STRUCTURAL EQUATION MODEL KEY CONCEPTS

Farid Najafi
MD, PhD
School of Public Health
Kermanshah University of Medical Sciences, Kermanshah, Iran
2020

SUBJECTS TO BE COVERED

Definition of SEM and its characteristics

Latent variable

SEM vs. path analysis

WHAT IS SEM

SEM is not one statistical technique

It integrates a number of different multivariate techniques into one model fitting framework

It is an integration of:

- Measurement theory
- Factor (latent variable) analysis
- Path analysis
- Regression
- Simultaneous equation

USEFUL FOR RESEARCH QUESTIONS THAT…

Involve complex, multi-faceted constructs that are measured with error

That specify 'systems' of relationship rather than a dependent variable and a set of predictors

- It suitable for analysis of a causal systems with many dependent and independent variables
- It is useful to show the mechanism of actions which is important for both (direct and indirect effect)
  1. Our knowledge regarding the biology of chain of causation
  2. To help policy makers to have a clear idea about the list of required actions for controlling the health outcomes

ALSO KNOWN AS

Covariance structure analysis (analysis of covariance and not just variable directly)

Analysis of Moment Structure (AMOS)

- The modern SEMs are not just analyzing covariance but means

Analysis of Linear Structural Relationship (LISREL)

Causal modeling (controversial name)

- Causal model is not related to just the type of analysis
- It is largely dependent to the design of study

SEM CAN BE THOUGHT OF AS PATH ANALYSIS USING LATENT VARIABLES

SEM HAS TWO OVERALL STEPS

SEM is path analysis with latent variables

This is a distinction between:

- Measurement of constructs
- Relationships between these constructs

First step: measure constructs

Second step: estimate how construct are related to one another

دانشگاه علوم پزشکی و خدمات بهداشتی،درمانی کرمانشاه
Kermanshah University Of Medical Sciences

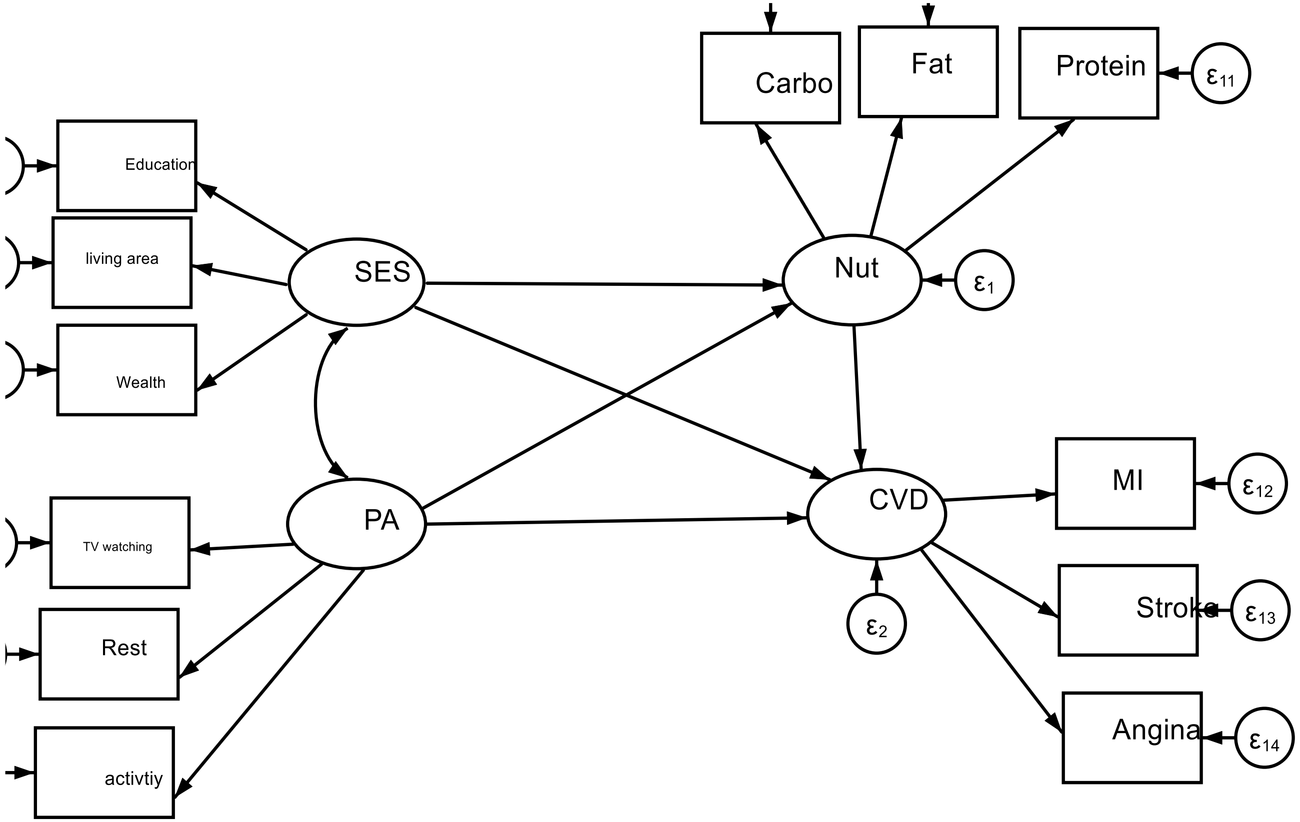
DETAILS OF STEPS IN SEM

1. Establish satisfactory measurement model for key concepts using latent variables
2. Fit regression path between concepts
   1. These are the 'structural equations' which specify how concept are related to each other
3. Test hypotheses on model parameters
4. Assess model fit

OTHER SPECIFICATIONS FOR STEPS IN SEM

1. Specification: What variables are in your data and how they are related to each other
2. Identification: Is there enough information for model estimation (mean, variance,…)
3. Estimation: Estimation of parameters
4. Evaluation: Is your model fitted?
5. Respecification
6. Interpretation

دانشگاه علوم پزشکی و خدمات بهداشتی،درمانی کرمانشاه
Kermanshah University Of Medical Sciences



WHAT ARE LATENT VARIABLES

Most social scientific concepts are not directly observable, e.g. intelligence, social capital, happiness

