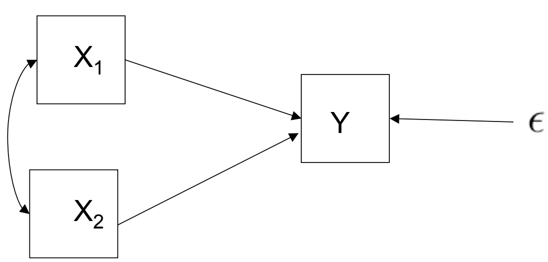
Introduction to Structural Equation Models

Jerry Brunner
Department of Statistics

Structural equation models are extensions of multiple regression

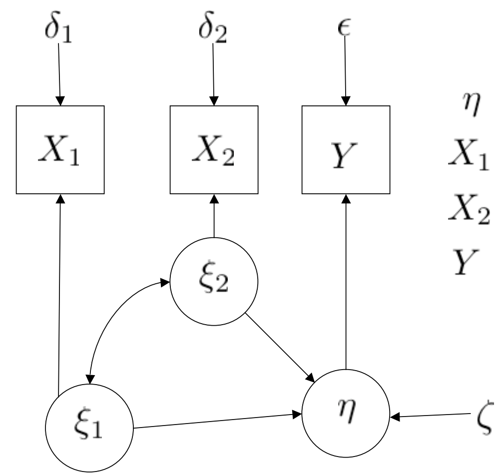
- Simultaneous equations
- Include latent variables
- DV in one equation can be IV in another

Multiple Regression



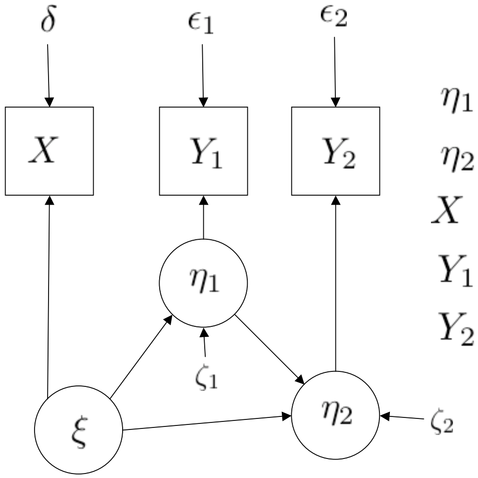
$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

An Extension



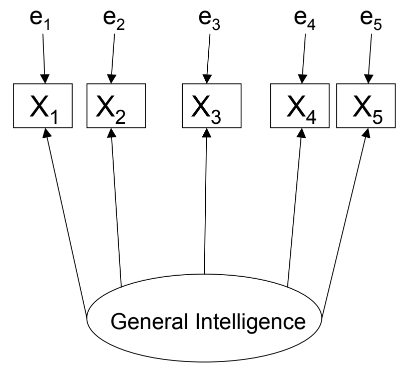
$$\begin{aligned} \eta &= \gamma_1 \xi_1 + \gamma_2 \xi_2 + \zeta \\ X_1 &= \xi_1 + \delta_1 \\ X_2 &= \xi_2 + \delta_2 \\ Y &= \eta + \epsilon \end{aligned}$$

A Further Extension



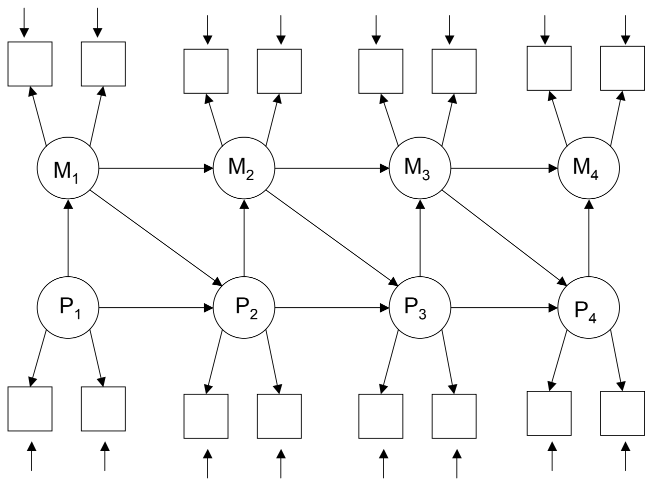
$$\begin{aligned}
\eta_1 &= \gamma_1 \xi + \zeta_1 \\
\eta_2 &= \beta \eta_1 + \gamma_2 \xi + \zeta_2 \\
X &= \xi + \delta \\
Y_1 &= \eta_1 + \epsilon_1 \\
Y_2 &= \eta_2 + \epsilon_2
\end{aligned}$$

