Key ideas, terms & concepts in SEM

Professor Patrick Sturgis

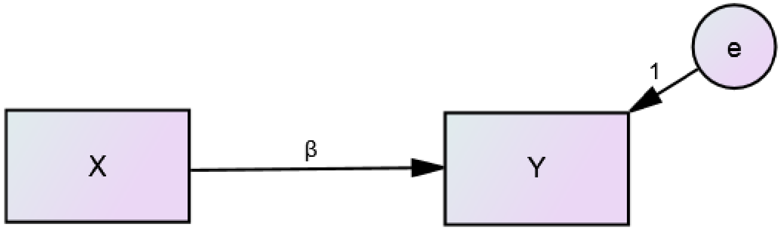
Plan

- Path diagrams
- Exogenous, endogenous variables
- Variance/covariance matrices
- Maximum likelihood estimation
- Parameter constraints
- Nested Models and Model fit
- Model identification

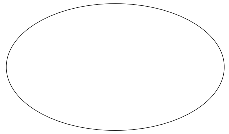
Path diagrams

- An appealing feature of SEM is representation of equations diagrammatically

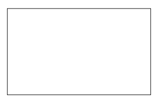
e.g. bivariate regression Y= bX + e



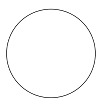
Path Diagram conventions



