

Structural Equation Model (SEM) Using LISREL

Summarized by Hun Myoung Park (Joint Public Policy

Structural equation modeling takes a confirmatory, rather than an exploratory, approach to the data analysis. Whereas traditional multivariate procedures are incapable of either assessing or correcting for measurement error, SEM provides explicit estimates of these parameters. Whereas the former methods are based on observed measurements only, SEM can incorporate both unobserved and observed variables (Byrne 1998: 3-4) observed or manifest variables serve as indicators of the underlying constructs (latent variables or factors) that they are presumed to represent.

Components

- **Structural Model:** $\mathbf{h} = \mathbf{B}\mathbf{h} + \mathbf{\Gamma}\mathbf{x} + \mathbf{z}$, which specifies the causal relationships among the latent endogenous (η) and exogenous (ξ) variables, describes the causal effects, and assigns the explained and unexplained variance (disturbance term). Latent variables (hypothetical constructs) are underlying causes of multiple observed behaviors.
- **Measurement Model:** $y = \Lambda_y \mathbf{h} + \mathbf{e}$ and $x = \Lambda_x \mathbf{x} + \mathbf{d}$, which specifies how latent variables (hypothetical constructs) depend upon or are indicated by the observed variables.
- The third model, for example, has two η s, one ξ , two y s (y_1 and y_2) associated with η_1 , three y s (y_3 through y_5) associated with η_2 , three x s associated with ξ ?

$$\begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{b}_{12} \\ \mathbf{b}_{21} & 0 \end{bmatrix} \times \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{g}_{11} \\ \mathbf{g}_{21} \end{bmatrix} \times [\mathbf{x}_1] + \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \end{bmatrix}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_p \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{y11} & 0 \\ \mathbf{I}_{y21} & 0 \\ \dots & \dots \\ 0 & \mathbf{I}_{yp2} \end{bmatrix} \times \begin{bmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \dots \\ \mathbf{e}_p \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_q \end{bmatrix} = \begin{bmatrix} \mathbf{I}_{x11} \\ \mathbf{I}_{x21} \\ \dots \\ \mathbf{I}_{xq1} \end{bmatrix} \times [\mathbf{x}_1] + \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \\ \dots \\ \mathbf{d}_q \end{bmatrix}$$

Notations

- \mathbf{h} (eta) is a $m \times 1$ random vector of latent dependent, or endogenous, variables
- \mathbf{x} (ksi) is a $n \times 1$ random vector of latent independent, or exogenous, variables
- \mathbf{y} is a $p \times 1$ vector of observed (endogenous) indicators of the dependent latent variables η
- \mathbf{x} is a $q \times 1$ vector of observed (exogenous) indicators of the independent latent variables ξ
- \mathbf{e} (epsilon) is a $p \times 1$ vector of measurement errors in an observed endogenous variable y
- \mathbf{d} (delta) is a $q \times 1$ vector of measurement errors in an observed exogenous variable x
- \mathbf{L}_y (lamda y) is a $p \times m$ coefficients matrix of the regression of y on η
- \mathbf{L}_x (lamda x) is a $q \times n$ coefficients matrix of the regression of x on ξ

- \mathbf{G} (gamma) is a $m \times n$ coefficients matrix of the ξ in the structural relationship
- \mathbf{B} (beta) is a $m \times m$ coefficients matrix of the η in the structural relationship. (\mathbf{B} has zeros in the diagonal, and $\mathbf{I} - \mathbf{B}$ is required to be non-singular)
- \mathbf{z} (zeta) is a $m \times 1$ vector of equation errors (residual) in the structural relationship between η and ξ

