Introduction to Structural Equation Models

Chuan Zhou, PhD.
Department of Pediatrics, University of Washington

BIOSTAT 578 April 21st, 2009

Outline

- Outline
- A Motivating Example
- Specification of SEMs
  - Components of SEMs
  - Path Diagrams
  - Recursive models
- Instrumental-Variables
- Estimation of SEMs
- Example with R sem Package

Motivating example

Klein’s macroeconomic model


- Consumption = f(Private profits, Private wages, Government wages)
- Investment = f(Private profits, Capital stock)
- Private wages = f(Time trend, Spending Demand)
Klein’s economic model can be expressed in the following set of regression models,

\[ C_t = \gamma_{10} + \gamma_{11} P_t + \gamma_{12} (W^p_t + W^g_t) + \zeta_{1t} \]
\[ I_t = \gamma_{20} + \gamma_{21} P_t + \gamma_{22} P_{t-1} + \gamma_{23} K_{t-1} + \zeta_{2t} \]
\[ W^p_t = \gamma_{30} + \gamma_{31} A_t + \beta_{31} X_t + \beta_{32} X_{t-1} + \zeta_{3t} \]
\[ X_t = \gamma_{40} + \gamma_{41} X_t + \zeta_{4t} \]
\[ P_t = X_t - T_t - W^p_t \]
\[ K_t = K_{t-1} + I_t \]

- **Consumption (in year t)**
- **Investment**
- **Private wages**
- **Equilibrium demand**
- **Private profits**
- **Capital stock**
- **Government non-wage spending**
- **Indirect business taxes and net exports**
- **Government wages**
- **Time trend, year-1831**

Structural-equation models (SEMs) are multiple-equation regression models in which the response variable in one regression equation can appear as an explanatory variable in another equation. Structural-equation models can include variables that are not measured directly, but rather indirectly through their effects (indicators) or, sometimes, through observed causes (manifest variables). Model structural-equation methods represent a confluence of work in many disciplines, including biostatistics, econometrics, psychometrics, etc.

Some cautionary notes

- SEMs are multiple-equation regression models representing putative causal (and hence structural) relationships among a number of variables, some of which may affect one another mutually.
- Design is rarely explicitly taken into account, mostly on observational data.
- Lack of sound conceptual framework for causal effects.
- Claiming that a relationship is causal based on observational data is intrinsically problematic and requires support beyond the data at hand.
Two classes of variables

- **Endogenous variables** are the response variables of the model
  - In path diagram, they are the nodes with directional arrows going into
  - One structural equation per endogenous variable
  - An endogenous variable may also be an explanatory variable in other structural equations
- **Exogenous variables** appear only as explanatory variables in the SEMs
  - In path diagram, they are the nodes without arrows going into
  - The values of exogenous variables are therefore determined outside of the model
  - Assumed to be measured without error (unless latent)
  - Can be categorical while endogenous variables are mostly continuous

Structural errors

- Aggregated omitted causes of the endogenous variables plus measurement error (and possibly intrinsic randomness) in the endogenous variables
- One error variable per endogenous variable
- Assumed to have zero expectation and to be independent of exogenous variables
- Errors for different observations are assumed to be independent, but maybe correlated within observation
- Each error variable is assumed to have constant variance across observations, although the variances may differ across error variables
- Sometimes normality is assumed

Structural coefficients and covariance

- Structural coefficients represent the direct (partial) effect
  - On directed edge in path diagram
  - Of an exogenous on an endogenous variable
  - Of an endogenous on another endogenous variable
- Covariances can be either between two exogenous variables or two error variables (unanalyzed associations)

Path diagrams

Path diagram is a causal graph commonly used in SEMs. Some conventions are

- Nodes: observed variables in boxes, latent variables in circles
- Edges: a directed (single headed) arrow represent a direct effect of one variable on another; a bidirectional arrow represents a covariance (no causal interpretation given)
- Labels: unique subscripts on variables are helpful
A path diagram example

Duncan, Haller, and Portes’s (1968) study of peer influence on the aspiration of high school students.

Structural equations

- The structural equations of a model can be read straightforwardly from the path diagram.
  
  \[ y_5 = \gamma_{51} x_1 + \gamma_{52} x_2 + \beta_{56} y_6 + \epsilon_7 \]
  
  \[ y_6 = \gamma_{63} x_3 + \gamma_{64} x_4 + \beta_{65} y_5 + \epsilon_8 \]

- With some manipulation, including centering the exogenous variables at the means

  \[
  \begin{bmatrix}
  1 & -\beta_{56} & y_6 \\
  -\beta_{56} & 1 & \gamma_{52} & x_2 \\
  0 & 0 & \gamma_{63} & 0 \\
  0 & 0 & \gamma_{64} & 0 \\
  \end{bmatrix}
  \begin{bmatrix}
  y_1 \\
  x_1 \\
  x_2 \\
  \epsilon_7 \\
  \end{bmatrix}
  =
  \begin{bmatrix}
  \gamma_{51} \\
  \gamma_{52} \\
  y_5 \\
  x_6 \\
  \epsilon_8 \\
  \end{bmatrix}
  \]

