Path Analysis and Structural Equation Models (SEM): Options for Advanced Analysis of Relationships

Mark Fossett
Sociology, Texas A&M University
m-fossett@tamu.edu

Sociology Quantitative Methods Series, Summer 2021
Overview of the Presentation

The goal is to provide a brief introduction to the methods of path analysis and structural equation modeling to give perspective on the potential value of these methods.

Note what is involved in Path Analysis and SEM.
Review the potential benefits of the methods.
Note the limitations of the methods.
Give attention to a few of the useful applications of these frameworks including, for example:
  - decomposing effects
  - making “hidden” measurement assumptions explicit
  - making “hidden” causal assumptions explicit
  - using simple models to estimate effects
Review selected results from analyses performed using demonstration programs.
Anticipating Conclusions

Even if one does not use the methods, knowledge of Path Analysis and SEM can provide perspective that can help one estimate and interpret analysis results in a more careful and thoughtful way.

This somewhat obviously applies to quantitative analysis.

It applies equally, if less obviously, to qualitative analysis.

Benefits

SEM provides a frame work for expressing theory and refining empirical analysis.

- graphical representations provide useful heuristics
- opportunities for quantitative refinements

Highlights issues of measurement and measurement error

Highlights issues of causal associations among X’s

Highlights distinctions between “controls” and “mediators”

Highlights distinctions between Direct and Indirect Effects

Can be used to investigate complex models that cannot be investigated with a single equation analysis
Assumptions

This presentation presumes familiarity with:
  Multiple regression analysis

A full introduction to SEM methods would require one or two courses.
  That is not the goal here.

The goal here is to raise awareness about the methods and highlight some of the things that can be accomplished using just knowledge of OLS multiple regression analysis
What are Path Analysis and SEM?

Structural Equation Modeling (SEM): A broad modeling framework that enables researchers to investigate an expanded set of issues not normally considered in “standard” regression analyses.

The following are a few of the more important features of the SEM framework.

- It permits investigation of measurement models for dependent and independent variables
- It permits investigation of direct and indirect effects within a system of inter-related variables
- It permits investigation of reciprocal causation
- It permits investigation of feedback loops within systems of variables

Path Analysis: A subset of SEM methods that can be used to investigate some of the issues addressed in SEM.

Methods of path analysis are less complex and can be used without mastering the full SEM framework.
Features of “Regular” Single-Equation Regression

The basic multiple regression

\[ Y = b_0 + b_1X_1 + b_2X_2 \ldots b_kX_k \]

A single Dependent Variable (DV) denoted by Y
One or more Independent Variables (IV’s) denoted by X’s
X’s effect Y
Impacts of X’s are reflected in estimated values of b’s
Implicit, But Usually Unstated Assumptions

Variables are measured without error

A single-equation analysis reveals all that is relevant
   The main concern is correct specification of the single equation

No causal associations among independent variables (X’s)
   Associations (correlations) among X’s has no implications for substantive interpretation and can be ignored.

b’s reflect the impact on Y when X’s change

X’s affect Y, Y does not affect X’s
   No reciprocal causation or feedback loops

Errors are uncorrelated

IMPORTANTLY, THE ASSUMPTIONS ARE ASSERTED, NOT TESTED.
   They often are questionable. SEM forces an explicit review.
Added Features/Possibilities with SEM

Variables MAY be measured with error

Independent variable MAY be connected via causal associations
   One or more X’s is a DV relative to one or more other X’s

Changes in X’s MAY bring about changes in Y beyond what is indicated by b’s
   X’s MAY have both Direct Effects and Indirect Effects

Effects MAY be non-recursive (e.g., may involve reciprocal causation and/or feedback loops)

Errors MAY be correlated
The SEM framework brings major changes to analyses.

1. Estimation of a “Structural Model”
   This roughly corresponds to a regular regression
   However, Y and X’s are “Latent Variables”
   Latent variables are not observed (they are not in the data set)

2. Estimation of an associated set of “Measurement Models”
   There is one measurement model for each latent variable
   There is no analog in regular regression
   Indicator variables (y and x’s) are viewed as dependent variables, “products” caused by latent variables
   Indicator variables are observed (they are in the data set)

3. The Structural and Measurement Models are estimated together as part of a single analysis of a full “system”
Specialized Software for SEM Keeps Improving

Historically, researchers have had to master complex, esoteric software to conduct analysis using SEM.