This makes them hypothetical or "latent" construct

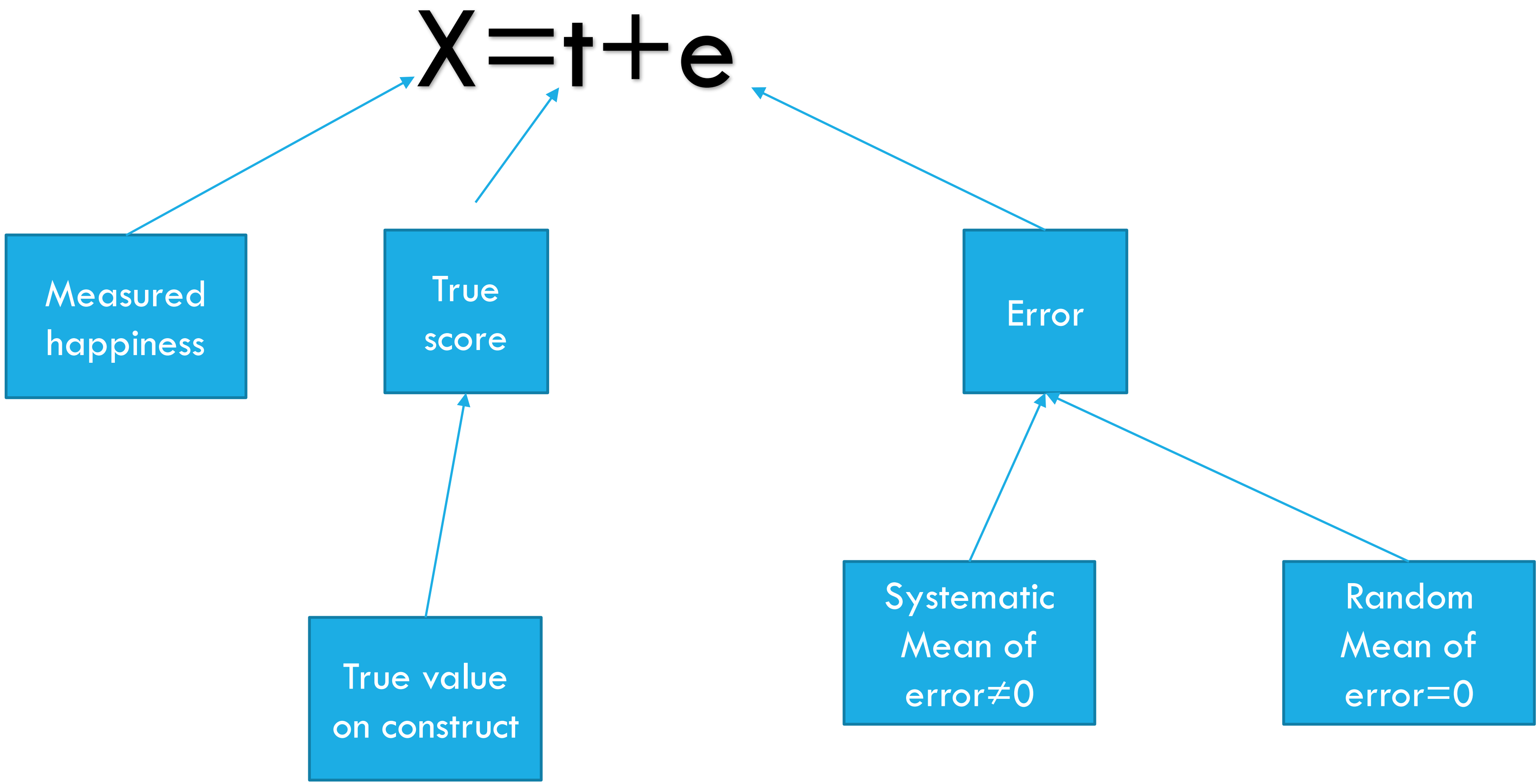
We can measure latent variables using observable indicators

We can think of the variance of a questionnaire item (e.g. happiness) as being caused by:

- The latent construct we want to measure (true variability in happiness among people)
- Other factors (error/unique variance) such as, temperature of room, type of questions and whether the questionnaire is self-administered or by interviewer

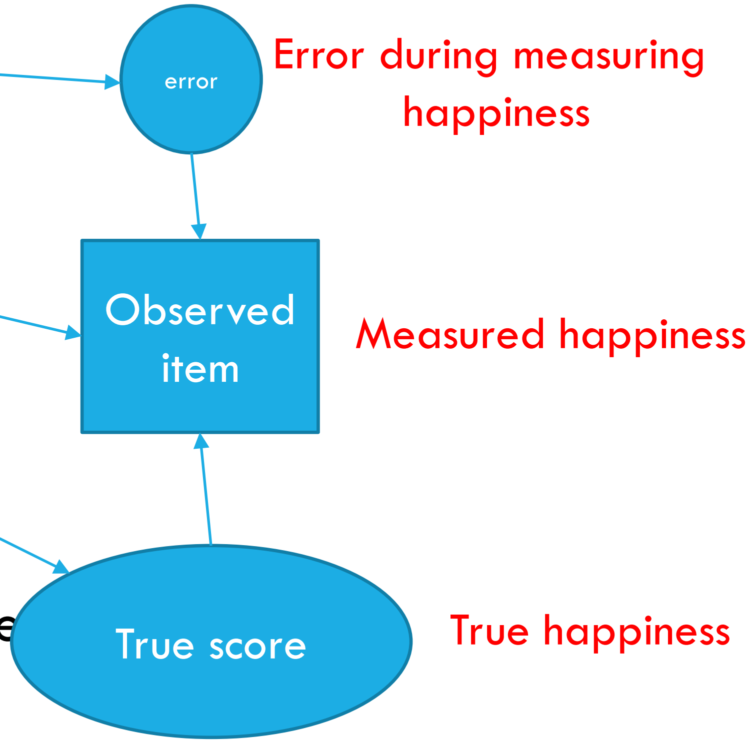
دانشگاه علوم پزشکی و خدمات بهداشتی،درمانی کرمانشاه
Kermanshah University Of Medical Sciences

TRUE SCORE AND MEASUREMENT ERROR



TRANSLATION OF TRUE SCORE AND ERROR

$$X=t+e$$



Problem – with one indicator, the equation is unidentified

We cannot separate true score and error

We have two unknowns (t and e) and one known (X) and therefore not possible to model it

MULTIPLE INDICATOR LATENT VARIABLE

To identify t & e components we need multiple indicators of the latent variable

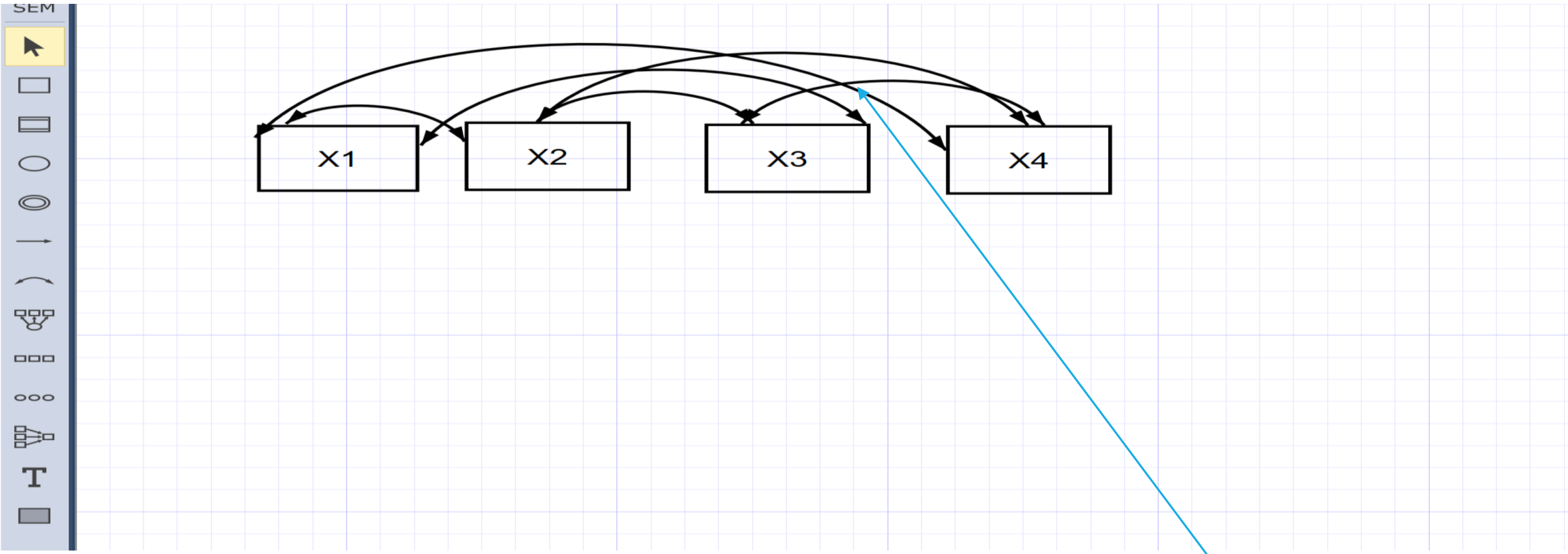
With multiple indicators we can use a latent variable model to partition variance