Matrix form of the model

- More generally, when there are \( q \) endogenous variables, \( q \) errors, and \( m \) exogenous variables, the model for an individual observation is

  \[
  B_i y_i + \Gamma x_i = \epsilon_i 
  \]

- For all \( n \) observations,

  \[
  Y' B' + X' \Gamma' = E 
  \]
Recursive models

- An important type of SEM, called a recursive model, has two defining characteristics:
  1. Different error variables are independent
  2. There are no reciprocal directed paths or feedback loops in the path diagram
- Put another way, the error covariance matrix $\Sigma_\epsilon$ is diagonal, while $B$ matrix is lower-triangular

Estimation for recursive models

- As a consequence of the two properties of recursive models, the predictors are always independent of the error, and the model can be estimated by a sequence of OLS regressions
- SEMs that are not recursive are termed nonrecursive
- There are also block recursive SEMs

Instrumental Variables

- Instrumental-variable (IV) estimation serves two purposes: check whether the model is identifiable and estimate the structural coefficients if it is
- An instrument variable is a variable uncorrelated with the error of a structural equation AND correlated with an exogenous variable
Simple regression

- To understand the IV approach to estimation, consider the following simple linear regression

\[ y = \beta x + \epsilon \]

where \( E(\epsilon) = 0 \), \( \text{var}(\epsilon) = \sigma^2_\epsilon \), \( x \) and \( \epsilon \) are independent.

- Now multiply both sides of the model by \( x \) and take expectations,

\[ \text{cov}(x, y) = \beta \text{var}(x) + \text{cov}(x, \epsilon) \]

\[ \sigma_{xy} = \beta \sigma^2_x + 0 \]

- Plug in consistent sample estimates and solve for \( \beta \)

\[ b = \frac{s_{xy}}{s^2_x} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \]

Instrumental-variable estimation in matrix form

- Now consider

\[ y = X \beta + \epsilon \]

where \( \epsilon \sim N_n(0, \sigma^2_\epsilon I_n) \).

- When \( X \) and \( \epsilon \) are independent, \( b_{\text{OLS}} = (X'X)^{-1}X'y \)

- When \( X \) and \( \epsilon \) are NOT independent, suppose we have observations on \((k + 1)\) instrumental variables \( Z \), that are independent of \( \epsilon \), then follow the scalar treatment,

\[ b_{\text{IV}} = (Z'X)^{-1}Z'y \]

is a consistent estimator of \( \beta \)

IV with simple regression

- Imagine, alternatively, that \( x \) and \( \epsilon \) are not independent, but \( \epsilon \) is independent of some other variable \( z \)

- Suppose further that \( z \) and \( x \) are correlated, that is, \( \text{cov}(z, x) \neq 0 \)

- Then, proceed as before but with \( z \),

\[ \text{cov}(z, y) = \beta \text{cov}(z, x) + \text{cov}(z, \epsilon) \]

\[ \sigma_{zy} = \beta \sigma_{zx} + 0 \]

\[ \beta = \frac{\sigma_{zy}}{\sigma_{zx}} \]

\[ b_{\text{IV}} = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})} \]

Identification problem

- SEM is under-identified if there are fewer instrumental variables than predictors

- SEM is just-identified if number of IVs is the same as predictors

- SEM is over-identified if there are more IVs than predictors, we can either discard surplus IVs, or use better method such as two-stage least squares

For \( b_{\text{IV}} \) to be defined, in addition to at least \( (k + 1) \) IVs, we also need \( Z'X \) to be non-singular

It requires IVs are correlated with predictors plus there is no perfect collinearity
Estimation of recursive SEMs

- By its definition, pool of IVs for recursive SEMs contains exogenous variables and prior endogenous variables.
- Always have at least as many IVs as predictors, therefore necessarily identified.
- To understand this, consider Blau and Duncan’s basic-stratification model, The American Occupational Structure (1967).

Two-stage least squares (2SLS) estimation

- Using combination of IVs for estimation in over-identified non-recursive SEMs.
- First stage, regress predictors $X$ on the IVs $Z$, obtaining fitted values:
  $$\hat{X} = X(Z'Z)^{-1}Z'X$$
- Second stage, the response $y$ is regressed on $\hat{X}$, producing the 2SLS estimator of $\beta$:
  $$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$
- Column of $X$ are uncorrelated with the structural disturbance in the probability limit.
- Very similar to weighted least squares!