A Factor Analysis Model

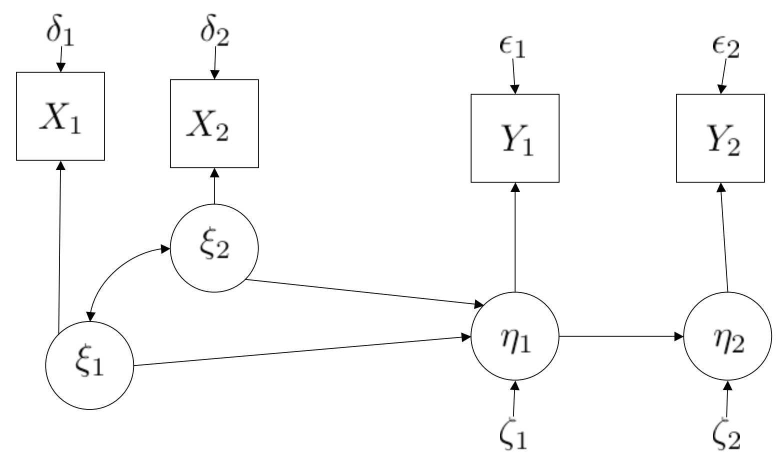


$$X_1 = \lambda_1 F + e_1$$
$$X_2 = \lambda_2 F + e_2$$
$$X_3 = \lambda_3 F + e_3$$
$$X_4 = \lambda_4 F + e_4$$
$$X_5 = \lambda_5 F + e_5$$

A Longitudinal Model



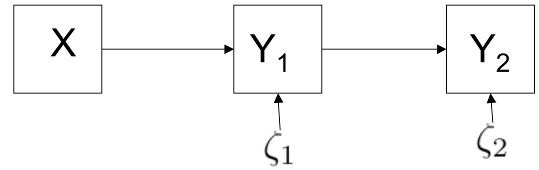
Vocabulary: Latent v.s. Manifest, Endogenous v.s. Exogenous



Notation

- LISREL (Bollen, 1989)
- Classical Factor Analysis (Lawley, 1971)
- LINEQS (Bentler and Weeks, 1980)
- RAM (McArdale, 1980)
- COSAN (McDonald 1978, 1980)

Estimation and Testing



$$
\begin{aligned}
Y_1 &= \gamma X + \zeta_1 \\
Y_2 &= \beta Y_1 + \zeta_2
\end{aligned}
$$

$E(X) = E(\zeta_1) = E(\zeta_2) = 0$

$V(X) = \phi, V(\zeta_1) = \psi_1, V(\zeta_2) = \psi_2$

X, ζ_1, ζ_2 are independent

Everything is normal

Distribution of the data

$\begin{bmatrix} X_1 \\ Y_{1,1} \\ Y_{1,2} \end{bmatrix} \ldots \begin{bmatrix} X_n \\ Y_{n,1} \\ Y_{n,2} \end{bmatrix}$ are independent normal with mean zero

and covariance matrix

$$
\boldsymbol{\Sigma} = \begin{bmatrix} \phi & \gamma\phi & \beta\gamma\phi \\ \gamma\phi & \gamma^2\phi + \psi_1 & \beta(\gamma^2\phi + \psi_1) \\ \beta\gamma\phi & \beta(\gamma^2\phi + \psi_1) & \beta^2(\gamma^2\phi + \psi_1) + \psi_2 \end{bmatrix}
$$

$$
\boldsymbol{\theta} = (\gamma, \beta, \phi, \psi_1, \psi_2)
$$

Maximum Likelihood

$$
L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-n/2}(2\pi)^{-nk/2} \exp -\frac{n}{2}\left[tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}^{-1}) + (\overline{\mathbf{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu})\right]
$$

Minimize $-2\ell(\boldsymbol{\mu}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$

$$
\begin{aligned}
= n\,[\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| &+ k\log(2\pi) + tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) \\
&+ (\overline{\mathbf{x}} - \boldsymbol{\mu}(\boldsymbol{\theta}))'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\overline{\mathbf{x}} - \boldsymbol{\mu}(\boldsymbol{\theta}))]
\end{aligned}
$$

Likelihood Ratio Test for Goodness of Fit

$$
\begin{aligned}
G &= -2\log\frac{L(\overline{\mathbf{X}}, \boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}}))}{L(\overline{\mathbf{X}}, \widehat{\boldsymbol{\Sigma}})} \\
&= n[\log|\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})| + k\log(2\pi) + tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1}] \\
&\quad -n[\log|\widehat{\boldsymbol{\Sigma}}| \quad + k\log(2\pi) + k] \\
&= n[\log|\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})| - \log|\widehat{\boldsymbol{\Sigma}}| + tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})^{-1} - k]
\end{aligned}
$$