For doing this we can use factor analysis, principal analysis, latent class models

- e.g. principal component analysis transforms correlated variables into uncorrelated components

We can then use a reduced set of components (factors or indicators) to summarize the observed associations (by creating a latent variable)

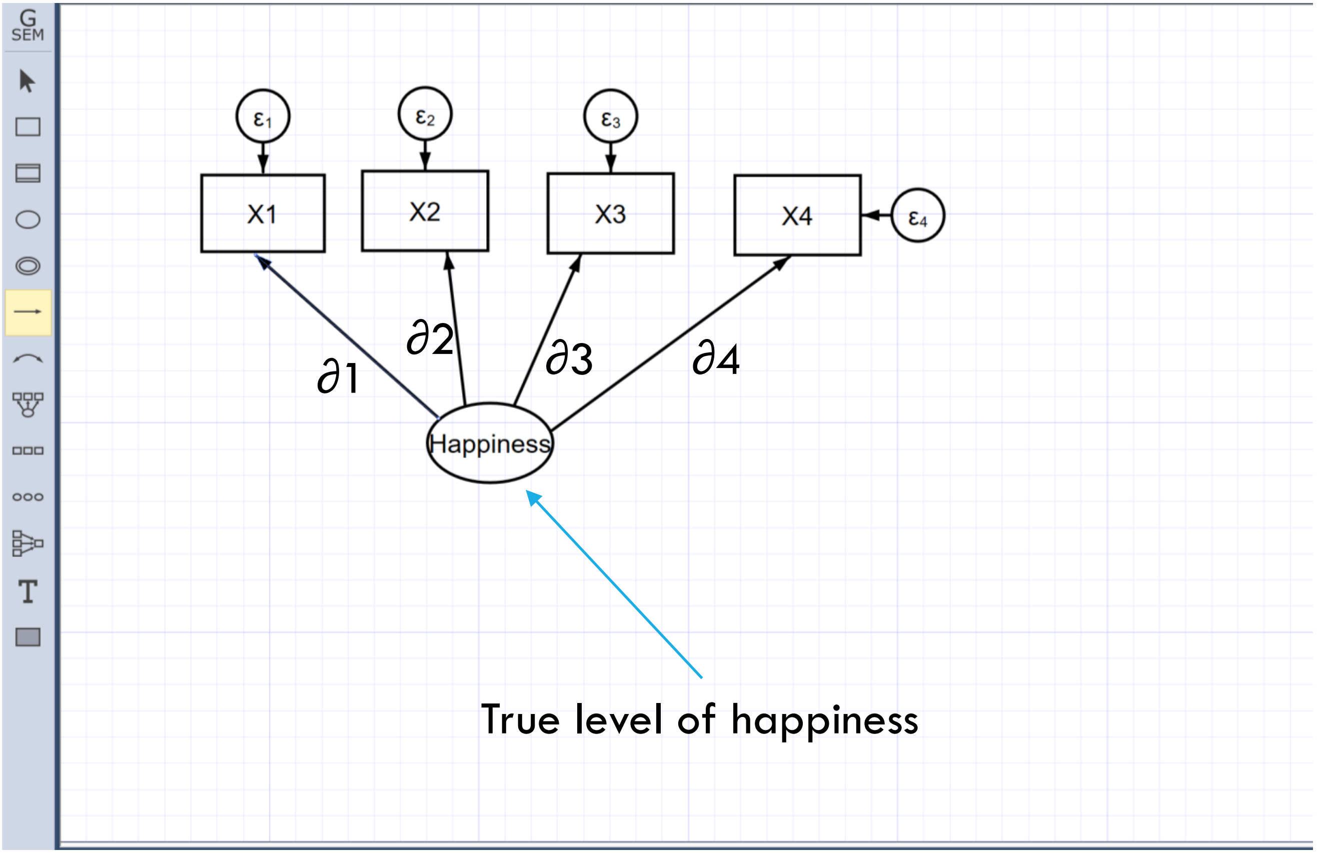
A COMMON FACTOR MODEL (ONE WAY OF PRESENTATION)



Four measures of a latent variable:
happiness in family, work, society and school
All measure a single latent variable “happiness”

Correlated measures

A COMMON FACTOR MODEL



∂ are factor loading=correlation between factor and indicators

BENEFIT OF LATENT VARIABLE

-Most social concepts are complex and multi-faceted

-Using single measures will not adequately cover the full conceptual map

-Remove/reduces random error in measured construct

Random error in dependent variable -> estimates unbiased but less precise

Random error in independent variables -> attenuates regression coefficient toward zero

REMEMBER

SEM CAN BE THOUGHT OF AS PATH ANALYSIS USING LATENT VARIABLES

We know about latent variables, what about path analysis

PATH ANALYSIS

The diagrammatic representation of a theoretical model using standardized notation

Regression equations specified between measured variables (and not latent variable)

“Effects” of predictor variables on criterion/dependent variables can be:

- Direct
- Indirect
- Total

PATH DIAGRAM NOTATION

Observed/manifest variable

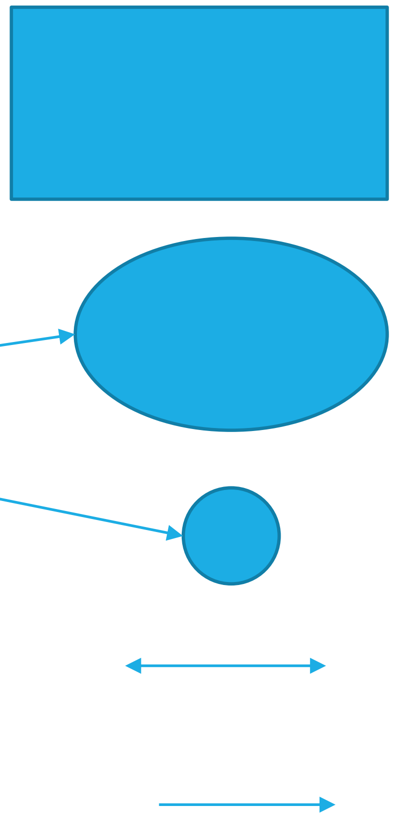
Measured latent variable

Similar to each other

Error variance/disturbance term

Covariance/non directional path

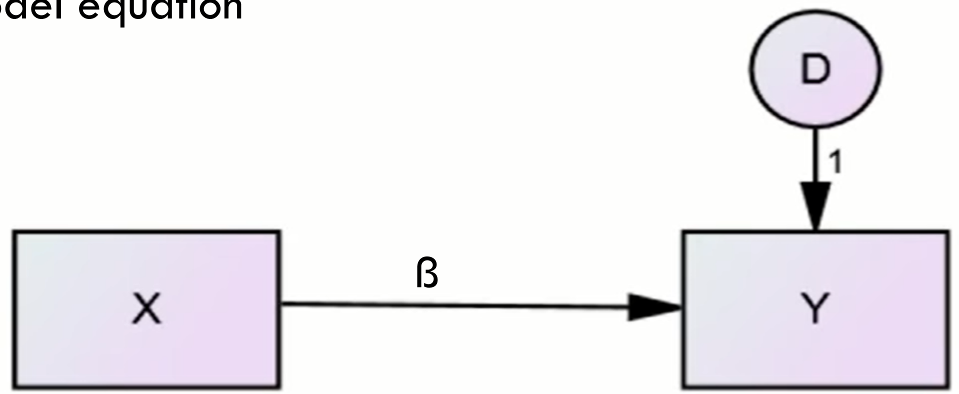
Regression/directional path



دانشگاه علوم پزشکی و خدمات بهداشتی،درمانی کرمانشاه
Kermanshah University Of Medical Sciences

