A Gentle Introduction to Structural Equation Modeling

Matthew Sigal

Department of Psychology
York University

November 10, 2012

Introduction

- Structural Equation Modeling (SEM) is a relatively new statistical methodology that can be thought of as an extension of multiple regression.
- It is highly versatile and easily adopted for a variety of research designs, and hypotheses.
- It is a confirmatory approach to the analysis of an a priori structural theory pertaining to some phenomenon of interest.

The Purpose of Latent Variable Analysis

- One of the most notable features of SEM is the inclusion of *latent variables*.
- In research, we often know that we are not measuring exactly what we intend to, but we hope that our methods potentially tap that construct.
- Latent variables are exciting because they allow us to get a sense of the underlying construct through the use of multiple indicators.
- The inclusion of latent variable accounts for measurement error in the underlying construct!

Major Questions

The fundamental question posed in the SEM framework is: To what extent does our hypothesized model, which we have set up to account for all sorts of relationships between our variables of interest, actually fit the data?

How is that accomplished?

In SEM, the researcher provides two fundamental things:

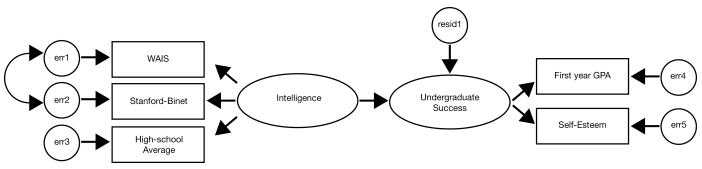
- A theoretical model, usually in the form of a path diagram.
- A sample covariance or correlation matrix.

Using the path diagram and the sample covariance matrix, SEM software will produce an estimated population covariance matrix.

The question then becomes: how well does your data match the estimated population covariance matrix from the model you have specified?

Path Diagrams

An example path diagram:



Summary of Model Information

Ensure that the model makes sense by looking at:

1. The Measurement Model and Structural Model separately and in conjunction.
2. Directionality between parameters.
3. The number of Fixed versus Freed loadings.

How is SEM different from other methodologies?

- Unlike Multiple Regression, SEM allows for more than one dependent variable.
- Unlike in Multivariate Multiple Regression, multiple DVs do not have to be a linear combination of each other.
- Unlike in Factor Analysis, SEM allows for the simultaneous examination of multiple relationships between factors.
- SEM is better than Multilevel Modeling at handling latent variables and larger (or more complex) models.

What other benefits does SEM allow for?

On top of the ability to model latent variables, SEM has other benefits for researchers:

- Some SEM software allows for graphical methods of modeling.
- We can test both global fit and individual parameters.
- We can have multiple outcomes and model relationships between error terms.
- We can test coefficients across groups.

Statistical Assumptions

As with any statistical procedure, appropriate inference for SEM requires researchers meet a few basic assumptions, pertaining to:

1. Sample Size.
2. Distributional Issues.
3. Complete Data.

Sample Size

- A popular area of research for QM, so kind of contentious.
- "Rule of Thumb": $N \geq 8 * K$, where K refers to the number of observed variables in the model.
- At a minimum, researchers should strive for at least 100 cases, although 200 is better.

Problems with Small n

- Estimation methods of population parameters will often fail to converge.
- May produce meaningless or impossible results (e.g. negative error variance estimates).
- May reduce the accuracy of parameter estimates and their standard errors.
- Will reduce the power for the tests of individual parameters.

Univariate and Multivariate Distributions

In a perfect world...

- Each dependent and mediating variable is assumed to be continuous and normally distributed.
- Each observed variable should be univariate normal.
- All observed variables should be jointly multivariate normal.

Yeah, right!

Distributional Solutions

Since all of those requirements are never met in practice, researchers often will:

1. Transform variables to help normalize them.
2. Use robust statistics instead of traditional parametric methods (only available in select software).
3. Use boot-strapping to get more precise estimates of standard errors (only available in select software).
4. Use methods that account for categorical mediators and outcomes (such as WLS, and polychoric correlations)

Complete Data

The final basic requirement of SEM pertains to having a dataset with no missing values.

- If there are, can we assume that the data missing at random?
- If so, researchers will often use algorithms to estimate the missing values (e.g. FIML).
- If not, this poses a grave problem for inference.
- **Never** just delete data or use mean substitution methods!

Model Specification

Once those assumptions have been looked at, a researcher will want to specify their model. This entails:

- Creating a path diagram, ensuring that each latent variable has, *at minumum*, two indicators;
- Determining a modeling strategy (confirmatory, assessing competing models, or exploratory).

and

- Ensuring that the model is "identifiable" (has positive degrees of freedom).