Blau and Duncan’s basic-stratification model

- Along with other standard assumptions of SEMs, FIML estimates are calculated under the assumption that the structural errors are multivariately normally distributed.
- Under this assumption, the log-likelihood for the model is:
  $$\log p(B, f, \Sigma_x) = -\frac{nq}{2} \log (2\pi) - \frac{n}{2} \log \det (\Sigma_x) - \frac{1}{2} \sum_{i=1}^{n} (y_i - x_i\beta-y_i)^'\Sigma^{-1}_{\epsilon}\epsilon_i (y_i - x_i\beta-y_i)$$
- The full general machinery of MLE is available if the model is identifiable.
Klein's model revisited

```r
library(sem)
data(Klein)
head(Klein)
```

```
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 35.4 5.1 192.7 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
```

```
P.lag <- c(NA, P[-length(P)])
X.lag <- c(NA, X[-length(X)])
A <- Year - 1931
cbind(Year, A, P, P.lag, X, X.lag)
```

```
Year A P P.lag X X.lag
[1,] 1920 -11 12.7 NA 44.9 NA
[2,] 1921 -10 12.4 12.7 45.6 44.9
[3,] 1922 -9 16.9 12.4 50.1 45.6
[4,] 1923 -8 18.4 16.9 57.2 50.1
[5,] 1924 -7 19.4 18.4 57.1 57.2
[6,] 1925 -6 20.1 19.4 61.0 57.1
```

```
eqn.1 <- tsls(C~P+P.lag+I(Wp+Wg),
+ instruments=~G+T+Wg+A+P.lag+K.lag+X.lag, data=Klein)
summary(eqn.1)
```

```
2SLS Estimates
Model Formula: C ~ P + P.lag + I(Wp + Wg)
Instruments: ~G + T + Wg + A + P.lag + K.lag + X.lag
Residuals:
  Min. 1st Qu. Median Mean 3rd Qu. Max.
-1.89e+00 -6.16e-01 -2.46e-01 -2.74e-12 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
```

Duncan, Haller, and Portest peer influence model

```
R.DHP <- read.moments(diag=FALSE, names=c('ROccAsp', 'REdAsp',
+ 'FOccAsp', 'FEdAsp', 'RParAsp', 'RIQ', 'RSES', 'FSES', 'FIQ',
+ 'FParAsp'))
```

```
1: .6247
2: .3269 .3669
4: .4216 .4275 .6404
7: .2197 .2742 .1124 .0893
11: .3605 .3903 .2998 .1839
16: .3240 .4047 .3054 .2766 .0489 .2220
22: .2930 .2407 .4105 .3607 .0186 .1861 .2707
29: .2995 .2863 .5191 .5007 .0782 .3355 .2202 .2950
37: .0760 .0702 .2784 .1988 .1147 .1021 .0931 -.0438 .2087
46:
Read 45 items
```
Duncan, Haller, and Portest peer influence model

> model.dhp <- specify.model()
1: RParAsp -> RGenAsp, gam11, NA
2: RIQ -> RGenAsp, gam12, NA
3: RSES -> RGenAsp, gam13, NA
4: FSES -> RGenAsp, gam14, NA
5: RSES -> FGenAsp, gam23, NA
6: FSES -> FGenAsp, gam24, NA
7: FIQ -> FGenAsp, gam25, NA
8: FParAsp -> FGenAsp, gam26, NA
9: FGenAsp -> RGenAsp, beta12, NA
10: RGenAsp -> FGenAsp, beta21, NA
11: RGenAsp -> ROccAsp, NA, 1
12: RGenAsp -> REdAsp, lam21, NA
13: FGenAsp -> FOccAsp, NA, 1
14: FGenAsp -> FEdAsp, lam42, NA
15: RGenAsp <-> RGenAsp, ps11, NA

> sem.dhp <- sem(model.dhp, R.DHP, 329,
+ fixed.x=c('RParAsp', 'RIQ', 'RSES', 'FSES', 'FIQ', 'FParAsp'))
> summary(sem.dhp)

Model Chisquare = 26.697 Df = 15 Pr(>Chisq) = 0.031302
Chisquare (null model) = 872 Df = 45
Goodness-of-fit index = 0.98439
Adjusted goodness-of-fit index = 0.94275
RMSEA index = 0.048759 90% CI: (0.014517, 0.07831)
Bentler-Bonnett NFI = 0.96938
Tucker-Lewis NNFI = 0.95757
Bentler CFI = 0.98586
SRMR = 0.020204
BIC = -60.244

Parameter Estimates

Estimate Std Error z value Pr(>|z|)
gam11 0.161224 0.038487 4.1890 2.8019e-05 RGenAsp <--- RParAsp
...

A path diagram

The following path diagram was generated by path.diagram() function in sem package

General structural equation models

- Include unobserved exogenous or endogenous variables (also termed factors or latent variables) in addition to unobservable disturbances
- Sometimes called LISREL models (linear structural relations), after first widely available computer program (Jøreskog, 1973)
- Mainly likelihood based estimation
- No simple general solution towards identification
- There are many ways to fool yourself with SEMs