Measured latent variable

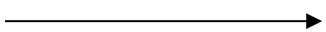


Observed / manifest variable



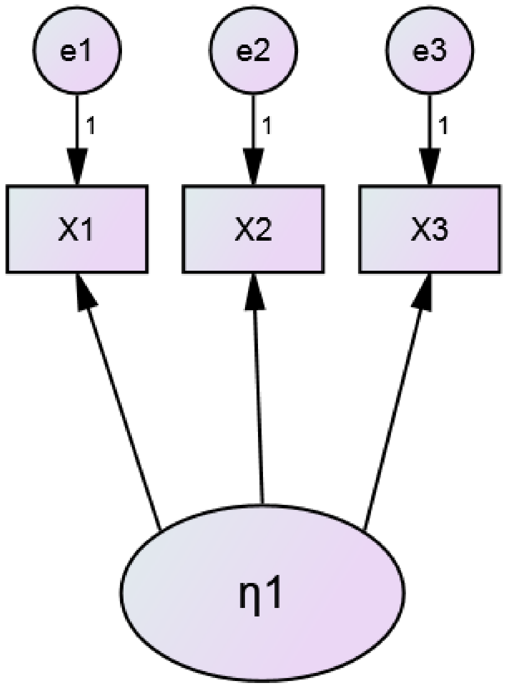
Error variance / disturbance term

Covariance / non-directional path



Regression / directional path

Reading path diagrams

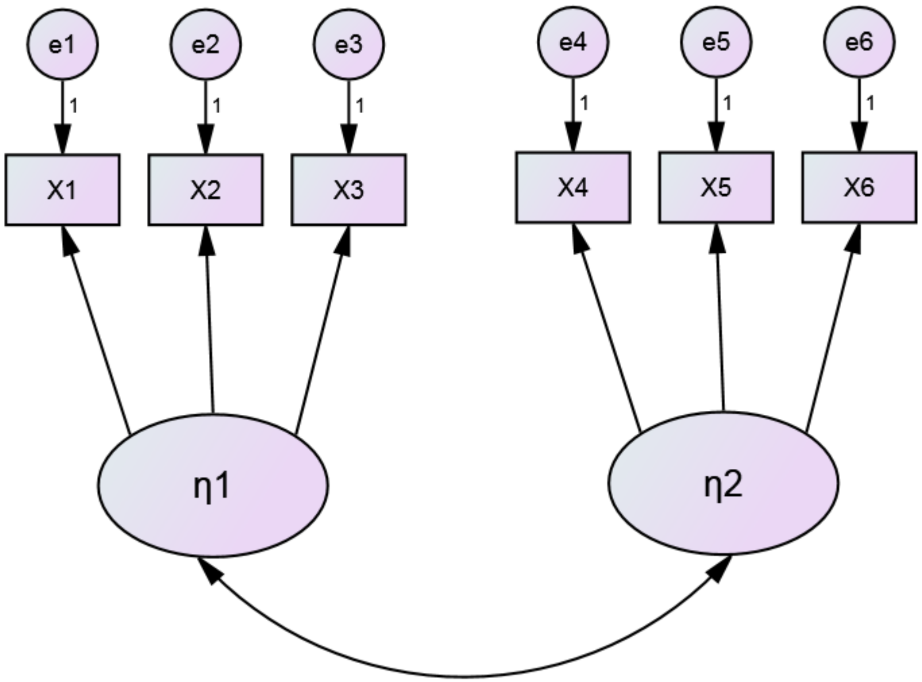


With 3 error variances

Causes/measured by
3 observed variables

A latent variable

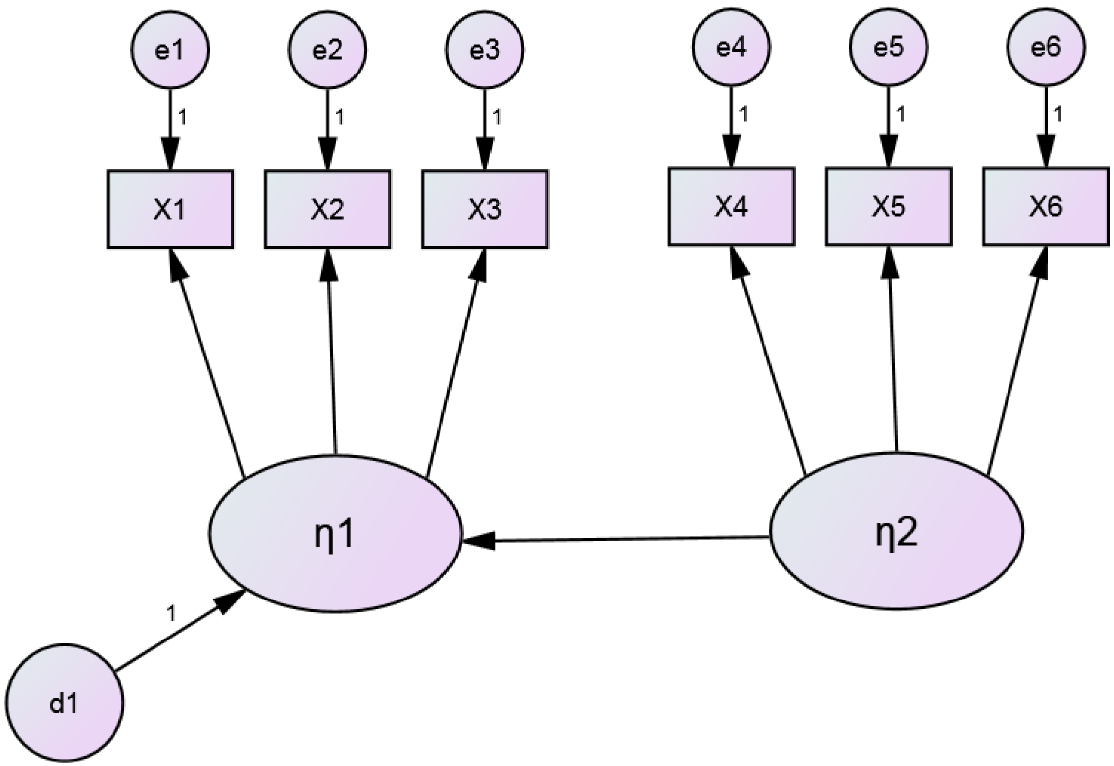
Reading path diagrams



Correlated

2 latent variables, each measured by 3 observed variables

Reading path diagrams