Assumptions of Random Components: OLS Assumptions

- ε is uncorrelated with η : $\text{Cov}(\varepsilon, \eta) = E(\varepsilon\eta) = 0$
- δ is uncorrelated with ξ : $\text{Cov}(\delta, \xi) = E(\delta\xi) = 0$
- ζ (zeta) is uncorrelated with ξ : $\text{Cov}(\zeta, \xi) = E(\zeta\xi) = 0$
- ε , δ , and ζ are mutually uncorrelated with each other:
 $\text{Cov}(\varepsilon, \delta) = E(\varepsilon\delta) = 0$, $\text{Cov}(\delta, \zeta) = E(\delta\zeta) = 0$, and $\text{Cov}(\zeta, \varepsilon) = E(\zeta\varepsilon) = 0$

Covariance Matrices

- $\text{Cov}(y, x) = \mathbf{S}$, a $(p+q)$ by $(p+q)$ matrix, "sigma," whose estimate is s
- $\text{Cov}(\varepsilon) = \mathbf{T}_\varepsilon$ ($p \times p$) "theta-epsilon," a v-cov matrix for measurement error ε
- $\text{Cov}(\delta) = \mathbf{T}_\delta$ ($q \times q$) "theta-delta," a v-cov matrix for measurement error δ
- $\text{Cov}(\xi) = \mathbf{\Phi}$ ($n \times n$) "phi," a variance-covariance matrix of ξ
- $\text{Cov}(\zeta) = \mathbf{\Psi}$ ($m \times m$) "psi," a variance-covariance matrix of ζ

Estimating and Degree of Freedom

- $\Lambda_y, \Lambda_x, \Gamma, \mathbf{B}, \Theta_\varepsilon, \Theta_\delta, \Phi, \Psi$
- $s = \Sigma(\hat{\mathbf{q}})$, an estimator of $\Sigma(\mathbf{q})$, which represents the restricted covariance matrix implied by the model.
- Estimate parameters so that estimators can minimize residuals, $s - \Sigma(\hat{\mathbf{q}})$, where s represent the sample covariance matrix of observed variables (known), \mathbf{S} the population covariance matrix, and \mathbf{q} a (unknown) vector that comprises the model parameters.
- The goal of covariance structure analysis is to estimate parameters of the model, which minimize the residuals.
- Maximum number of parameters to be estimated: $(p+q)(p+q+1)/2$
- The degree of freedom (the mean of a Chi square distribution): maximum - the number of parameter to be actually estimated.

Matrix	Elements	Type	Greek	LISREL	Default
Λ_x ($q \times n$)	\mathbf{I}_x	REG Coefficients	Lambda x	LX	FU,FI
Λ_y ($p \times m$)	\mathbf{I}_y	REG Coefficients	Lambda y	LY	FU,FI
\mathbf{T}_δ ($q \times q$)	\mathbf{q}_d	Variance-Covariance	Theta delta	TD	DI,FR
\mathbf{T}_ε ($p \times p$)	\mathbf{q}_e	Variance-Covariance	Theta epsilon	TE	DI,FR

Γ ($m \times n$)	\mathbf{g}_{ij}	REG Coefficients	Gamma	GA	FU,FR
\mathbf{B} ($m \times m$)	\mathbf{b}_{ij}	REG Coefficients	Beta	BE	ZE,FI
Φ ($n \times n$)	\mathbf{f}_{ij}	Variance-Covariance	Phi	PH	SY,FR
Ψ ($m \times m$)	\mathbf{y}_{ij}	Variance-Covariance	Psi	PS	DI,FR
$\Theta_{\delta\epsilon}(\rho+\dots)$	\mathbf{q}_{de}	Variance-Covariance	Theta delta epsilon	TH	ZE,FI

Source: adapted from Byrne 1998

* Matrix Forms: FU (Full), SY (Symmetric), DI (Diagonal), ID (Identity), and ZE (Zero)

** Matrix Estimation Modes: FI (Fixed) and FR (Freely estimated)

First-order and Second-order Factor Models

First-order factor models are those in which correlations among the observed variables can be described by a small number of latent variables, each of which may be considered to be one level, or one unidimensional arrow away from the observed variables (18). Second-order factor models are those in which correlations among the first-order factors can be represented by a single factor, or at least a smaller set of factors (19).

