A Bit of SEM

Advanced Master in Agricultural Economics and Policy

Giovanbattista Califano

University of Naples Federico II, giovanbattista.califano@unina.it

Attitudinal Scales for the Study of Consumer Preferences

April 30, 2026

Roadmap

20/04	Hello, Psychometrics!	✓
23/04	The Questionnaire	✓
24/04	Reliability and Validity of a Measure	✓
27/04	Latent Variables: Reflective or Formative?	✓
30/04	A Bit of SEM	◯
07/05	Stata Stata Stata Stata Stata Stata Stata	◯

Recommended readings:

- Chapters 4 and 7 from Jhangiani et al.2019)
- Chapters 7, 10 and 14 from Olivero and Russo2022)
- Chapter 12 from Mehmetoglu and Jakobsen2022)

Jhangiani, R. S., Chiang, I. A., Cuttler, C., and Leighton, D. C. (2019). *Research Methods in Psychology*. Kwantlen Polytechnic University, 4th edition.

Mehmetoglu, M. and Jakobsen, T. G. (2022). *Applied Statistics Using Stata: A Guide for the Social Sciences*. SAGE Publications Ltd, 2nd edition.

Olivero, N. and Russo, V. (2022). *Psicologia dei Consumi*. McGraw-Hill Education, 3rd edition.

Before we begin...

https://www.stata.com/manuals/sem.pdf

Before we begin...

Remarks and examples

Structural equation modeling encompasses a broad array of models from linear regression to measurement models to simultaneous equations, including along the way confirmatory factor analysis (CFA), correlated uniqueness models, latent growth models, multiple indicators and multiple causes (MIMIC) models, and item response theory (IRT) models.

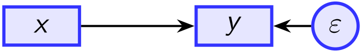
Structural equation modeling is not just an estimation method for a particular model in the way that Stata's `regress` and `probit` commands are, or even in the way that `stcox` and `mixed` are. Structural equation modeling is a way of thinking, a way of writing, and a way of estimating.

Before we begin...

Remarks and examples

Structural equation modeling encompasses a broad array of models from linear regression to measurement models to simultaneous equations, including along the way confirmatory factor analysis (CFA), correlated uniqueness models, latent growth models, multiple indicators and multiple causes (MIMIC) models, and item response theory (IRT) models.

Structural equation modeling is not just an estimation method for a particular model in the way that Stata's `regress` and `probit` commands are, or even in the way that `stcox` and `mixed` are. Structural equation modeling is a way of thinking, a way of writing, and a way of estimating.

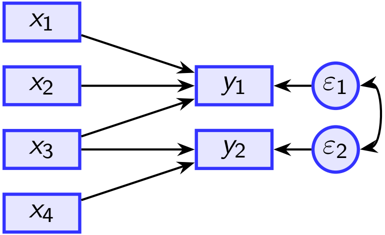


Before we begin...

Remarks and examples

Structural equation modeling encompasses a broad array of models from linear regression to measurement models to simultaneous equations, including along the way confirmatory factor analysis (CFA), correlated uniqueness models, latent growth models, multiple indicators and multiple causes (MIMIC) models, and item response theory (IRT) models.

Structural equation modeling is not just an estimation method for a particular model in the way that Stata's `regress` and `probit` commands are, or even in the way that `stcox` and `mixed` are. Structural equation modeling is a way of thinking, a way of writing, and a way of estimating.



Moving on to...

A Bit of Terminology

A Bit of History

A Bit of Theory

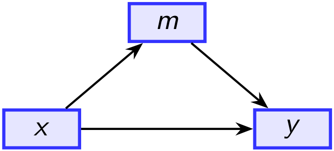
Types of SEM

- Structural equation modelling allows us to estimate relationships between observed and/or latent variables through the study of the variance-covariance matrix. For this reason, it is also called **Covariance Based (CB) SEM**[1].
- Since it is used to estimate multiple equations simultaneously, models are described and reported in the form of block diagrams.
- As the Stata manual also noted, under the umbrella of SEM fall many different models...