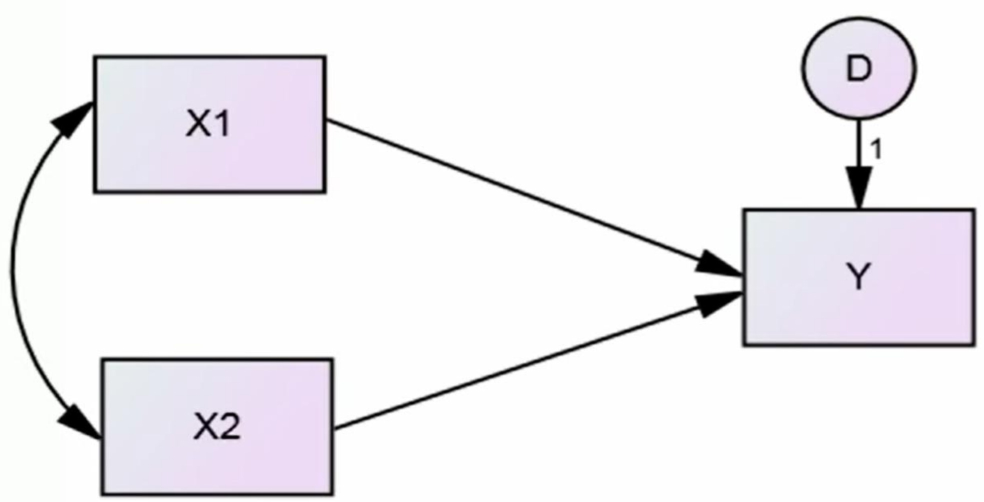
PDI
SINGLE CAUSE

Y=a+ßX+D simple linear regression model equation



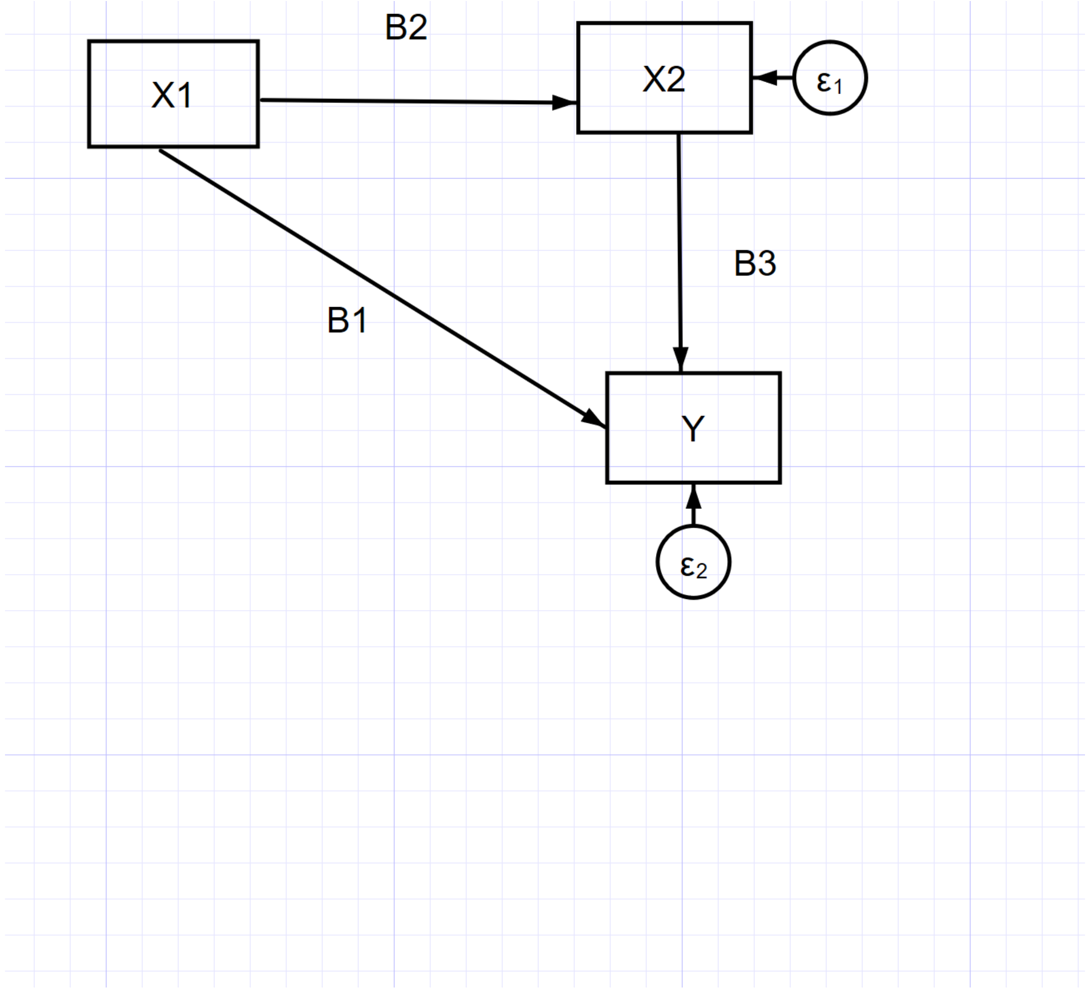
A bivariate regression model

TWO CORRELATED CAUSES



Multiple linear regression with two independent variables

INDIRECT EFFECT



B1=direct effect of X1 on Y

B2=direct effect of X1 on X2

B3=direct effect of X2 on Y

B2*B3=indirect effect of X1 on Y

B1+(B2*B3)=total effect of X1 on Y

SO THINK ABOUT SEM AS A PATH DIAGRAM INCLUDING LATENT VARIABLE

KEY IDEAS, TERMS AND CONCEPTS IN SEM

PLAN

Path diagram

Exogenous, endogenous variables

Variance/covariance matrices

Maximum likelihood estimation

Parameter constraints

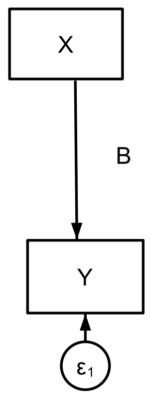
Nested models and model fit

Model identification