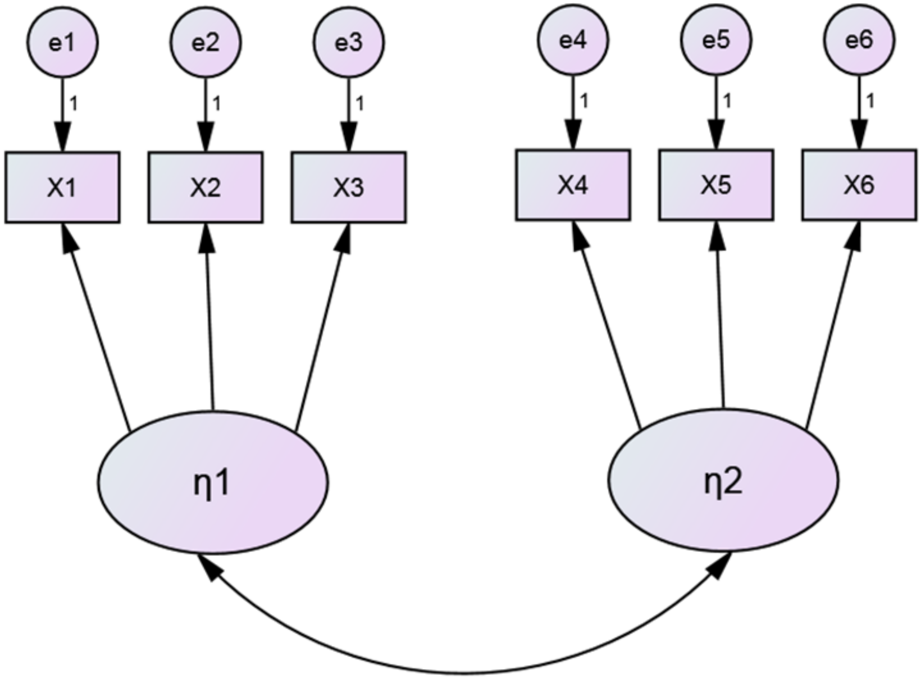
2 latent variables, each measured by 3 observed variables

Error/disturbance

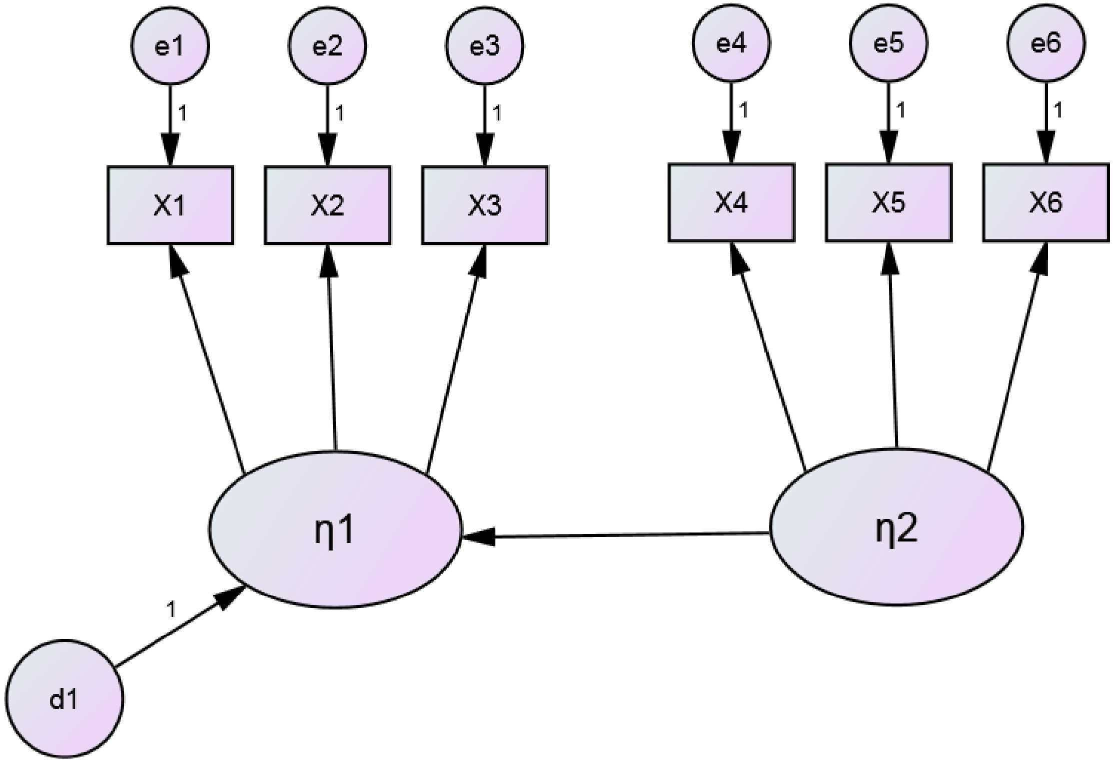
Exogenous/Endogenous variables

- Endogenous (dependent)
  - caused by variables in the system
- Exogenous (independent)
  - caused by variables outside the system
- In SEM a variable can be a predictor <u>and</u> an outcome (a mediating variable)

2 (correlated) exogenous variables



η1 endogenous, η2 exogenous



Data for SEM

- In SEM we analyse the variance/covariance matrix (S) of the observed variables, not raw data

