

Structural Equation Modeling

- SEMs are linear equations for specifying phenomena in terms of cause-and-effect variables.
- In general form, SEM involves variables that are *unobservable*.
- SEM is a commonly used tool in the social and behavioral sciences.
- Applications: determination of profitability of a firm, efficacy of social programs, etc.

SEM is a linear model with a twist:

Linear Model

DATA → MODEL

What is the best linear model that can be fitted to the data?

Empirical model.

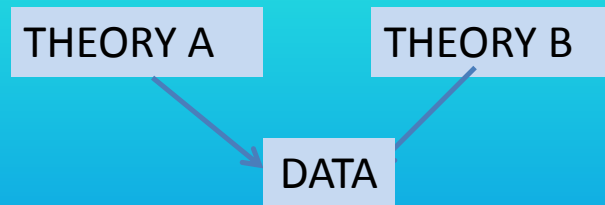
LISREL

MODEL → DATA

Could the hypothesized model lead to the data?

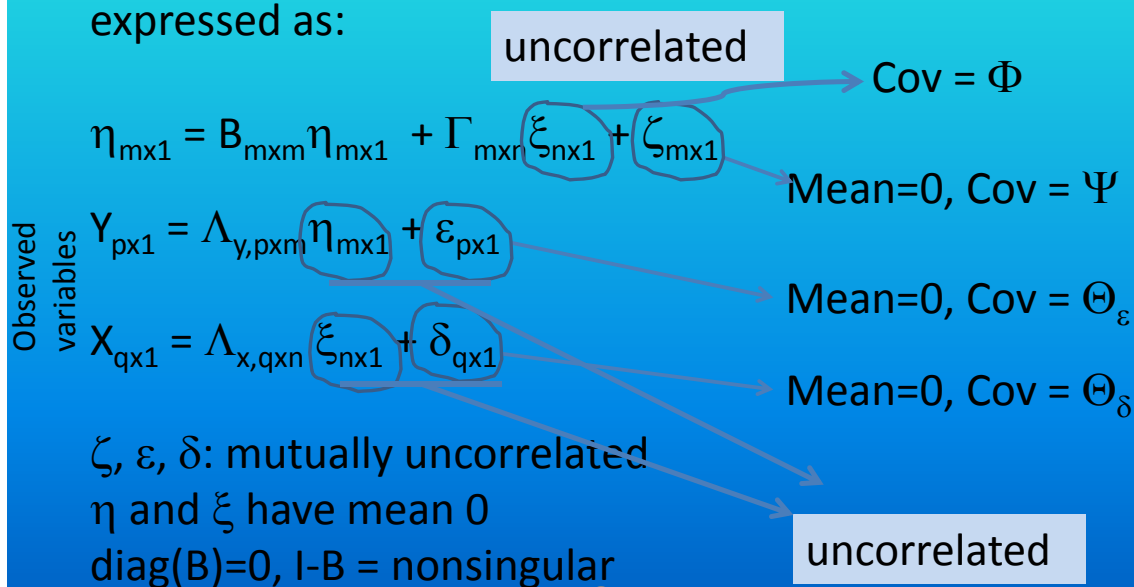
Theory expressed as a model.

SEM can be used to test alternative models:



K. G. Jöreskog and D. Sörbom (LISREL 8: User's Reference Guide (1996).

LISREL (Linear Structural Relationships) model is expressed as:

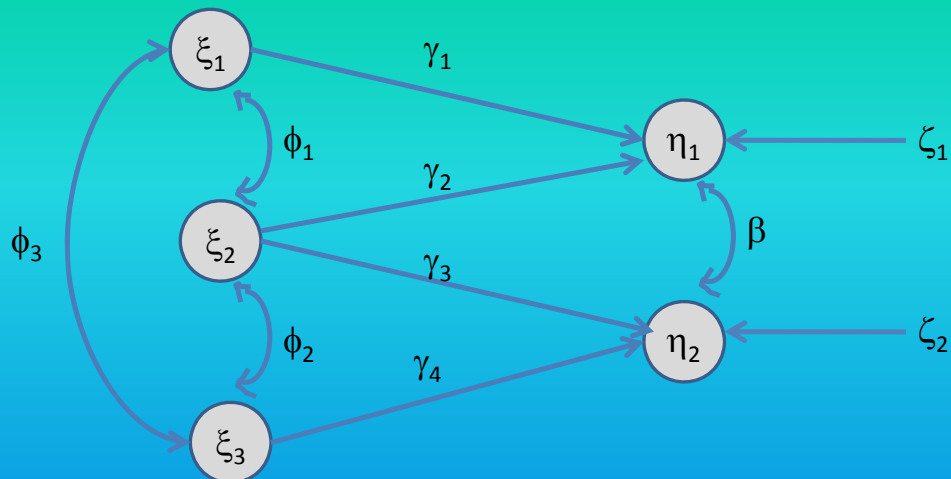


Path Diagram:

Exogenous variable – not influenced by other variables (IV)

Endogenous variable – affected by other variables (DV)

- Draw straight arrow to - each DV (endogenous var) from each of its source + each DV from its error term.
- Draw double-headed curved arrow between each pair of uncorrelated IV (exogenous) variables.



$$m = 2, n = 3$$

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & \beta \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \gamma_1 & \gamma_2 & 0 \\ 0 & \gamma_3 & \gamma_4 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix} + \begin{bmatrix} \zeta_1 \\ \zeta_2 \end{bmatrix}$$

- Since the (latent) variables η and ξ are unobservable, we cannot verify the LISREL model.
- LISREL model imposes a covariance structure which can be verified.
- At this point, we need to talk a little bit about the covariance matrix of a random vector (multivariate random variable), and also learn some covariance matrix algebra.

COVARIANCE MATRIX:

$X = (X_1, X_2, \dots, X_q)$ = q-variate random variable

$q=2$, $X = (X_1, X_2)$ is bivariate

Population mean of $X_j = \mu_j$ ($j=1,2$)

Population mean of $(X_j - \mu_j)^2 = (X_j - \mu_j) \times (X_j - \mu_j)$
 $= \text{var}(X_j) = \sigma_j^2$

Population mean of $(X_i - \mu_i) \times (X_j - \mu_j) = \text{Cov}(X_i, X_j)$

$X = (X_1, X_2)$

Population covariance matrix of X is:

$$\Sigma_{2 \times 2} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Example:

```
ab <- read.csv("K:/DataMining/Data/multinorm.csv",header=TRUE)
```

```
> mean(ab)
```

```
  X1    X2
```

```
9.952870 8.252316
```

```
> cov(ab)
```

```
  X1    X2
```

```
X1 3.504535 -2.885386
```

```
X2 -2.885386 8.337694
```

Note: The sample of size 50 was generated from bivariate normal with mean (10,8), and covariance

```
  4  -3  
 -3  9
```

In order to understand SEM, the following formula is needed:

If X_{qx1} has mean vector μ_x and covariance matrix Σ_x then CX has mean $C \mu_x$ and covariance matrix $C\Sigma_x C^T$ where C is a matrix with q columns.

Assuming $B = 0$ in LISREL equations:

$$\eta_{mx1} = B_{mxm} \eta_{mx1} + \Gamma_{m \times n} \xi_{nx1} + \zeta_{mx1}$$

$$\eta_{mx1} = \Gamma_{m \times n} \xi_{nx1} + \zeta_{mx1}$$

$$Y_{px1} = \Lambda_{y,pxm} \eta_{mx1} + \varepsilon_{px1}$$

$$X_{qx1} = \Lambda_{x,qxn} \xi_{nx1} + \delta_{qx1}$$

$$\begin{aligned} \text{Cov}(Y) &= E(\Lambda_{y,pxm} \eta_{mx1} + \varepsilon_{px1})(\Lambda_{y,pxm} \eta_{mx1} + \varepsilon_{px1})^T = \\ &= \Lambda_y E(\eta \eta^T) \Lambda_y^T + E(\varepsilon \varepsilon^T) = \Lambda_y \text{Cov}(\eta) \Lambda_y^T + \Theta_\varepsilon = \Sigma_{YY} \end{aligned}$$

Similarly

$$\text{Cov}(X) = \Lambda_x \text{Cov}(\xi) \Lambda_x^T + \Theta_\delta = \Sigma_{XX}$$

$$\begin{aligned} \text{Cov}(Y,X) &= E(YX^T) = E(\Lambda_y (\Gamma_{m \times n} \xi_{nx1} + \zeta_{mx1})(\Lambda_{x,qxn} \xi_{nx1} + \delta_{qx1})^T) \\ &= \Lambda_y \Gamma \Phi \Lambda_x = \Sigma_{XY} \end{aligned}$$

$$\Sigma = \text{Cov} \begin{pmatrix} Y \\ X \end{pmatrix} = \begin{bmatrix} \Sigma_{XX} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{YY} \end{bmatrix}$$