The first dominant program

LISREL
Well-deserved reputation for obscure notation, obtuse syntax, uninformative error messages, and fragile estimation routines.
BUT IT WORKED!

Second-Generation programs, easier, specialized capabilities

EQS
Amos
CALIS (SAS)
MPLUS

Latest iterations, easier still, simplified model building tools

SEM (Stata)
To investigate of these additional possibilities, the mathematical formulation of the analysis is elaborated with many more terms. One influential implementation of SEM uses the following terms.

<table>
<thead>
<tr>
<th>LISREL notation</th>
<th>LISREL parameter specification</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x )</td>
<td>–</td>
<td>An observed exogenous variable</td>
</tr>
<tr>
<td>( y )</td>
<td>–</td>
<td>An observed endogenous variable</td>
</tr>
<tr>
<td>( \eta )</td>
<td>eta</td>
<td>An endogenous latent variable</td>
</tr>
<tr>
<td>( \xi )</td>
<td>ksi</td>
<td>An exogenous latent variable</td>
</tr>
<tr>
<td>( \zeta )</td>
<td>zeta</td>
<td>A residual of ( f )</td>
</tr>
<tr>
<td>( \epsilon )</td>
<td>epsilon</td>
<td>A residual measurement error component of ( y )</td>
</tr>
<tr>
<td>( \delta )</td>
<td>delta</td>
<td>A residual measurement error component of ( x )</td>
</tr>
<tr>
<td>( \lambda_y )</td>
<td>lambda ( y ) (LY)</td>
<td>A factor loading of ( f ) on ( y )</td>
</tr>
<tr>
<td>( \lambda_x )</td>
<td>lambda ( x ) (LX)</td>
<td>A factor loading of ( F ) on ( x )</td>
</tr>
<tr>
<td>( \Theta_e )</td>
<td>theta epsilon (TE)</td>
<td>Dispersion matrix of ( e )</td>
</tr>
<tr>
<td>( \Theta_d )</td>
<td>theta delta (TD)</td>
<td>Dispersion matrix of ( d )</td>
</tr>
<tr>
<td>( \beta )</td>
<td>beta (BE)</td>
<td>Matrix of path coefficients between ( f )</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>gamma (GA)</td>
<td>Matrix of path coefficients from ( F ) to ( f )</td>
</tr>
<tr>
<td>( \Psi )</td>
<td>psi (PS)</td>
<td>Dispersion matrix of ( z )</td>
</tr>
<tr>
<td>( \Phi )</td>
<td>phi (PH)</td>
<td>Dispersion matrix of ( F )</td>
</tr>
</tbody>
</table>
### Notation for Variables (per Stata SEM)

**Variable Notation**

- **Observed endogenous variables** \( y \)
- **Observed exogenous variables** \( x \)
- **Latent endogenous variables** \( \eta \)
- **Latent exogenous variables** \( \xi \)
- **Error variables for observed endogenous variables** \( e_y \)
- **Error variables for latent endogenous variables** \( e_\eta \)

**Vectors**

<table>
<thead>
<tr>
<th>Endogenous Variables</th>
<th>Exogenous Variables</th>
<th>Error Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y = (y_\eta) )</td>
<td>( X = (x_\xi) )</td>
<td>( \zeta = (e_y e_n) )</td>
</tr>
</tbody>
</table>
Notation for the Basic Model (per Stata SEM)

The Basic Model

\[ Y = BY + \Gamma X + \alpha + \zeta \]

where

\[ B = [\beta_{ij}] \text{ matrix of coefficients on endogenous variables} \]
\[ \Gamma = [\gamma_{ij}] \text{ matrix of coefficients on exogenous variables} \]
\[ \alpha = [\alpha_i] \text{ vector of intercepts for endogenous variables} \]
\[ \zeta \text{ has mean 0 and } Cov(X, \zeta) = 0 \]

Notation for other aspects of the full estimation model

There are more elements in the full SEM estimation model. These elements are important. But they are omitted here in the interests of brevity.

(See Stata SEM documentation “Methods and Formulas” for a more complete review. Alternatively, see Fox (2012: p8-11).
Proponents of Path Analysis and SEM wisely adopted graphical methods – flowgraph analysis – to make the complex SEM models intelligible to the researcher and especially to broader audiences.

Flowgraph conventions are only norms and, like model notation, they are not adopted universally. But most approaches adopt the following conventions.