Do it all at once: Minimize

$$G(\boldsymbol{\theta}) = n[\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \log|\widehat{\boldsymbol{\Sigma}}| + tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) - k]$$

Actually, SAS minimizes the “Objective Function”

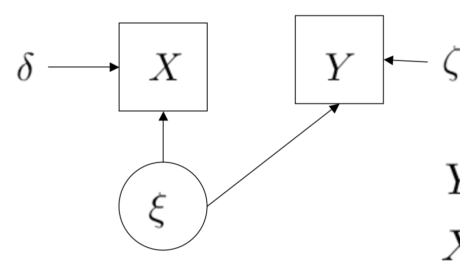
$$\log|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \log|\widehat{\boldsymbol{\Sigma}}| + tr(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}) - k$$

Chi-square and Z Tests

- “Chisquare” is (n-1) times minimum objective function.
- Test nested models by difference between chi-square values
- Z tests are produced by default; Asymptotic Covariance matrix is available

- Likelihood ratio tests perform better

Simple Regression with measurement Error



$$
\begin{aligned}
Y &= \gamma\xi + \zeta \\
X &= \xi + \delta
\end{aligned}
$$

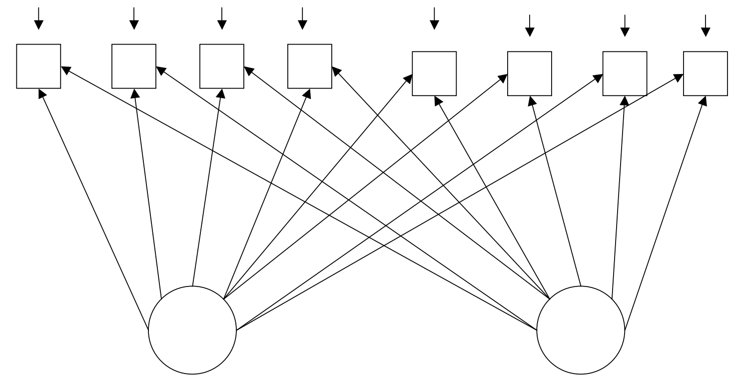
$$V(\xi) = \phi, V(\zeta) = \psi, V(\delta) = \theta_\delta$$

The Model is not Identified

$$\boldsymbol{\theta} = (\gamma, \phi, \theta_\delta, \psi)$$

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} \sigma_{1,1} & \sigma_{1,2} \\ & \sigma_{2,2} \end{array} \right] = \left[\begin{array}{cc} \phi + \theta_\delta & \gamma\phi \\ & \gamma^2\phi + \psi \end{array} \right]$$

Unconstrained (Exploratory) Factor Analysis



Model Identification

$$\begin{array}{lcl} \theta & \in & \Theta \\ \mathbf{D} & \sim & P_\theta = g(\theta) \in \mathcal{P} \end{array}$$

$$g : \Theta \rightarrow \mathcal{P}$$

If the function g is one to one, then the model is identified.

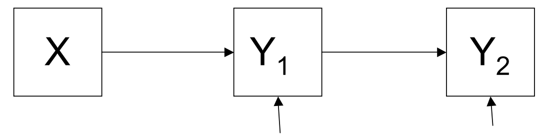
Consistent Estimation is Impossible

Suppose $\theta_1 \neq \theta_2$ with $P_{\theta_1} = P_{\theta_2}$

To prove model identification

- Show that the parameter can be recovered from the distribution of the observed data.
- In practice, recover it from the moments (usually, the covariance matrix).
- Sometimes, only a function of the parameter is identified.

Remember the example



$$\boldsymbol{\theta} = (\gamma, \beta, \phi, \psi_1, \psi_2)$$

$$\boldsymbol{\Sigma} = \left[\begin{array}{ccc} \phi & \gamma\phi & \beta\gamma\phi \\ \gamma\phi & \gamma^2\phi + \psi_1 & \beta(\gamma^2\phi + \psi_1) \\ \beta\gamma\phi & \beta(\gamma^2\phi + \psi_1) & \beta^2(\gamma^2\phi + \psi_1) + \psi_2 \end{array} \right]$$

Solve the identifying equations

$$\boldsymbol{\Sigma} = \left[\begin{array}{ccc} \sigma_{1,1} & \sigma_{1,2} & \sigma_{1,3} \\ & \sigma_{2,2} & \sigma_{2,3} \\ & & \sigma_{3,3} \end{array} \right] = \left[\begin{array}{ccc} \phi & \gamma\phi & \beta\gamma\phi \\ & \gamma^2\phi + \psi_1 & \beta(\gamma^2\phi + \psi_1) \\ & & \beta^2(\gamma^2\phi + \psi_1) + \psi_2 \end{array} \right]$$

$$\text{For } \boldsymbol{\theta} = (\gamma, \beta, \phi, \psi_1, \psi_2)$$

An Identified model can be

- Just-Identified (saturated): Same number of parameters and identifying equations
- Over-identified: More equations than unknowns

Identification Rules

- A necessary condition for all Models
- Sufficient conditions for models with just observed variables
- Sufficient conditions for measurement models (factor analysis)
- Sufficient conditions for combined models

Parameter Count

- The number of parameters must be no more than the number of unique elements in the covariance matrix of the observed variables.