Given n multivariate observations on variables (X,Y), we can estimate the above population covariance matrix by the sample covariance matrix:

$$S = \text{Sample Cov of} \begin{pmatrix} Y \\ X \end{pmatrix} = \begin{bmatrix} S_{XX} & S_{YX} \\ S_{XY} & S_{YY} \end{bmatrix}$$

Set population cov matrix equal to the sample cov matrix and solve to estimate LISREL parameters.

$$\begin{bmatrix} \Sigma_{XX} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{YY} \end{bmatrix} = \begin{bmatrix} S_{XX} & S_{YX} \\ S_{XY} & S_{YY} \end{bmatrix}$$

Example 1: Klein's (1950) macroeconomic model of the U. S. economy.

C_t = Consumption (in year t)

I_t = Investment

W_t^p = Private wages

W_t^g = Government wages

X_t = Equilibrium demand

P_t = Private profits

K_t = Capital stock

G_t = Government non-wage spending

T_t = Indirect business taxes and net exports

A_t = Time trend, year - 1931

Year	C	P	Wp	I	K.lag	X	Wg	G	T	
1	1920	39.8	12.7	28.8	2.7	180.1	44.9	2.2	2.4	3.4
2	1921	41.9	12.4	25.5	-0.2	182.8	45.6	2.7	3.9	7.7
3	1922	45.0	16.9	29.3	1.9	182.6	50.1	2.9	3.2	3.9
4	1923	49.2	18.4	34.1	5.2	184.5	57.2	2.9	2.8	4.7
5	1924	50.6	19.4	33.9	3.0	189.7	57.1	3.1	3.5	3.8
6	1925	52.6	20.1	35.4	5.1	192.7	61.0	3.2	3.3	5.5
7	1926	55.1	19.6	37.4	5.6	197.8	64.0	3.3	3.3	7.0
8	1927	56.2	19.8	37.9	4.2	203.4	64.4	3.6	4.0	6.7
9	1928	57.3	21.1	39.2	3.0	207.6	64.5	3.7	4.2	4.2
10	1929	57.8	21.7	41.3	5.1	210.6	67.0	4.0	4.1	4.0
11	1930	55.0	15.6	37.9	1.0	215.7	61.2	4.2	5.2	7.7
12	1931	50.9	11.4	34.5	-3.4	216.7	53.4	4.8	5.9	7.5
13	1932	45.6	7.0	29.0	-6.2	213.3	44.3	5.3	4.9	8.3
14	1933	46.5	11.2	28.5	-5.1	207.1	45.1	5.6	3.7	5.4
15	1934	48.7	12.3	30.6	-3.0	202.0	49.7	6.0	4.0	6.8
16	1935	51.3	14.0	33.2	-1.3	199.0	54.4	6.1	4.4	7.2
17	1936	57.7	17.6	36.8	2.1	197.7	62.7	7.4	2.9	8.3
18	1937	58.7	17.3	41.0	2.0	199.8	65.0	6.7	4.3	6.7
19	1938	57.5	15.3	38.2	-1.9	201.8	60.9	7.7	5.3	7.4
20	1939	61.6	19.0	41.6	1.3	199.9	69.5	7.8	6.6	8.9
21	1940	65.0	21.1	45.0	3.3	201.2	75.7	8.0	7.4	9.6
22	1941	69.7	23.5	53.3	4.9	204.5	88.4	8.5	13.8	11.6

Endogenous
vars

SEM in R

$$C_t = \gamma_{10} + \gamma_{11}P_t + \gamma_{12}P_{t-1} + \beta_{11}(W_t^p + W_t^g) + \zeta_{1t} \quad (1)$$

$$I_t = \gamma_{20} + \gamma_{21}P_t + \gamma_{22}P_{t-1} + \beta_{21}K_{t-1} + \zeta_{2t}$$

$$W_t^p = \gamma_{30} + \gamma_{31}A_t + \beta_{31}X_t + \beta_{32}X_{t-1} + \zeta_{3t}$$

Structural
disturbances

+ 3 identities

$$X_t = C_t + I_t + G_t$$

$$P_t = X_t - T_t - W_t^p$$

$$K_t = K_{t-1} + I_t$$

In Klein's model, endogenous variables appear on the RHS of the equation (e.g., P_t in eqn (1)) as well, and therefore OLS may not yield consistent estimates.