**Widely Adopted Conventions**
- Latent Variables – Ovals or circles (bolded)
- Indicator Variables – Squares or rectangles (thin)
- Error Terms – small circles (thin)
- Causal Effects – straight lines with arrow heads to indicate direction of effects
- Non-Causal Association – curved lines with double-arrow heads
Graphical representation helps make the model easier to grasp.
A multi-indicator measurement model. Models can be evaluated using confirmatory factor models to test hypothesized measurement relationships.
An influential model from early days of SEM methods; Contribution is to facilitate the estimation of reciprocal effects.
Graphical Representation – An Example

An influential model from early days of SEM methods. Contribution is to estimate relationships more reliably via multi-indicator measurement models.
Stata provides the “SEM Builder” tool to aid in developing SEM models via graphical representation of the model. (Within Stata, use “help sem” to access relevant documentation.)

Description

The SEM Builder lets you create path diagrams for SEMs, fit those models, and show results on the path diagram. Here we discuss standard linear SEMs; see [SEM] Builder, generalized for information on using the Builder to create models with generalized responses and multilevel structure.

Remarks and examples

Launch the SEM Builder by selecting the menu item Statistics > SEM (structural equation modeling) > Model building and estimation. You can also type sembuilder in the Command window.
# Estimation, Effect Testing, and Model Evaluation

<table>
<thead>
<tr>
<th></th>
<th>&quot;Regular&quot; Single-Equation Analysis</th>
<th>Structural Equation Model Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Estimation Method</strong></td>
<td>Least Squares and Maximum Likelihood</td>
<td>Maximum Likelihood</td>
</tr>
<tr>
<td><strong>Effect evaluation</strong></td>
<td>t and Z tests</td>
<td>Z tests</td>
</tr>
<tr>
<td><strong>Model evaluation</strong></td>
<td>$R^2$ and likelihood ratio $X^2$</td>
<td>same; but global and also for submodels</td>
</tr>
<tr>
<td><strong>Model objective</strong></td>
<td>predict $y$ for individual cases</td>
<td>predict associations among observed variables (terms of the var/cov matrix)</td>
</tr>
<tr>
<td><strong>Predictions</strong></td>
<td>values of $y$</td>
<td>values of var/cov terms</td>
</tr>
<tr>
<td><strong>Errors</strong></td>
<td>misprediction of $y$ for cases</td>
<td>misprediction of values of var/cov terms</td>
</tr>
</tbody>
</table>
Simple Path Analysis

Here “simple” path analysis involves a full model with two or more multiple regression equations representing the relationships among a system of variables. The system has multiple stages; that is, some variables are both exogenous (IV’s) and endogenous (DV’s).

The Model is Fully Recursive

Causation is unidirectional (no reciprocal causation or loops)

The effects can be estimated by regular OLS regression. Thus,

Each regression involves “single-indicator” variables

This implies a perfect one-to-one correspondence between the latent variable and a single observed variable

Path coefficients correspond to regression coefficients (unstandardized or standardized).

Error terms are assumed to be uncorrelated.
“Simple” Path Analysis – Opportunities

The full model specification requires explicit examination of possible causal associations among independent variables (X’s).

Single equation regression analyses often neglect this issue.

Ability to reveal Direct, Indirect, and Total Effects

(D) Direct Effects on Y – captured by b’s in regression of Y

(I) Indirect Effects on Y – captured by multiplying b’s in a sequence of regression equations

(T) Total Effects on Y – the sum of direct and indirect effects

Value of D-I-T

Highlights potential for unanticipated consequences.

Highlights hidden causal effects in single-equation analyses.

Creates opportunities for theory refinement and hypothesis testing by investigating effects of mechanisms.

Ability to decompose associations among variables.

Spurious Association (S) can be identified. Total Association (correlation) can be partitioned into D-I-T-S components.
"Simple" Path Analysis – I

The example program reviews a sequence of path models involving three variables measured in standard (z-score) form:
- $X_1 = \text{EDUC} (\text{educ6} – \text{education on 0-5 scale})$
- $X_2 = \text{OPPINT} (\text{oppint} – \text{oppose integration (0 no–yes 1)})$
- $X_3 = \text{POLVIEW} (\text{polview} – \text{political view (0 lib–con 1)})$

Model 0, latent variable = observed variables, no causal links

Associations (Correlations)
```
. corr x1 x2 x3
(obs=6,769)
```

<table>
<thead>
<tr>
<th></th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>-0.3082</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td>-0.0455</td>
<td>0.1158</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
"Simple" Path Analysis – II

Non-causal associations are estimated by OLS correlations; effects (causal links) are estimated by OLS regression coefficients. Regression equations reflect links as depicted in the model.