Model Identification

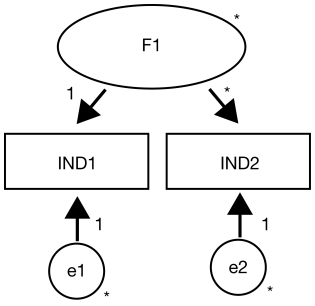
Test of degrees of freedom:

- $df = [k(k+1)/2\text{-}q]$, where k is the number of observed variables and q is the number of estimated parameters in the model. SEM requires over-identification for tests of global fit.

Latent Variable Scaling:

- As previously mentioned, each latent variable should either have a) its variance fixed to 1; or b) a regression coefficient from the factor to one of its observed indicators to 1.

Example of Model Identification



Information Available:

- $[k(k+1)/2] = 2(2+1)/2 = 3$.

Parameters to be Estimated:

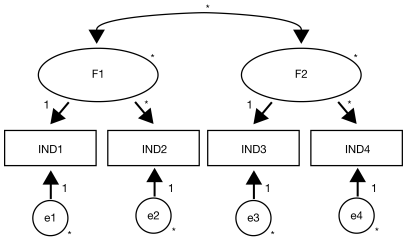
- 1 factor variance.
- 1 factor loading.
- 2 indicator residuals.
- Total: 4 parameters.

Model Degrees of Freedom:

- $df = 3 - 4 = -1$.
- The model is under-identified.

Example of Model Identification, Cont.

However, with more information...



Information Available:

- $4(4+1)/2 = 10$.

Parameters to be Estimated:

- 2 factor variances.
- 1 factor covariance.
- 2 factor loadings.
- 4 indicator residuals.
- Total: 9 parameters.

Model Degrees of Freedom:

- $df = 10 - 9 = 1$.
- The model is over-identified!

Model Assessment

Once we have our model, parameter estimates are calculated by the SEM software (typically using maximum likelihood estimation). It is then up to the researcher to investigate two areas:

- Whether any element of the residual covariance matrix differs from zero;
- And, if any individual coefficients differ from zero.

The first is the test of global fit, which is tested via a likelihood ratio test and distributed as a χ^2 with $(k - q)$ degrees of freedom.

Global Fit

However, the above test is overly sensitive to sample size. Some better tests of global fit are:

- The Goodness of Fit Index (GFI), akin to multiple R^2 in multiple regression (want a value $> .95$).
- The Adjusted Goodness of Fit Index (AGFI), similar to the GFI but accounts for model parsimony (want a value $> .90$).
- The Root Mean Squared Error of Approximation (RMSEA), accounts for differences between the variances and covariances of sample and estimated population matrix; adjusts for model parsimony; and, confidence intervals are available (want a value $< .05$).

Global Fit, Cont.

If you are comparing models, you should look at the Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC). These values will be produced by the statistical software, and you will want to look for the model with the smallest values on these scales.

Model Modification

Model Modification is **a purely exploratory technique!** The goal is to find a better fitting model than the one you originally hypothesized. To do so:

- Always start with changes that make theoretical sense - not because the software tells you to do so.
- Only make one change at a time (for instance, adding or removing a path; fixing a coefficient)
- To find values to change: look at residuals, modification indices.

Model Modification, Cont.

However, some questions arise from this procedure:

- Do the modifications you've made make theoretical sense?
- How many modifications should a researcher be allowed to make?
- Should multiplicity control be imposed?
- How do the modifications made influence other parameters in the model?

Demonstration and Conclusion

A LGC Example Using AMOS

This is an example of a Latent Growth Curve model that is especially useful for developmental research. In this example, the change in anti-social behaviour over a four year is of clinical importance. We primarily are concerned with three questions:

1. What is the over-time trajectory for the group? (e.g. what is the mean trajectory of child anti-social behaviour?)
2. Do we need distinct trajectories for individual cases? (e.g. do children differ in their initial levels and rates of change?)
3. If we do need distinct trajectories, can we find variables to predict them for groups of individuals?

Other Programs

SEM procedures are also available in:

- M+
- LISREL
- SAS (via PROC CALIS)
- R (via the "sem", "laavan", or "OpenMX" packages)
- EQS
- STATA (as of version 12)

Introductory SEM Resources

- Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York, NY: Wiley.
- Byrne, B. B. (2010). *Structural Equation Modeling with AMOS (2nd ed.)*. New York, NY: Routledge Press.
- Kline, R. B. (2010). *Principles and Practices of Structural Equation Modeling (3rd ed.)*. New York, NY: Guilford Press.
- Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. (2008). *Latent Growth Curve Modeling*. Thousand Oaks, California: SAGE Publications.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariate Statistics (5th ed.)*, Chapter 14. New York, NY: Allyn and Bacon.