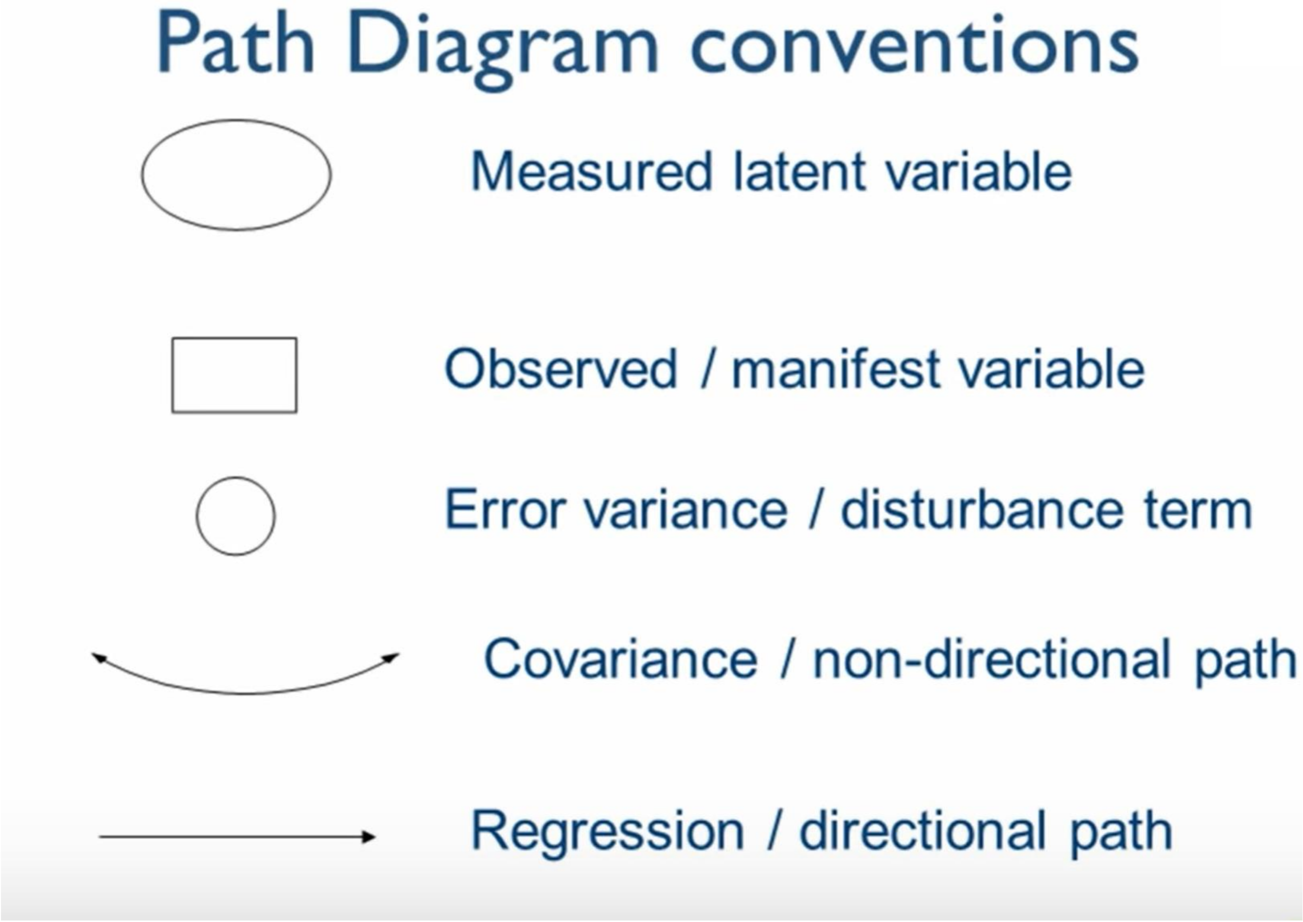
PATH DIAGRAMS

An appealing feature of SEM is representation of equations diagrammatically

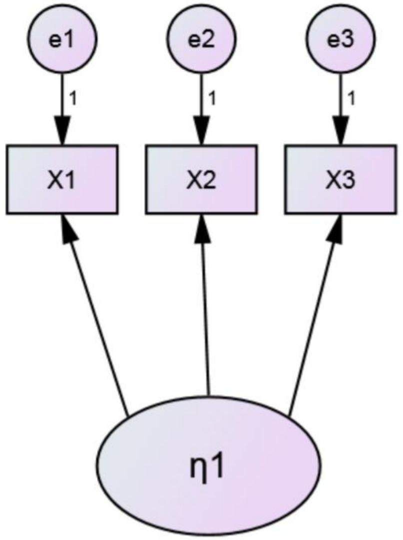
e.g. bivariate regression Y=bx+e



PATH DIAGRAM CONVENTIONS



Reading path diagrams

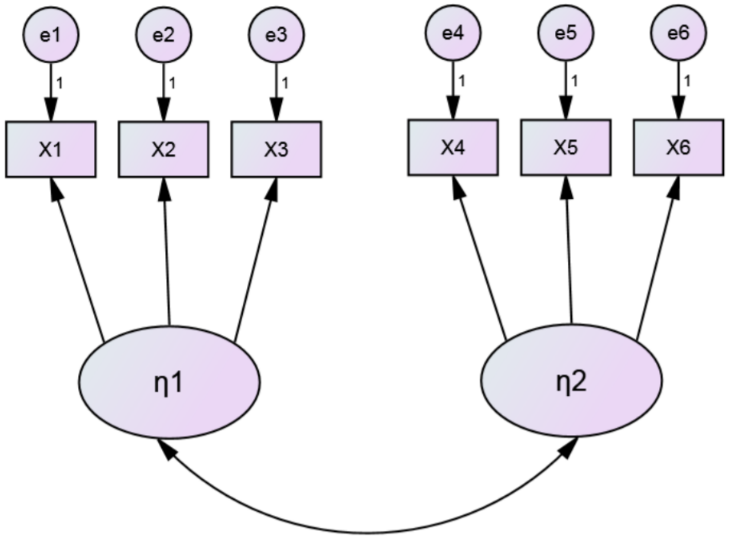


With 3 error variances

Causes/measured by 3 observed variables

A latent variable

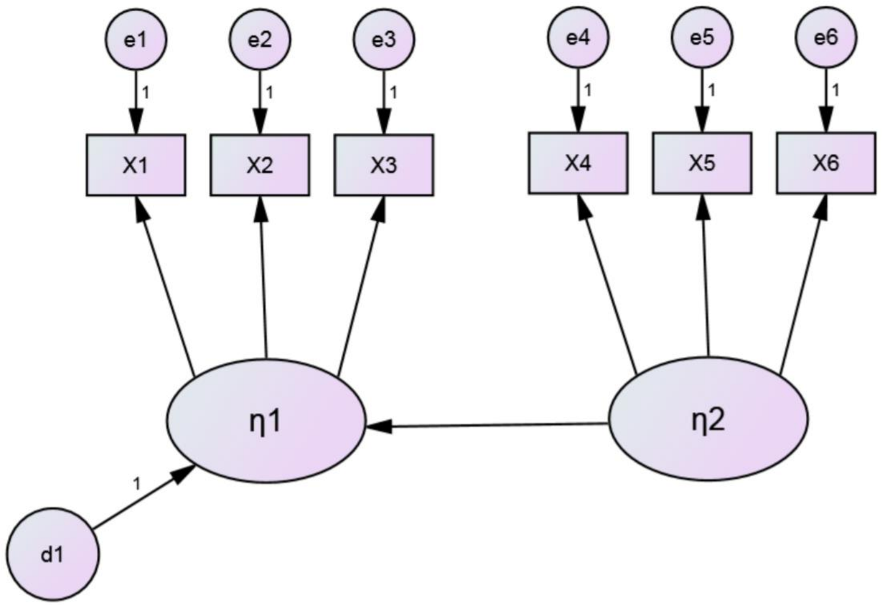
Reading path diagrams



Correlated

2 latent variables, each measured by 3 observed variables

Reading path diagrams



Error/disturbance

2 latent variables, each measured by 3 observed variables

EXOGENOUS/ENDOGENOUS VARIABLES

Endogenous (dependent)

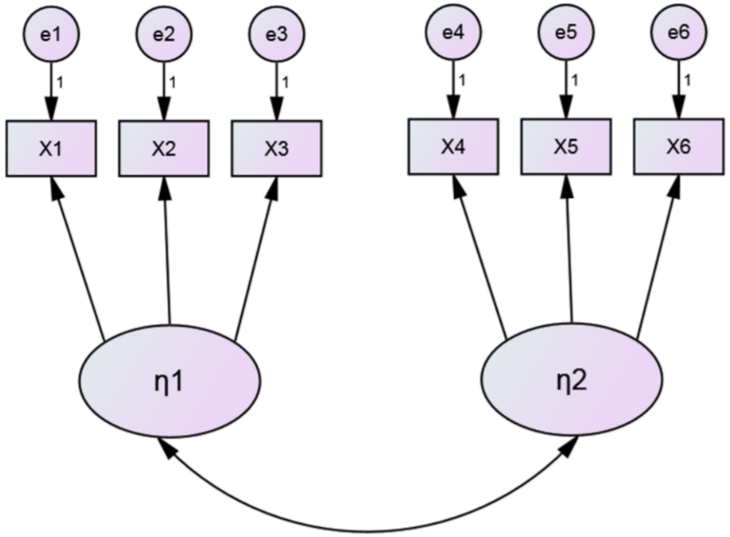
- Caused by variables in the system

Exogenous (independent)