Comparisons with Other Analysis

- Path analysis, factor analysis, and regression are special cases of SEM
- SEM is basically just path analysis with latent variables
- Path analysis contains only observed variable, and has a more restrictive set of assumptions than SEM

Statistical Significance

- Each parameter estimate of \mathbf{I}_x , \mathbf{I}_y , \mathbf{q}_d , or \mathbf{q}_e in the LISREL output consists of three components: the estimate, its parenthesized standard error, and corresponding t-value.
- Because standard errors are influenced by the units of measurement in observed variable, latent variables, or both, as well as the magnitude of the parameter estimate itself, no definitive criterion of small and large has been established (103-104).
- Squared multiple correlations of observed variables, so called R square, serve as reliability indicators of the extent to which each adequately measures its respective underlying construct (104). (R square) % of the variance of an observed variable can be explained by its latent variable.

Robust MLE and WLS

Robust MLE using asymptotic covariance matrix ->many data are required

WLS unless ME=ML; NCP does not appear.

Goodness of Fit: χ^2 and NCP

- H_0 of the Chi square test is that the model exactly fits the data or $\Sigma = \Sigma(\mathbf{q})$. In contrast to traditional approaches, *SEM researchers hope not to reject H_0* (107). The null hypothesis postulates that specification of the Λ_y , Λ_x , Θ_ϵ , Θ_δ , and Φ matrices of the model is valid

(true). The probability value associated with χ^2 represents the likelihood of obtaining a χ^2 value that exceeds the (critical) χ^2 value when the null hypothesis is true (110).

- Chi square, $nF \sim \chi^2$ (where F is the minimum fit function value), is in fact the Likelihood Ratio test statistic, which tests the closeness of fit between the unrestricted sample covariance matrix S and the restricted covariance matrix $\Sigma(\mathbf{q})$. The degree of freedom, the mean of the χ^2 distribution, is the difference between the maximum number of parameters to be estimated, $(p+q)(p+q+1)/2$, and the number of parameters that are actually estimated.
- Statistical significance testing with respect to the analysis of covariance structure is somewhat different in that it is driven by degrees of freedom involving the number of elements in the sample covariance matrix and the number of parameters to be estimated (109).
- **NCP** (Noncentrality Parameter) $\chi^2_{df,1}$ is a fixed parameter with associated degree of freedom. It also serves as a measure of the discrepancy, thus can be regarded as a natural measure of badness-of-fit of a covariance structure model (111). If the model fits data, a central distribution is confirmed. Otherwise (H_0 is not valid), we get a noncentral distribution with noncentral parameters.
- The central χ^2 is a special case of the noncentral χ^2 distribution when $NCP=0$ (111).
 $NCP = nF_0 = \chi^2 - df$.

Goodness of Fit: RMSEA

- **RMSEA** (Root Mean Square Error of Approximation) takes into account the error of approximation in the population and asks "How well would the model, with unknown but optimally chosen parameter values, fit the population covariance matrix if it were available?"
- This discrepancy is expressed per degree of freedom. RMSEA less than .05 indicates good fit, and values as high as .08 represent reasonable errors of approximation in the population.

- $F_0 = \max\{F - \frac{df}{n}, 0\}$ and $RMSEA = e = \sqrt{\frac{F_0}{df}}$, where F_0 is a population discrepancy

function value. $e^2 = \frac{F - df/n}{df}$, $df \times e^2 = F - \frac{df}{n}$,

$$df \times n \times e^2 = nF - df = \chi^2 - df = NCP = nF_0$$

- The more elaborate a model with many parameters to be estimated, the less its approximate error (difference between a perfect approximation and df/n); simpler models with less parameters are prone to more approximation errors.
- $df \times n \times e^2$ reflects "error of approximation," difference between true mean (df of correct chi square distribution) and the mean of the wrong chi square distribution ($df + NCP$). As

degrees of freedom increases (simpler model), the error becomes large, shifting the wrong (or close fit) chi square distribution to the right.