Desiderata for Structural Equation Modeling	Manuscript Section(s)*
1. Substantive theories that led to the model(s) being investigated are synthesized; a set of a priori specified competing models is generally preferred.	I
2. Path diagrams are presented to facilitate the understanding of the conceptual model(s) and the specification of the statistical model(s).	I
3. If applicable, latent factors are defined and their status as latent (vs. emergent) is justified	I, M
4. Measured variables are defined and, if applicable, their appropriateness as indicator variables of associated factors is justified.	M
5. Latent factors are indicated by a sufficient number of appropriately measured variables; how the latent factors are given scale within the model(s) is addressed.	M
6. How theoretically relevant control variables are integrated into the model is explained,	M
7. Sampling method(s) and sample size(s) are explicated and justified.	M
8. The treatment of missing data and outliers is addressed.	M, R
9. The name and version of the utilized software package is reported; the parameter estimation method is justified and its underlying assumptions are addressed.	M, R
10. Problems with model convergence, offending estimates, and/or model identification are reported and discussed.	R
11. Summary statistics of measured variables are presented; if raw data were analyzed, information on how to gain access to the data is provided.	R
12. For models involving structural relations among latent variables, a two-phase (measurement, structural) analysis process is followed and summarized.	R
13. Recommended data-model fit indices from multiple classes are presented and evaluated using literature-based criteria.	R
14. For competing models, comparisons are made using statistical tests (for nested models) or information criteria (for non-nested models).	R
15. For any post hoc model re-specification, theoretical and statistical justifications are provided.	R
16. Latent factor quality is addressed in terms of validity and reliability.	R
17. Standardized and unstandardized parameter estimates together with information regarding their statistical significance are provided; R^2 values for key structural outcomes are presented.	R, D
18. Appropriate language regarding model tenability and structural relations is used.	D

* Note: I = Introduction, M = Methods, R = Results, D = Discussion

Hancock, G. R., & Mueller, R. O. (2010). "Chapter 26: Structual Equation Modeling" (p. 372) in *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (Hancock, G.R., & Mueller, R.O. [eds]). New York, NY: Routledge.

LGCA Analysis Summary

Groups - Group number 1 (Group number 1) CONSTRAINED VARIANCES
The model is recursive.
Sample size = 221

Variable Summary (Group number 1)
Your model contains the following variables (Group number 1)
Observed, endogenous variables
anti1 anti2 anti4 anti3

Unobserved, endogenous variables
int slope

Unobserved, exogenous variables
e1 e2 e3 e4 ivar svar

Variable counts (Group number 1)

Number of variables in your model:	12
Number of observed variables:	4
Number of unobserved variables:	8
Number of exogenous variables:	6
Number of endogenous variables:	6

Parameter summary (Group number 1)

	Weights	Covariances	Variances	Means	Intercepts	Total
Fixed	12	0	2	0	0	14
Labeled	0	0	4	0	2	6
Unlabeled	2	1	0	0	0	3
Total	14	1	6	0	2	23

Computation of degrees of freedom (Default model)

Number of distinct sample moments:	14
Number of distinct parameters to be estimated:	6
Degrees of freedom (14 - 6):	8

Result (Default model)
Minimum was achieved
Chi-square = 5.559
Degrees of freedom = 8
Probability level = .696

Scalar Estimates (Group number 1 - Default model) - Maximum Likelihood Estimates
Regression Weights: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P
int	<---	ivar	.982	.105	9.317	***
slope	<---	svar	.310	.070	4.424	***
anti1	<---	int	1.000			
anti2	<---	int	1.000			
anti3	<---	int	1.000			
anti2	<---	slope	1.000			
anti3	<---	slope	2.000			
anti4	<---	slope	3.000			
anti1	<---	slope	.000			
anti4	<---	int	1.000			

Intercepts: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P	Label
int	1.554	.096	16.163	***	intm
slope	.176	.043	4.128	***	slm

Covariances: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
svar	<-->	ivar	.494	.341	1.447	.148	

Variances: (Group number 1 - Default model)