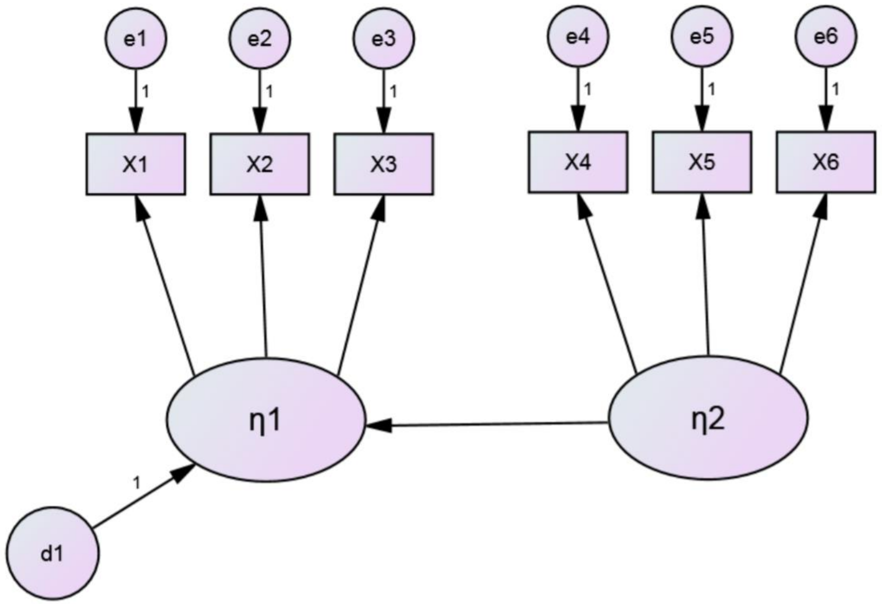
- Caused by variables outside the system

In SEM a variable can be a predictor and an outcome (a mediating variable)

TWO EXOGENOUS (LATENT) AND CORRELATED VARIABLES



η1 endogenous, η2 exogenous



DATA FOR SEM

In SEM we analyze the variance/covariance matrix (S) of the observed variables, not row data (surprising for people)

Some SEM also analyze means

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

The SEM yields an implied var/covar matrix in the suggested model which can be compared to the observed S

دانشگاه علوم پزشکی و خدمات بهداشتی،درمانی کرمانشاه
Kermanshah University Of Medical Sciences

$COV_{xy}=r_{xy}(SD_x)\ (SD_y)$

The variance (the covariance of a variable with itself)

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**

Correlation Matrix

MAXIMUM LIKELIHOOD (ML)

The main question in any relationship analysis is to estimate the parameter of association (ß)

In standard regression modeling we use ordinary (linear) least squares

ML estimate model parameter by maximizing the likelihood, L, of sample data

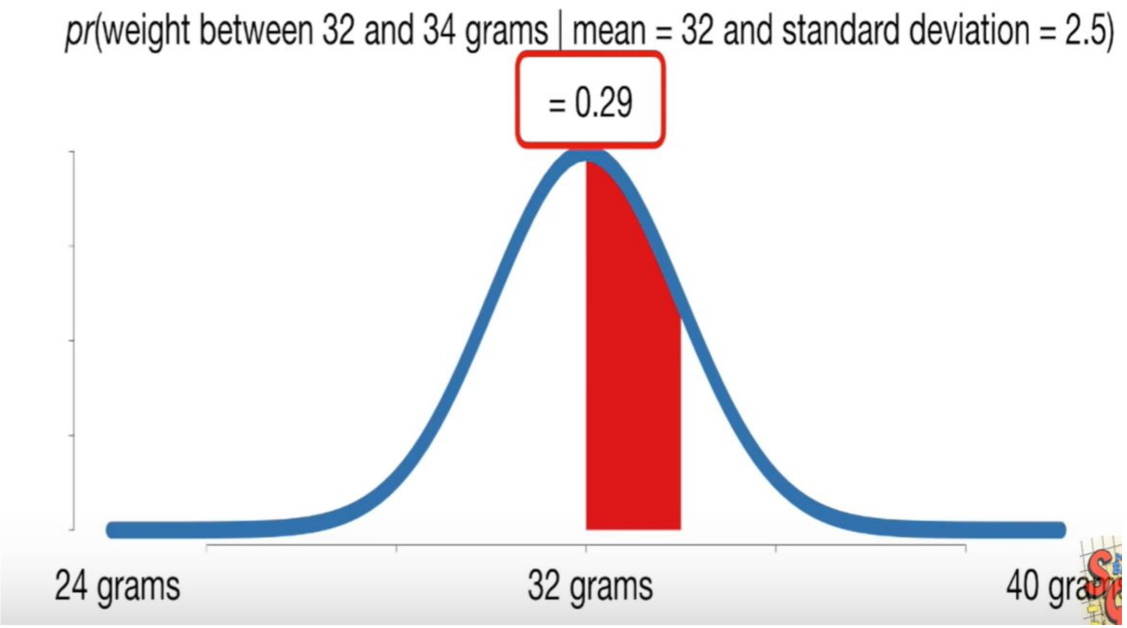
- What the maximum value of L is for particular sample of data (by iterating unknown parameters)

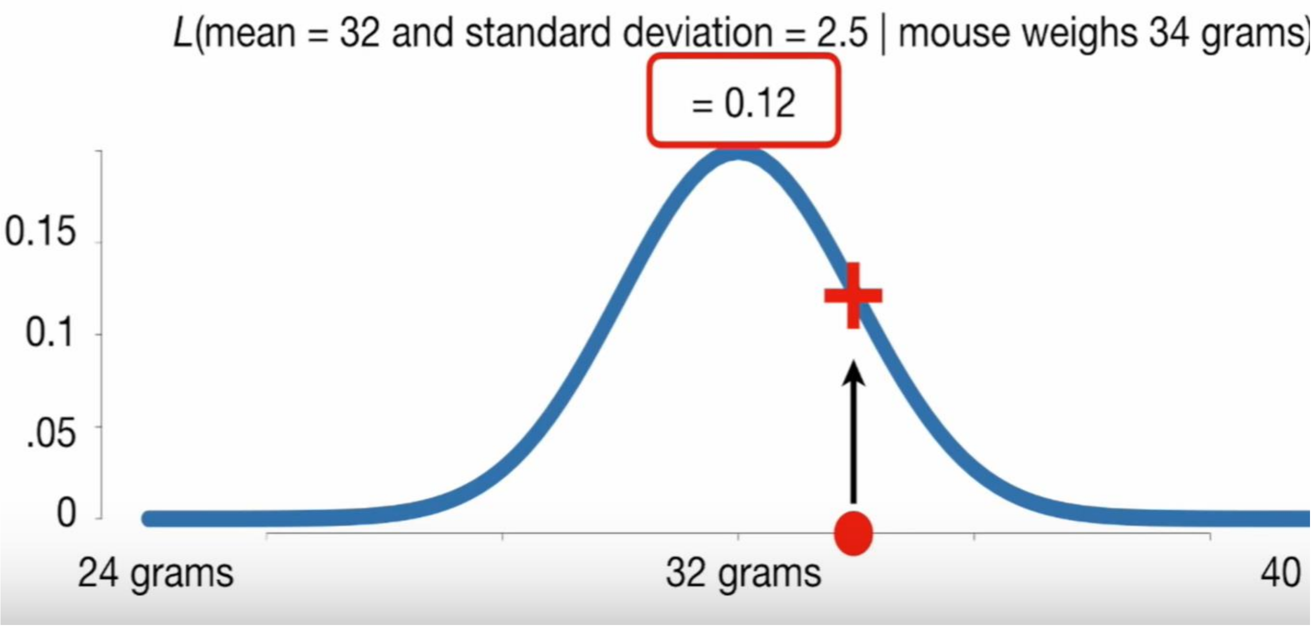
L is a mathematical function based on joint probability of continuous sample observations

Why ML?