- Confidence intervals can be influenced seriously by sample size as well as model complexity (the number of parameters estimated). Given a complex model, a vary large sample size would be required in order to obtain a reasonably narrow confidence interval (113).

Goodness of Fit: Other Statistics

- **ECVI** (Expected Cross-Validation Index) assesses the likelihood that the model cross-validates across similar-sized samples from the same population. It measures the discrepancy between the fitted covariance matrix in the analyzed sample, and the expected covariance matrix that would be obtained in another sample of equivalent size. The model having the smallest ECVI value (smaller than independence and saturated models) exhibits the greatest potential for replication.
- χ^2 for an independence model (a null model), the most restricted model, where all correlations among variables are zero (114).
- **AIC** (Akaike's Information Criterion) and **Consistent AIC (CAIC)** address the issues of parsimony in the assessment of model fit so that statistical goodness-of-fit as well as the number of estimated parameters are taken into account (115). Unlike AIC, CAIC takes sample sizes into account. A smaller (than saturated and independence model) CAIC of the model indicates cross-validated parameter estimators of the model.
- **RMR** (Root Mean Square Residual) represents the average residual value derived from the fitting of the variance-covariance matrix for the model to the variance-covariance matrix of the sample data. Standardized RMR represents the average value across all standardized residuals. In a well-fitting model, the value will be small (less than .05).
- **GFI** (Goodness-of-Fit Index) is a measure of the relative amount of variance and covariance in S. Adjusted GFI adjusts for the number of degrees of freedom in the specified model (115). They range from zero to 1.00, with values close to 1.0 being good fit.
- **PGFI** (Parsimony Goodness-of-Fit Index) takes into account the complexity of the model in the assessment of overall model fit (116). It incorporates a penalty for the inclusion of additional parameters.
- **NFI** (Normed Fit Index) and **CFI (Comparative Fit Index)** range from zero to 1 on the basis of the comparison of the model with the independence model (null model). A value greater than .90 indicates an acceptable fit to the data. **NNFI** (Non-normed Fit Index), although not normed, takes the complexity of model into account. If a model fits well, NNFI becomes close to 1 since $E\left(\frac{nF_i}{d_i}\right)$, otherwise larger than 1. **IFI** (Incremental Index of Fit) addresses both parsimoniousness (degree of freedom) and sample size. **RFI** (Relative Fit Index) or **RNI** (Relative Noncentrality Index) is algebraically equivalent to the CFI (117). **PNFI** (Parsimony Normed Fit Index) also takes the complexity of the model into account.

- CN (Critical N) is not a measure of goodness-of-fit, but a measure of adequate sample size. It estimates the sample size that would be sufficient to yield an adequate model of fit for a χ^2 test (118).

Memos

- WLS (Weighted Least Squares) should be used whenever AC or AV is used (in ordinal scale data). .dfs format has correlation and covariance matrix.

LISREL: DA (Data)

- `&& DA NI=4 NO=7000 MA=CM`
- **NI** is the number of input variables (observed variables) in a data set.
- **NO** is the number of observations.
- **MA** specifies a matrix to be analyzed. Several options are KM (correlation matrix), CM (covariance matrix), MM (moment matrix), AM (augmented moment matrix), and PM (polychoric or polyserial correlation matrix)
- **NG** is the number of groups
- **LA** specifies variables' labels. `&& LA; campus student q1 q2 q3;`
- Fixed format of the FORTRAN style. `&& KM SY; (5F3.0);` or `KM=corr_mat.txt FO; (5F3.0);` * FO indicates that the format follows on the specified FORTRAN style. SY represents a symmetric matrix.
- `RA=raw_data.txt FO` statement retrieves data in the form of a raw data matrix.
- **ME** (mean) and **SD** (standard deviation) `&& ME; .5 .8 .7 .8 .1;`
- **SE** enables us to alter the order in which the variables are read. The reference can be made either to their labels or their ordered number in the list of variables on the LA line. `&& SE; campus student q7 q10 q11;`