	Estimate	S.E.	C.R.	P	Label
ivar	1.000				
svar	1.000				
e1	1.529	.103	14.832	***	th1
e2	1.529	.103	14.832	***	th1
e3	1.529	.103	14.832	***	th1
e4	1.529	.103	14.832	***	th1

Unconstrained Variances (Unconstrained Variances)

Notes for Model (Unconstrained Variances)

Computation of degrees of freedom (Unconstrained Variances)

Number of distinct sample moments:	14
Number of distinct parameters to be estimated:	9
Degrees of freedom (14 - 9):	5

Result (Unconstrained Variances)

Minimum was achieved
Chi-square = 3.146
Degrees of freedom = 5
Probability level = .678

Scalar Estimates (Group number 1 - Unconstrained Variances) - Maximum Likelihood Estimates

Regression Weights: (Group number 1 - Unconstrained Variances)

			Estimate	S.E.	C.R.	P
int	<---	ivar	.996	.118	8.471	***
slope	<---	svar	.324	.082	3.959	***
anti1	<---	int	1.000			
anti2	<---	int	1.000			
anti3	<---	int	1.000			
anti2	<---	slope	1.000			
anti3	<---	slope	2.000			
anti4	<---	slope	3.000			
anti1	<---	slope	.000			
anti4	<---	int	1.000			

Intercepts: (Group number 1 - Unconstrained Variances)

	Estimate	S.E.	C.R.	P	Label
int	1.545	.096	16.117	***	intm
slope	.179	.043	4.191	***	slm

Covariances: (Group number 1 - Unconstrained Variances)

			Estimate	S.E.	C.R.	P
svar	<-->	ivar	.413	.386	1.070	.285

Variances: (Group number 1 - Unconstrained Variances)

	Estimate	S.E.	C.R.	P	Label
ivar	1.000				
svar	1.000				
e1	1.397	.234	5.972	***	th1
e2	1.737	.198	8.790	***	th2
e3	1.362	.181	7.544	***	th3
e4	1.572	.281	5.601	***	th4

Model Fit Summary

CMIN

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	6	5.559	8	.696	.695
Unconstrained Variances	9	3.146	5	.678	.629
Saturated model	14	.000	0		
Independence model	8	255.507	6	.000	42.584

Baseline Comparisons

Model	NFI Delta1	RFI rho1	IFI Delta2	TLI rho2	CFI
Default model	.978	.984	1.010	1.007	1.000
Unconstrained Variances	.988	.985	1.007	1.009	1.000
Saturated model	1.000		1.000		1.000
Independence model	.000	.000	.000	.000	.000

Parsimony-Adjusted Measures

Model	PRATIO	PNFI	PCFI
Default model	1.333	1.304	1.333
Unconstrained Variances	.833	.823	.833
Saturated model	.000	.000	.000
Independence model	1.000	.000	.000

NCP

Model	NCP	LO 90	HI 90
Default model	.000	.000	6.534
Unconstrained Variances	.000	.000	5.858
Saturated model	.000	.000	.000
Independence model	249.507	200.886	305.544

FMIN

Model	FMIN	F0	LO 90	HI 90
Default model	.025	.000	.000	.030
Unconstrained Variances	.014	.000	.000	.027
Saturated model	.000	.000	.000	.000
Independence model	1.161	1.134	.913	1.389

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.000	.000	.061	.907
Unconstrained Variances	.000	.000	.073	.862
Independence model	.435	.390	.481	.000

AIC

Model	AIC	BCC	BIC	CAIC
Default model	17.559	17.838		
Unconstrained Variances	21.146	21.564		
Saturated model	28.000	28.651		
Independence model	271.507	271.879		

ECVI

Model	ECVI	LO 90	HI 90	MECVI
Default model	.080	.091	.121	.081
Unconstrained Variances	.096	.105	.131	.098
Saturated model	.127	.127	.127	.130
Independence model	1.234	1.013	1.489	1.236

HOELTER

Model	HOELTER.05	HOELTER.01
Default model	614	796
Unconstrained Variances	775	1056
Independence model	11	15

Nested Model Comparisons - Assuming model Unconstrained Variances to be correct:

Model	DF	CMIN	P	NFI Delta-1	IFI Delta-2	RFI rho-1	TLI rho2
Default model	3	2.414	.491	.009	.010	.002	.002

Execution time summary

Minimization:	.032
Miscellaneous:	.452
Bootstrap:	.000
Total:	.484

NOTE: AMOS does not report GFI, AGFI, PGFI, RMR, or BIC fit statistics when means and intercepts are estimated (which they always are in LGC models). If this was not a growth curve model, they would be reported!