1. ML is asymptotically unbiased and efficient, assuming multivariate normal data
   - If there is a large sample then our estimate would be correct (unbiased)
   - No other way give us such precise estimate (efficient)
   - We need to have all data in continuous format (multivariate normal data)
2. The (log)likelihood of a model can be used to test fit against more/less restrictive baseline

PROBABILITY VS. LIKELIHOOD



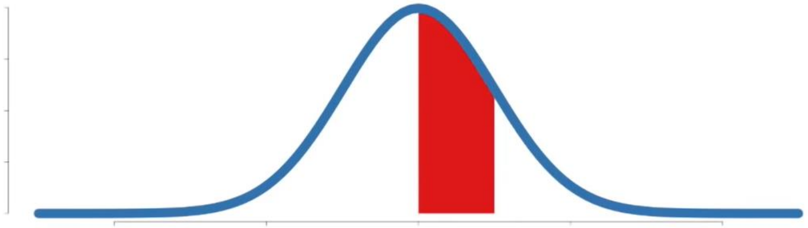


Probability is for the future events

In summary…

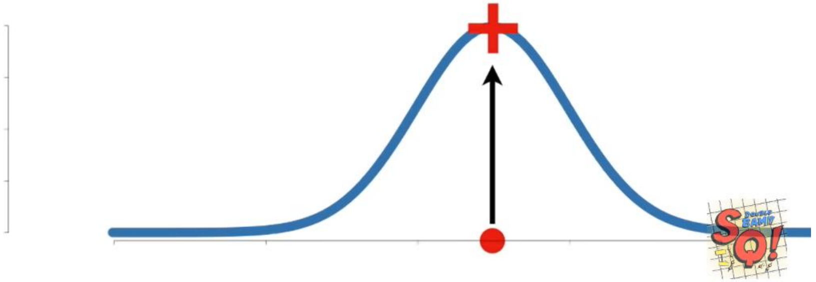
Probabilities are the areas under a fixed distribution…

pr(data | distribution)



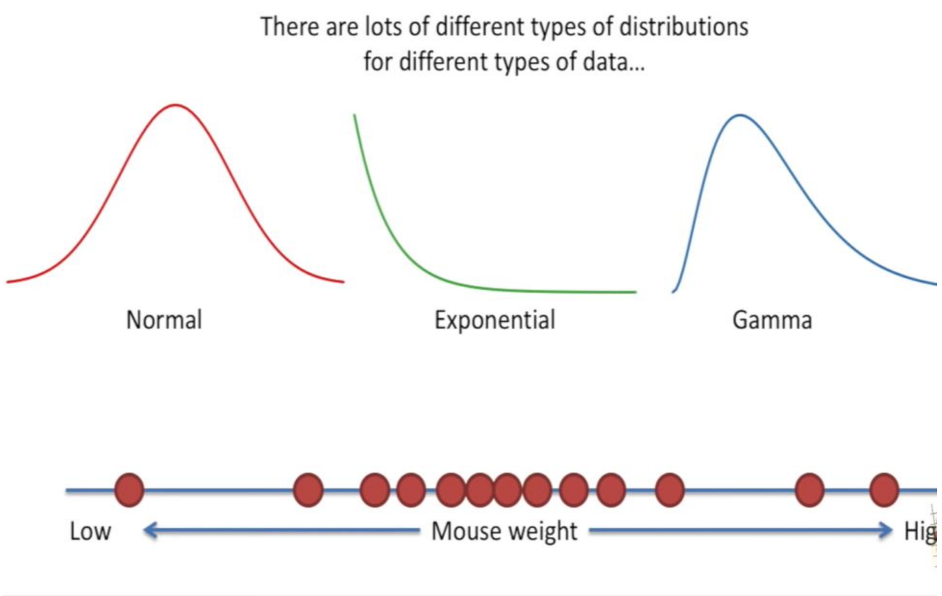
Likelihoods are the y-axis values for fixed data points with distributions that can be moved…

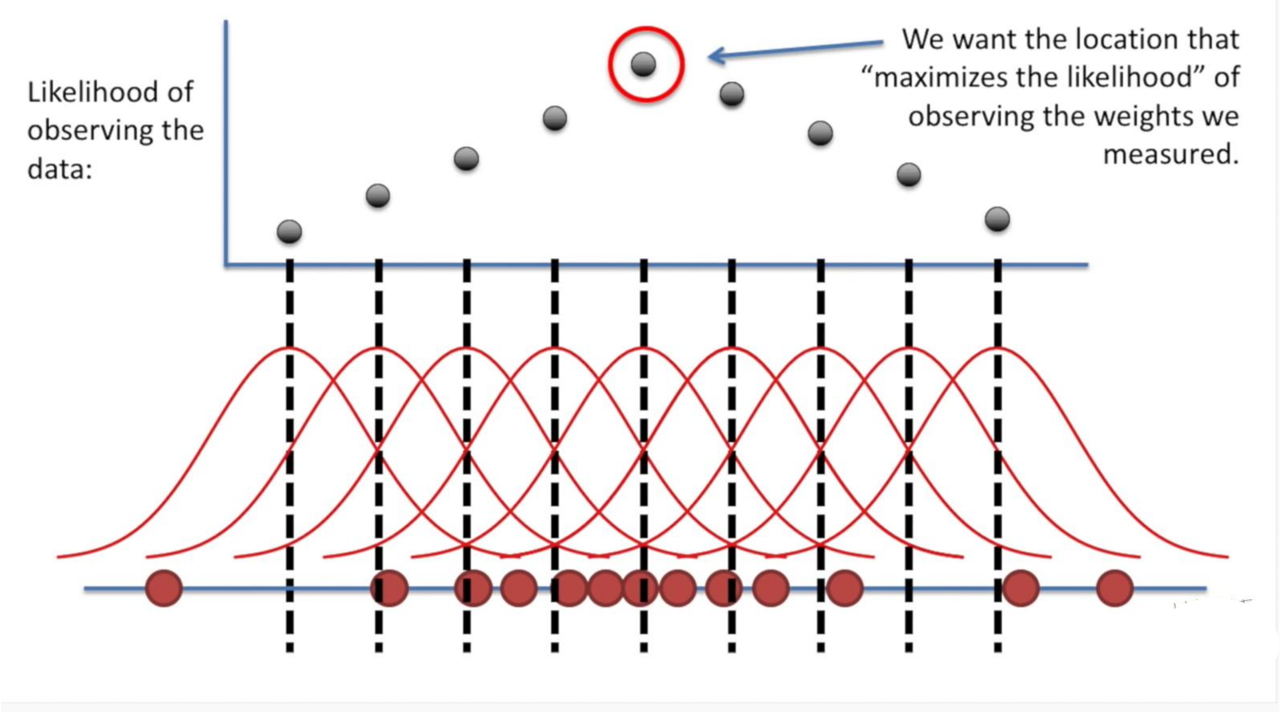
L(distribution | data)