[1] Also to distinguish it from Partial Least Squares (PLS) SEM.

Some examples

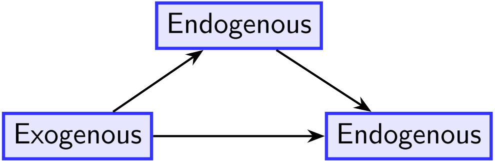
Path Analysis (PA)



Relationships between manifest variables are estimated.

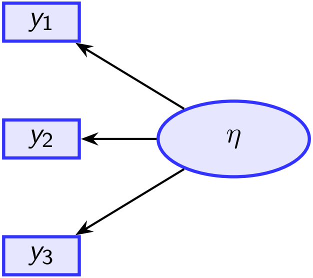
Some examples

Path Analysis (PA)



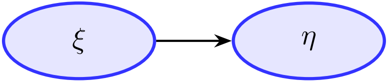
In these models it is complicated to speak of dependent and independent variables. Variables that are not explained within the model are called exogenous, while those that are explained are called endogenous.

Confirmatory Factor Analysis (CFA)



Relationships between manifest and latent variables are estimated.

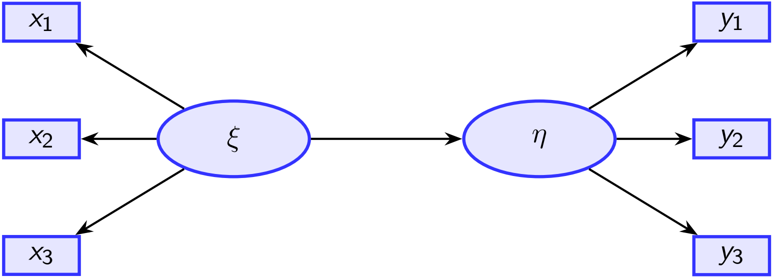
Latent Path Analysis (LPA)



Relationships between latent variables are estimated.

CFA + LPA

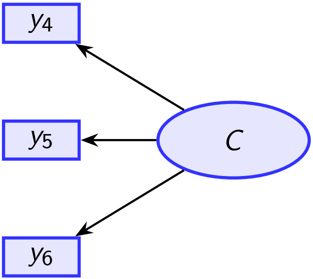
What is commonly meant by SEM



In these cases the CFA part is referred to as the **measurement model** *or* **outer model**. *The LPA (or PA) part, on the other hand, is referred to as the* **structural model** *or* **inner model**.

Some examples

Are latent variables only continuous?



No. We can hypothesise the existence of unobservable (or unobserved) classes and estimate them from observed indicators. SEM with categorical latent variables include Latent Class Analysis, Latent Profile Analysis, and other finite mixture models.

Moving on to...

A Bit of Terminology

A Bit of History

A Bit of Theory

Brief history of SEM

Origins

SEM analysis originates from the integration of:

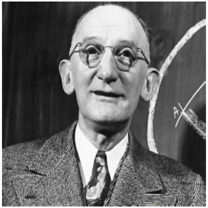
- **Path Analysis** (Sewall Wright, 1921)
- **Factor Analysis** (Spearman, 1904; Thurstone, 1930)
- **Latent variable models** (Jöreskog, 1969)

SEM unifies causal and measurement approaches within a single statistical framework.

Sewall Wright

Charles Spearman

Louis Thurstone



Karl Jöreskog

Brief history of SEM

CB-SEM: Covariance-Based SEM

- Formalised by Karl Jöreskog (1969) with LISREL;
- Objective: reproduce the observed **covariance matrix**;
- Based on **maximum likelihood**.

PLS-SEM: Partial Least Squares SEM

- Introduced by Herman Wold (1970s);
- **Component-based**, iterative approach;
- Focused on **maximising explained variance**;
- Preferable in case of: small samples, formative models, non-normality.

Moving on to...