Model 1, $X_3$-POLVIEW = $X_1$-EDUC

Equations & Regressions

$$[EQ1] X_3 = b_{30} + b_{31}X_1$$

reg x3 x1

$b_{31} = -0.0455$ (p < 0.01)

$X_1$-EDUC has significant causal effect on $X_3$-POLVIEW

Note: Load the example “.stsem” files in Stata to view the model in the Stata SEM Builder tool.
“Simple” Path Analysis – III

Non-causal associations are estimated by OLS correlations; effects (causal links) are estimated by OLS regression coefficients. Regression equations reflect links as depicted in the model.

Model 2, $X_3$-POLVIEW = $X_1$-EDUC, $X_2$-OPPINT

Equations & Regressions

$[\text{EQ1}] \quad X_3 = b_{30} + b_{31}X_1 + b_{32}X_1$

$\text{reg } x3 \ x1 \ x2$

$b_{31} = -0.0095 \ (p \approx 0.45)$

$b_{32} = 0.1196 \ (p < 0.01)$

$X_1$-EDUC has no causal effect on $X_3$-POLVIEW
Non-causal associations are estimated by OLS correlations; effects (causal links) are estimated by OLS regression coefficients. Regression equations reflect links as depicted in the model.

Model 3, $X_3 = X_1, X_2 \text{ and } X_2 = X_1$

Equations & Regressions

[EQ1] $X_3 = b_{30} + b_{31}X_1 + b_{32}X_1$

[EQ2] $X_2 = b_{20} + b_{21}X_1$

$\text{reg } x3 \ x1 \ x2; \text{ reg } x2 \ x1$

$b_{31} = -0.0095 \ (p \approx 0.45)$

$b_{32} = 0.1196 \ (p < 0.01)$

$b_{21} = -0.3010 \ (p < 0.01)$

$X_1$-EDUC has total causal effect on $X_3$-POLVIEW of $-0.0455$ (from $b_{31} + b_{21}b_{32} = -0.0095 + (-0.3010)(-0.1196) = -0.0455$)
“Simple” Path Analysis – V

Takeaway point

How effects are estimated and interpreted depends critically on the assumptions regarding the nature of the associations among the variables in the model. EDUC has no effect on POLVIEWS under Model 2 and negative effects under M

Policy Relevance

The effect of EDUC on POLVIEWS involves indirect as well as direct effects.

Practical Relevance

The total effect of EDUC on POLVIEWS under Model 3 of -0.0455 is exactly the same as the direct effect of EDUC on POLVIEW in Model 1.

This is because, under Model 3, Model 1 provides a “reduced form” estimate of the total effect of EDUC on POLVIEW. (Under Model 3, there is no need to control for OPPINT when estimating the total effect of EDUC on POLVIEW.)
Does EDUC affect POLVIEW? Under Model 2, the answer is **NO**. The association of EDUC and OPPINT is viewed as spurious. So the regression OPPINT = EDUC – **IS NOT** relevant. Thus, OPPINT *should be* controlled when estimating total effect of EDUC on POLVIEW.
Does EDUC affect POLVIEW? Under Model 2, the answer is **YES**. The association of EDUC and OPPINT is viewed as causal.

So the regression OPPINT = EDUC is relevant.

Also it is not necessary to control OPPINT to estimate the Total Effect of EDUC on POLVIEW (only to estimate Direct Effect).
Implications ...

It may be possible to estimate causal effects in an analysis where some valid predictors of the ultimate dependent variable are not included in the model.

If the omitted predictors are seen as mediating variables, the analysis can yield valid estimates of total effects.

Criticism that relevant predictors are omitted from the regression would be misplaced, *IF* the model of mediation.

Thus, OPPINT can be omitted in the reduced form regression for Model 1, if the goal is to estimate the Total Effect of EDUC on POLVIEW. (Because OPPINT mediates the effect of EDUC.)

Note:

The key is that the goal of the modeling effort is NOT to predict as much variation in the DV as possible. Instead, the goal of the modeling effort is to correctly estimate the causal effect of EDUC on POLVIEW.
Using SEM to Consider D-I-T in Detail

The full SEM goes beyond the “manual” calculations based on results of OLS regressions for Simple Path Analysis.