The *instrumental-variable* estimation is used for estimating structural equation parameters. An *instrumental* variable is uncorrelated with the error term.

STATS24x7.com © 2010 ADI-NV, INC

15

Identification and Instrumental-Variables Estimation

Structural equation of the model can be expressed as

$$V_{nx1} = W_{nxp} \delta_{px1} + \zeta_{nx1} ,$$

W = model matrix consisting of p endogenous and exogenous variables, including a column of 1's for the constant term, and δ consists of γ and β .

Let Z_{nxp} = matrix of instrumental variables (including a column of 1's)

STATS24x7.com © 2010 ADI-NV, INC

16

$$Z^T V_{nx1} = Z^T W_{nxp} \delta_{px1} + Z^T \zeta_{nx1}$$

$Z^T \zeta_{nx1} / n \rightarrow 0$ since Z and ζ_{nx1} are uncorrelated.

$$Z^T V_{nx1} = Z^T W_{nxp} \delta_{px1}$$

$$\delta_{px1} = (Z^T W_{nxp})^{-1} Z^T V_{nx1}$$

assuming (a) # (instrumental vars) = # (predictors) = p
(b) $Z^T W_{nxp}$ is nonsingular.

The estimating equations $Z^T V_{nx1} = Z^T W_{nxp} \delta_{px1}$ are:

Under-identified if # (instrumental variables) < # (predictors)

Just-identified if # (instrumental variables) = # (predictors)

Over-identified if # (instrumental variables) > # (predictors)

Moreover, for $Z^T W_{nxp}$ to be non-singular, the instrumental variables must be correlated with the predictors, but not perfectly.

Two-stage Least Squares (2SLS):

1. Regress W on the instrumental variables Z .
2. Regress response V on the fitted values of W from Step 1 above.

The function `tsls` in `sem` library performs the 2SLS estimation.

See `sem_Klein.txt`

```
summary(Klein.eqn1)
2SLS Estimates
```

Identity function I
used to protect
expression $W_p + W_g$

```
Model Formula: C ~ P + P.lag + I(Wp + Wg)
```

```
Instruments: ~G + T + Wg + I(Year - 1931) + K.lag + P.lag + X.lag
```

```
Residuals:
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.89e+00 -6.16e-01 -2.46e-01 -1.07e-10 8.85e-01 2.00e+00
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.55476   1.46798 11.2772 2.587e-09
P            0.01730   0.13120  0.1319 8.966e-01
P.lag       0.21623   0.11922  1.8137 8.741e-02
I(Wp + Wg)  0.81018   0.04474 18.1107 1.505e-12
```

```
Residual standard error: 1.1357 on 17 degrees of freedom
```

summary(Klein.eqn2)

2SLS Estimates

Model Formula: $I \sim P + P.lag + K.lag$

Instruments: $\sim G + T + Wg + I(\text{Year} - 1931) + K.lag + P.lag + X.lag$

Residuals:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-3.29e+00	-8.07e-01	1.42e-01	1.06e-11	8.60e-01	1.80e+00

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	20.2782	8.38325	2.4189	0.027071
P	0.1502	0.19253	0.7802	0.445980
P.lag	0.6159	0.18093	3.4044	0.003375
K.lag	-0.1578	0.04015	-3.9298	0.001080

Residual standard error: 1.3071 on 17 degrees of freedom

summary(Klein.eqn3)

2SLS Estimates

Model Formula: $Wp \sim X + X.lag + I(\text{Year} - 1931)$

Instruments: $\sim G + T + Wg + I(\text{Year} - 1931) + K.lag + P.lag + X.lag$

Residuals:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.29e+00	-4.73e-01	1.45e-02	-3.81e-11	4.49e-01	1.20e+00

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.5003	1.27569	1.176	2.558e-01
X	0.4389	0.03960	11.082	3.368e-09
X.lag	0.1467	0.04316	3.398	3.422e-03
I(Year - 1931)	0.1304	0.03239	4.026	8.764e-04

Residual standard error: 0.7672 on 17 degrees of freedom