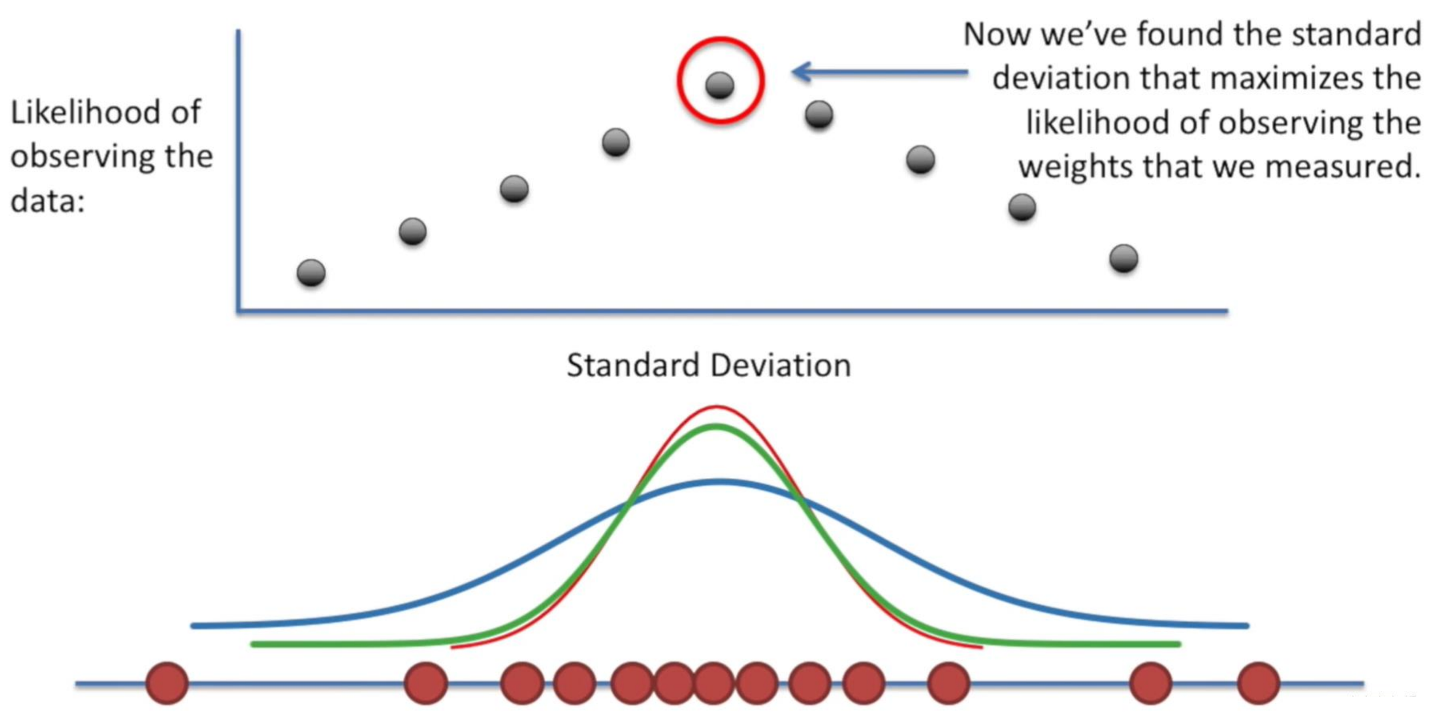


دانشگاه علوم پزشکی و خدمات بهداشتی،درمانی کرمانشاه
Kermanshah University Of Medical Sciences

The goal of maximum likelihood is to find the optimal way to fit a distribution to the data






Likelihood of observing the data:
Now we've found the standard deviation that maximizes the likelihood of observing the weights that we measured.
Standard Deviation

PARAMETER CONSTRAINTS

An important part of any association is estimating the unknown parameters

An important part of SEM is fixing or constraining model parameters which is unusual to some extent

We fix some model parameters to particular value, commonly 0, or 1

We constrain other model parameters to be equal to other model parameters

Parameter constraints are important for model identification

NESTED MODEL

Tow models, A and B, are said to be ‘nested’ when one is a subset of the other (or the parameter of one model is subset of the other)

(A=B+ Parameter restriction)

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=X^2 , with df=dfA-dfB

Where p of >0.05:

we prefer the more parasimonious model, A

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 variance/covariances/means of observed variables

Unknowns are the model parameters to be estimated

IDENTIFICATION STATUS

Model can be:

- Unidentified, knowns<unknowns
- Just identified, knowns=unknowns
  - We do not have any likelihood to use for assessing the model fit
- 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

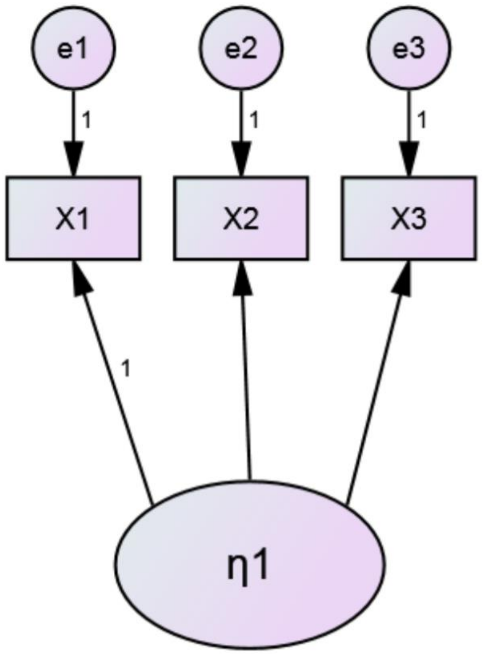
Number of non-redundant parameters=$1/2s(s+1)$

t=number of parameters to be estimated

$t>$ non-redundant parameters: model is unidentified

$t<$ non-redundant parameters: 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**

No degree of freedom and therefore although the parameters are estimated it is not possible to assess the fit of the model

CONTROLLING IDENTIFICATION

We can make an under/just identified model over-identified by:

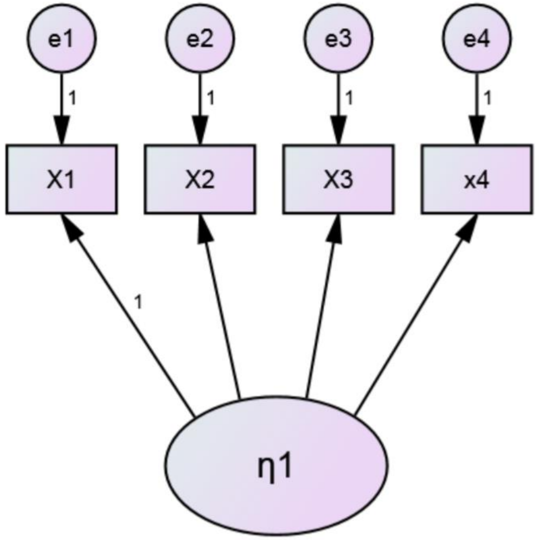
- Adding more knowns
- Removing unknowns

Including more observed variables can add more knowns

Parameter constraints remove unknowns

Constraint b1=b2 remove one unknown from the model (gain 1 df)

Example 2 – add knowns



Non-redundant parameters

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

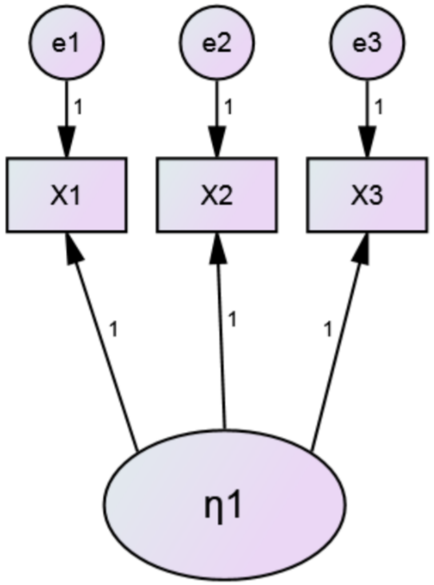
parameters to be estimated

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
- Analyzing variance/covariance matrix
- ML estimation
- Global 'test' of model fit
- Nested models
- Identification
- Parameter constraints/restriction

REFERENCE

Most of the materials for this presentation has been adapted from the following website:

https://www.ncrm.ac.uk/

با آرزوی موفقیت