- Necessary for all models

Observed variable models

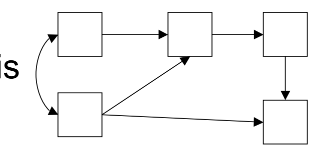
$$\mathbf{Y} = \boldsymbol{\beta}\mathbf{Y} + \boldsymbol{\Gamma}\mathbf{X} + \boldsymbol{\zeta}$$

Identified if $Cov(\mathbf{X}, \boldsymbol{\zeta}) = \mathbf{0}$ and

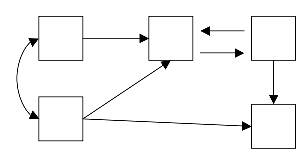
- $\boldsymbol{\beta} = \mathbf{0}$ (Regression model), or
- Model is Recursive, and $Var(\boldsymbol{\zeta})$ is diagonal

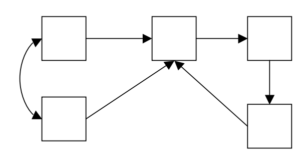
Recursive means no Feedback Loops

- Like this

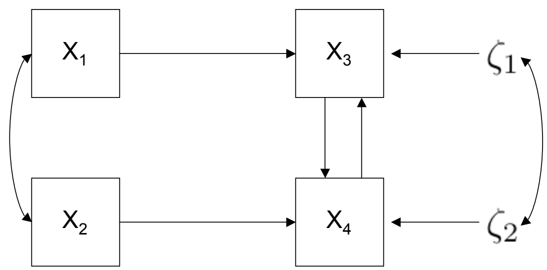


- Not this



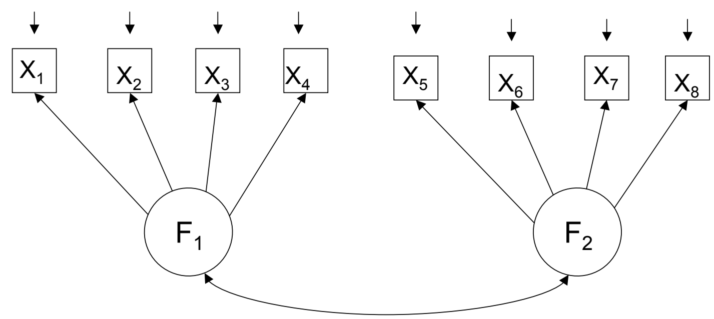


A just-identified, Non-recursive Model



Measurement Models (Confirmatory Factor Analysis)

$$\mathbf{X} = \mathbf{\Lambda F} + \mathbf{e}$$



Rules for Confirmatory Factor Analysis

- Three-indicator rules
- Two-indicator rules

"Pick a Scale" for each factor

- F is obesity
- X_1 is percent body fat from immersion
- X_2 is triceps skin fold
- X_3 is Body Mass Index

$$\begin{aligned} X_1 &= \lambda_1 F + e_1 \\ X_2 &= \lambda_2 F + e_2 \\ X_3 &= \lambda_3 F + e_3 \end{aligned} \longrightarrow \begin{aligned} X_1 &= F + e_1 \\ X_2 &= \lambda_2 F + e_2 \\ X_3 &= \lambda_3 F + e_3 \end{aligned}$$

Three-indicator rules

- At least three variables per factor
- Each variable caused by only one factor
- Errors uncorrelated with factors and with each other
- Pick a scale for each factor

- No restrictions on Var(**F**)

Two-Indicator Rules

- At least two variables per factor
- Each variable caused by only one factor
- Errors uncorrelated with factors and with each other
- Pick a scale for each factor
- At least one non-zero off-diagonal element in each row of Var(**F**)

Combined Models

- Consider the latent part of the model as a model for observed variables. Verify identification.
- Consider the measurement part of the model as a confirmatory factor analysis, ignoring structure in Var(**F**). Verify identification.

Fixing up non-identified models

- Negotiation
- Deeper study may be rewarded
- “Model” is not identified.

- Consider identification before collecting data!

Further Issues

- Normality
- Numerical problems
- Sample size
- Categorical data

Software

- LISREL, EQS, RAM
- Mplus
- Stata
- R
- AMOS (Graphical interface)
- SAS proc calis