SEM software can be used to:
- Estimate Direct, Indirect, and Total Effects
- Place confidence intervals on estimates of effects
- Permit statistical tests of effects

Example program “w14_path_sem1.do” provides examples.
- The regressions are estimated as part of a structural equation model specified with single-indicators and perfect measurement.
- Regression results are the same.
- The “estat teffects” post-estimation command yields results for calculations of direct, indirect, and total effects with associated statistical tests for each kind of effect.
Using SEM to Consider Measurement Error – I

In OLS regression, the measurement models are assumed to be perfect. Thus, in the SEM representation of the OLS regressions

For educ = EDUC, b=1.0, R²=1.0, & ε₃=0.0.
For oppint = OPPINT, b=1.0, R²=1.0, & ε₂=0.0.
For polview = POLVIEW, b=1.0, R²=1.0, & ε₁=0.0.

What if

educ has random measurement error? R²=0.60-0.90 and/or
oppint has random measurement error? R²=0.60-0.90 and/or
polview has random measurement error? R²=0.60-0.90.

Even if one is uncertain about measurement error, SEM permits investigation of the implications of different, possibly plausible, scenarios.

Example program “w14_path_sem1.do” provides examples
Using SEM to Consider Measurement Error – II

General Expectation

Associations among variables are attenuated (biased toward zero) by the presence of random measurement error.

All else equal, incorporating accurate assessments of random error in observed variables in SEM models will lead to larger estimates of effects.

A Major Caution

In a multivariate model, the consequences of random measurement error (the simplest kind), can be complicated. If measurement error is differential – that is, if it varies in degree across variables, consequences can be difficult to anticipate.
Using SEM to Consider Measurement Error – III

The example program compares the following two scenarios.

(A) No measurement error
   educ has NO random measurement error? \( R^2 = 0.99 \)
   oppint has NO random measurement error? \( R^2 = 0.99 \)
   polview has NO random measurement error? \( R^2 = 0.99 \)

(B) Differential measurement error
   educ has partial random measurement error? \( R^2 = 0.90 \)
   oppint has partial random measurement error? \( R^2 = 0.60 \)
   polview has partial random measurement error? \( R^2 = 0.75 \)

Impact on coefficients

EDUC => POLVIEW ...... (A) -0.009, (B) 0.024
EDUC => OPPINT ........ (A) -0.314, (B) -0.424
OPPINT >= POLVIEW ... (A) 0.117, (B) 0.187
Conclusions – Summing Up

Even if one does not use the methods, knowledge of Path Analysis and SEM can provide valuable perspective so analysis results are interpreted in a more careful and thoughtful way.

This applies to quantitative analysis.

It applies equally, if less obviously, to qualitative analysis.

Benefits

SEM provides a framework for expressing theory and refining empirical analysis.

- Graphical representations provide useful heuristics

- Opportunities for quantitative refinements

Highlights issues of measurement error

Highlights issues of causal associations among X’s

Highlights distinctions between “controls” and mediators

Highlights distinctions between Direct and Indirect Effects

Can be used to investigate complex models that cannot be investigated with a single equation analysis
Limitations

SEM is not a panacea
It cannot overcome poor theory and severely flawed data

SEM analysis is not easy
Drawing flowgraphs of complex models is relatively easy.
Obtaining sound estimates of the models is hard.
Software is getting better, but it is still hard to master

SEM estimation options are limited
The methods are most well developed for interval variables.
It is harder to work with categorical and bounded variables

SEM assumptions are demanding
SEM sample size requirements are demanding

Establishing the viability of innovative models (e.g., verifying the model is identified) can be demanding
Parting Advice

It is worth your while to learn about and explore SEM methods.

Knowledge of SEM methods expands imagination and perspective. So, learning about SEM brings benefits whether or not one uses it.

If using SEM in a serious way, have realistic expectations.

- Expect to take courses and/or workshops
- Expect a fairly steep learning curve for learning how to formulate models and estimate them using statistical software
- Expect to spend a long time with the modeling effort. There will be endless possibilities and tweaks to consider.

DON’T BE INTIMIDATED !!!

- Unfamiliar notation, complex modeling options, and esoteric software can all be intimidating.
- Remember. In the final analysis, it is just a more thoughtful form of regression analysis.
End of Slides

Thank you for your attention.