- Some SEMs also analyse means

- The goal is to summarise S by specifying a simpler underlying structure: the SEM

- The SEM yields an implied var/covar matrix which can be compared to S

Variance/Covariance Matrix (*S*)

	x1	x2	x3	x4	x5	X6
x1	**0.91**	-0.37	0.05	0.04	0.34	0.31
x2	-0.37	**1.01**	0.11	0.03	-0.22	-0.23
x3	0.05	0.11	**0.84**	0.29	0.14	0.11
x4	0.04	0.03	0.29	**1.13**	0.11	0.06
x5	0.34	-0.22	0.14	0.11	**1.12**	0.34
x6	0.31	-0.23	0.11	0.06	0.34	**0.96**

Maximum Likelihood (ML)

- ML estimates model parameters by maximising the Likelihood, L, of sample data
- L is a mathematical function based on joint probability of continuous sample observations
- ML is asymptotically unbiased and efficient, assuming multivariate normal data
- The (log)likelihood of a model can be used to test fit against more/less restrictive baseline

Parameter constraints

- An important part of SEM is fixing or constraining model parameters
- We fix some model parameters to particular values, commonly 0, or 1
- We constrain other model parameters to be equal to other model parameters
- Parameter constraints are important for identification

Nested Models

- Two models, A & B, are said to be ‘nested’ when one is a subset of the other

(A = B + parameter restrictions)

e.g. Model B:

$y_i = a + b_1X_1 + b_2X_2 + e_i$

- Model A:

$y_i = a + b_1X_1 + b_2X_2 + e_i$ (constraint: $b_1 = b_2$)

- Model C (not nested in B):

$y_i = a + b_1X_1 + b_2Z_2 + e_i$

Model Fit

- Based on (log)likelihood of model(s)
- Where model A is nested in model B:

  LLA-LLB = χ^2 , with df = dfA-dfB

- Where p of > 0.05, we prefer the more parsimonious model, A χ^2
- Where B = observed matrix, there is no difference between observed and implied
- Model ‘fits’!

Model Identification

- An equation needs enough ‘known’ pieces of information to produce unique estimates of ‘unknown’ parameters

  X + 2Y=7 (unidentified)

  3 + 2Y=7 (identified) (y=2)

- In SEM ‘knowns’ are the variances/ covariances/ means of observed variables
- Unknowns are the model parameters to be estimated

Identification Status

- Models can be:
  - Unidentified, knowns < unknowns
  - Just identified, knowns = unknowns
  - Over-identified, knowns > unknowns
- In general, for CFA/SEM we require over-identified models
- Over-identified SEMs yield a likelihood value which can be used to assess model fit

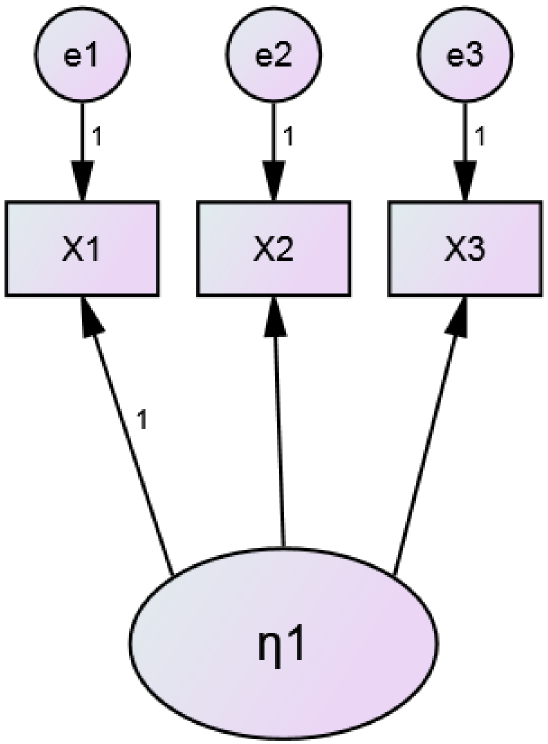
Assessing identification status

- Checking identification status using the counting rule
- Let s = number of observed variables in the model
- number of non-redundant parameters = $\frac{1}{2}s(s+1)$
- t=number of parameters to be estimated