LISREL: MO (Model)

- `&& MO NX=4 NK=2 LX=FU,FI PH=SY,FR TD=DI,FR`
- **NK** and **NE** are the numbers of exogenous and endogenous latent variables.
- **NX** and **NY** are the numbers of observed exogenous and endogenous variables.
- **LX** and **LY** respectively represents lambda x and lambda y.
- **LK** and **LE** respectively represents labels for exogenous and endogenous latent variables. `&& LK; externality;`
- Available matrix forms are FU (full), SY (symmetric), DI (diagonal), ID (identity), and ZE (zero)
- Matrix estimation modes comprises FR (free) and FI (fixed). Free parameters are those whose values are unknown, and thus will be estimated. Fixed parameter are assigned some values a priori using VA or ST statement.
- Individual parameters within a particular matrix may be specified as fixed (FI), free (FR), or constrained equal to some other parameters (EQ). `&& FR LY(2,2) LX(2,2)`

- **ST**art and **VA**lue are used to refer to start values for iterations and the fixed values for the parameters. *&& VA 1.0 LY(2,2) LX(2,2); ST 1.0 LY(2,1) LX(2,1);*

LISREL: OU and Others

- Estimation methods include IV (instrumental variables), TSLS (two-stage least squares), ULS (unweighted least squares), GLS (generalized least squares), ML (maximum likelihood), WLS (weighted least squares), DWLS (diagonally weighted least squares). WLS requires an asymptotic covariance (AC) matrix, whereas DWLS method requires the asymptotic variance (AV) from the AC matrix.
- PD indicates the PATH DIAGRAM.

Missing Values

- FIML (Full Information Maximum Likelihood) considers missing values without dropping them in computation.
- EM (Expectation Maximization algorithm): pick up a starting value; randomly generate cases with new parameters; fill in the missing values with new ones generated with high expectation. In LISREL, Use STATISTICS → MULTIPLE IMPUTATION (including output option to specify the ASCII file, in which the output is stored).
- MCMC algorithm actually generates multiple data sets that have alternate values for missing. Estimate the model with different data sets. Compute variability of the model to choose the best data set.
- But LISREL for some reasons cannot impute more than 27 variables.

Simple

- “*SYSTEM FILE FROM FILE DEPRESS.DSF*” import the data set specified.
- “*LATENT VARIABLES*” declares latent variables
- “*RELATIONSHIPS*” defines the structural and measurement models.
- “*PATH DIAGRAM*” produces a path diagram for the model.
- “*LISREL OUTPUT: ND=3 SC ME=ML*” ME stands for “Method”. The option is for robust MLE. Without it, the result will be standard MLE.
- “*SET THE ERROR COVARIANCE BETWEEN endog1 AND endog2 FREE*” to test whether two error terms of endogenous variables are correlated or not. The null hypothesis is that they are not correlated.
- Correlation option for ordinal data, whose covariance is not defined. (Statistics-->Output Options)

Robust MLE (Maximum Likelihood Estimators)

WLS Estimators

MI (Modification Index)

In Simplis, 1.00 multiplied generate standardized variance covariance (covariance=0) matrix of latent variable.

Compare using imputed data set and robust or WLS method on original data set.

Normal score; Statistics → Normal Scores

Mean and variance remain the same; covariance changes.

Skewness and Kurtosis are improved

Reference

Byrne, Barbara M. 1998. *Structural Equation Modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, Applications, and Programming*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Toit, Mathilda du and Stephen du Toit. 2001. *Interactive LISREL: User's Guide*. SSI (Scientific Software International).

<http://luna.cas.usf.edu/~mbrannic/files/regression/SEM.html>

<http://www.ssicentral.com/lisrel/define.htm>

<http://gsu.edu/~mkteer/sem2.html>

<http://giorgio.catchword.com/vl=16443501/cl=35/nw=1/rpsv/catchword/erlbaum/10705511/v7n3/s5/p442>