A Bit of Terminology

A Bit of History

A Bit of Theory

Analysis process

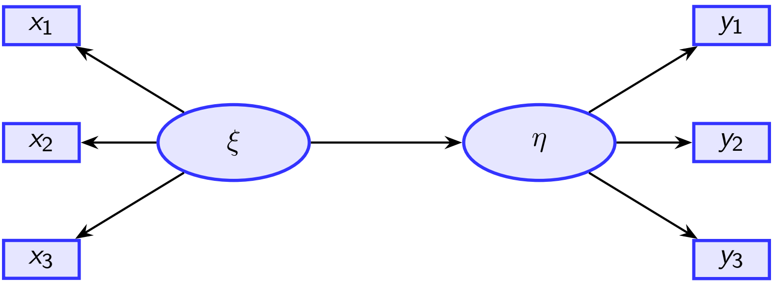
SEM analysis goes through four phases, the same for both **CB-SEM** and **PLS-SEM**. However, the specific criteria to be used differ, as these models differ in their functioning and purposes (e.g. Reflective or Formative LVs). We will focus on the four phases of the process with an emphasis on **CB-SEM**:

1. Model specification
2. Model identification
3. Model estimation
4. Model evaluation

Model specification

Phase 1:

The first phase consists of defining the theoretical model and the hypothesised relationships. In many software packages (including Stata) model specification can be done either by drawing diagrams or by writing code.




```
.  sem (Xi -> x1 x2 x3) (Eta -> y1 y2 y3) (Xi -> Eta)
```

Model identification

Phase 2:

Unfortunately, not everything that can be drawn can also be estimated. The **T-Rule**, for example, is a necessary (but not sufficient) condition for a model to be identifiable:

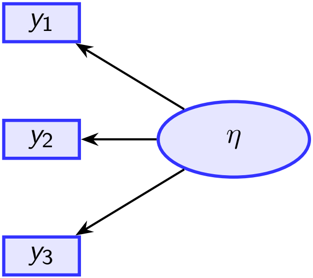
$$t \leq \frac{p(p+1)}{2}$$

- t is the number of parameters to be estimated (arrows in the diagrams);
- p is the number of manifest variables in the model.

- If $t < \frac{p(p+1)}{2}$ the model is *over-identified*: estimable and testable;
- If $t = \frac{p(p+1)}{2}$ the model is *just-identified*: estimable but not testable (degrees of freedom $= 0$);
- If $t > \frac{p(p+1)}{2}$ the model is *under-identified*: cannot be uniquely estimated.

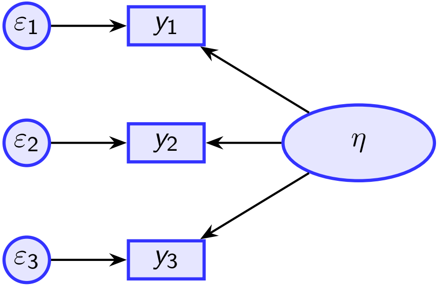
Model identification

$$t \leq \frac{p(p+1)}{2} \qquad \textbf{?}$$



Model identification

$$t \leq \frac{p(p+1)}{2} \qquad \textbf{?}$$



Model estimation

Phase 3:

Model estimation consists of calculating the parameter values that best reproduce the observed covariance matrix (typically through *maximum likelihood*).

Model evaluation

Phase 4:

Model evaluation takes place at three levels:

1. **Global goodness of fit**:
   - χ^2 test (preferably non-significant)
   - RMSEA < 0.06
   - CFI, TLI > 0.95
   - SRMR < 0.08
2. **Measurement model**:
   - λ_i (factor loadings) > 0.5 and significant
   - CR (Composite Reliability) > 0.7
   - AVE (Average Variance Extracted) > 0.5
   - Discriminant validity (e.g. Fornell-Larcker criterion)
3. **Structural model**:
   - β (path coeff.) significant
   - Acceptable R^2 for endogenous variables
   - Significant direct, indirect and total effects (if relevant)

Modification Indices

What are they?

Modification Indices (MI) signal potential improvements to the model by introducing unspecified parameters (e.g. residual covariances).

Caution: suggested modifications should be accepted only if theoretically justified, to avoid *overfitting*.

By introducing new parameters to be estimated, we go back to Phase 1!