t> $\frac{1}{2}s(s+1)$ model is unidentified

t< $\frac{1}{2}s(s+1)$ model is over-identified

Example 1 - identification



Non-redundant parameters

$$\frac{1}{2}s(s+1) = 6$$

parameters to be estimated

3 * error variance +

2 * factor loading +

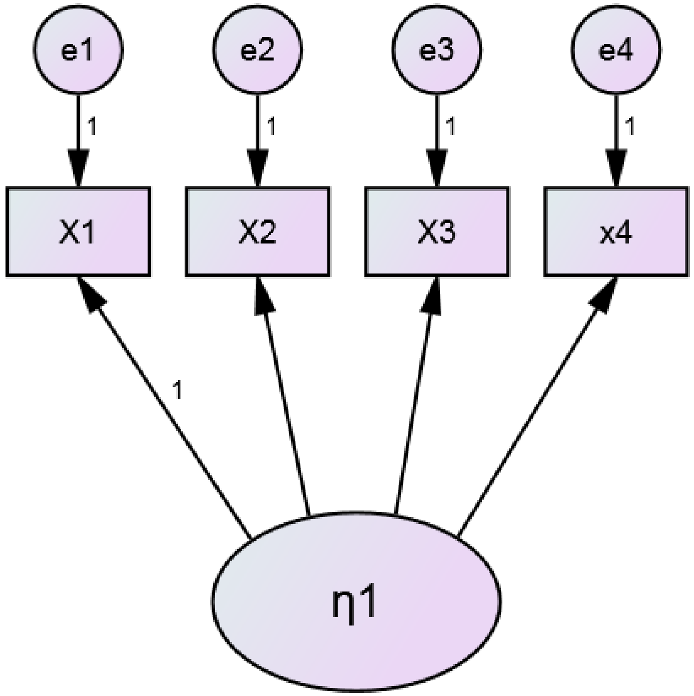
1 * latent variance = 6

6 - 6 = 0 degrees of freedom, model is **just-identified**

Controlling Identification

- We can make an under/just identified model over-identified by:
  - Adding more knowns
  - Removing unknowns
- Including more observed variables can add more knowns
- Parameter constraints remove unknowns
- Constraint $b_1=b_2$ removes one unknown from the model (gain 1 df)

Example 2 – add knowns



Non-redundant parameters

$$\frac{1}{2} s(s+1) = 10$$

parameters to be estimated

4 * error variance +

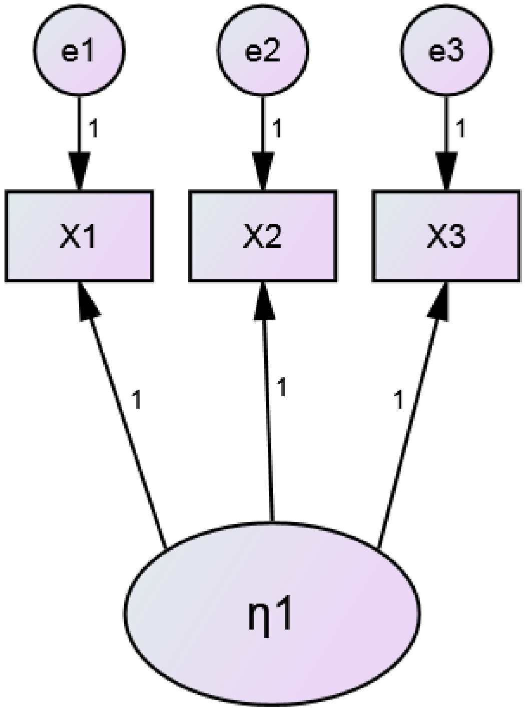
3 * factor loading +

1 * latent variance = 8

10 - 8 = 2 degrees of freedom, model is **over-identified**

Example 3 – remove unknowns

Constrain factor loadings = 1



Non-redundant parameters

$$\frac{1}{2}s(s+1) = 6$$

parameters to be estimated

3 * error variance +

0 * factor loading +

1 * latent variance = 4

6 - 4 = 2 degrees of freedom, model is **over-identified**

Summary

- SEM requires understanding of some ideas which are unfamiliar for many substantive researchers:
  - Path diagrams
  - Analysing variance/covariance matrix
  - ML estimation
  - global ‘test’ of model fit
  - Nested models
  - Identification
  - Parameter constraints/restrictions

NCRM National Centre for Research Methods

For more information contact

ncrm.ac.uk