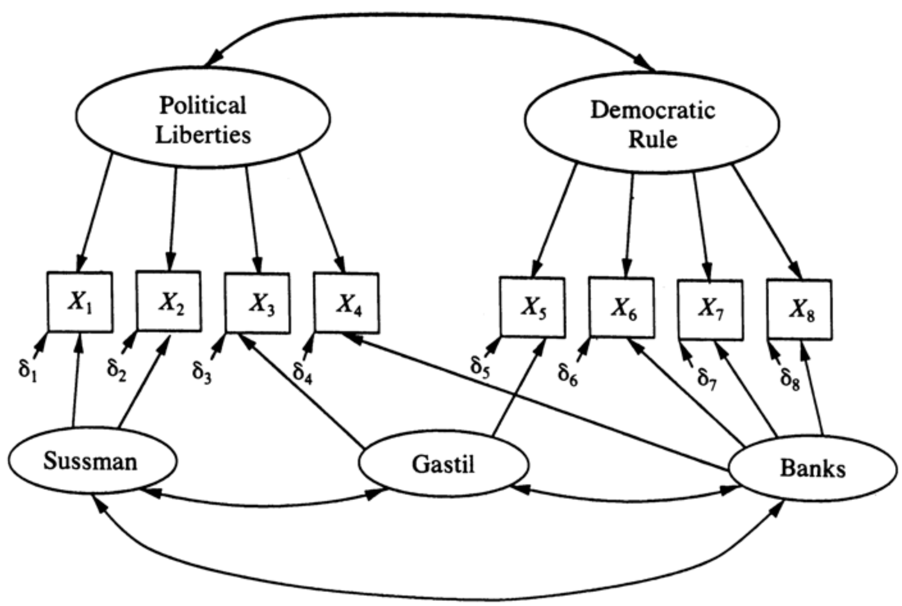


Table 1. Overall Fit of Confirmatory Factor Analysis Models for 1980 (N = 153)

Model	χ^2	*df*	*p*-value	Δ_1	Δ_2	GFI	AGFI
Initial model M_1	9.2	8	.33	.99	1.00	.99	.93
Trimmed model M_2	14.2	13	.36	.99	1.00	.98	.94
$M_2 - M_1$	5.0	5	>.25	.00	.00	−.01	+.01
No method factors M_3	217	19	<.001	.86	.88	.73	.48

Table 2. Variance in Indicators Due to Validity, Method Factor, and Random Measurement Error (1980 Data)

	Percent of Total Variance of Indicator Due to:		
Variable	Validity %	Method Factor Error %	Random Measurement Error %
X_1	68	14	18
X_2	71	22	7
X_3	78	16	6
X_4	92	0	8
X_5	93	7	0
X_6	62	38	0
X_7	14	22	64
X_8	79	9	12

Where:
X_1 = freedom of broadcast media (Sussman)
X_2 = freedom of print media (Sussman)
X_3 = civil liberties (Gastil)
X_4 = freedom of group opposition (Banks)
X_5 = political rights (Gastil)
X_6 = competitiveness of nomination process (Banks)
X_7 = chief executive elected (Banks)
X_8 = effectiveness of legislative body (Banks)