Structural Equation Modeling
An Econometrician's Introduction

PD Dr. Stefan Klößner

Winter Term 2018/19

Overview

- NOT: easy to digest introduction for practitioners
- instead: theoretical foundations of SEM
- many formulas, using a lot of the 'Greeks', i.e. Greek symbols used to denote quantities appearing in SEM models
- aims to make the econometrically trained learn what SEM 'theoretically' is
- aims to explain why (one branch of) SEM is called 'covariance-based'
- tries to gradually consider increasingly complex models, moving from simple linear regression models to SEM models
- NOT: detailed discussion of all aspects of SEM

Recap: Simple Linear Regression I

- the simple linear regression models looks as follows:

$$y = \alpha + \beta x + u,$$

  where y, x, u are random (!) variables which are called, respectively,
  - regressand, dependent variable, or endogeneous variable: y
  - regressor, independent variable, exogeneous variable: x
  - error term, disturbance: u
- α and β are unknown parameters which describe, respectively,
  - the intercept of the regression line: α
  - the slope of the regression line: β

Recap: Simple Linear Regression II

- given data $(y_1, x_1), \ldots, (y_T, x_T)$ (with T denoting the sample size), α and β are usually estimated by OLS, with formulas
  - $\hat{\beta} = \frac{\sum_{t=1}^{T}(x_t - \overline{x})(y_t - \overline{y})}{\sum_{t=1}^{T}(x_t - \overline{x})^2} = \frac{s_{xy}}{s_x^2}$
  - $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$
- notice that $\hat{\beta}$ can be calculated using demeaned data $(x_t - \overline{x},\, y_t - \overline{y})$ only, while $\hat{\alpha}$ depends on the means $\overline{y}, \overline{x}$ of y, x (as well as $\hat{\beta}$)
- therefore, if one is only interested in β, the parameter describing the relation between y and x, then one may calculate the estimate $\hat{\beta}$ using the demeaned versions of x and y

Recap: Multiple Linear Regression I

- the multiple linear regression model looks as follows:

$$y = \alpha + \beta_1 x_1 + \ldots + \beta_n x_n + u,$$

  where y, $x_1, \ldots, x_n$ are random variables which are called, respectively,
  - regressand, dependent variable, or endogeneous variable: y
  - regressors, independent variables, exogeneous variables: $x_1, \ldots, x_n$
  - error term, disturbance: u
- β_i ($i = 1, \ldots, n$) determines by how much y is expected to change if x_i changes by one unit and all other regressors stay unchanged

Recap: Multiple Linear Regression II

- given data $(y_1, x_{11}, \ldots, x_{n1}), \ldots, (y_T, x_{1T}, \ldots, x_{nT})$ (with T denoting the sample size), α, $\beta_1, \ldots, \beta_n$ are usually estimated by OLS, satisfying the following formulas:
  - $$\begin{pmatrix} \widehat{\beta_1} \\ \vdots \\ \widehat{\beta_n} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}^{-1} \begin{pmatrix} s_{1y} \\ \vdots \\ s_{ny} \end{pmatrix},$$
  - $\widehat{\alpha} = \overline{y} - (\widehat{\beta_1}\overline{x_1} + \ldots + \widehat{\beta_n}\overline{x_n})$, with
  - $s_{ij} := \frac{1}{T}\sum_{t=1}^{T}(x_{it} - \overline{x_i})(x_{jt} - \overline{x_j}) \quad (i, j = 1, \ldots, n)$,
  - $s_{iy} := \frac{1}{T}\sum_{t=1}^{T}(x_{it} - \overline{x_i})(y_t - \overline{y}) \quad (i = 1, \ldots, n)$.

Recap: Multiple Linear Regression III

- notice that $\widehat{\beta}_1, \ldots, \widehat{\beta}_n$ can be calculated using demeaned data ($x_{it} - \overline{x}_i$ ($i = 1, \ldots, n$), $y_t - \overline{y}$) only, while $\widehat{\alpha}$ depends on the means $\overline{y}, \overline{x}_i$ ($i = 1, \ldots, n$) of the variables (as well as $\widehat{\beta}_1, \ldots, \widehat{\beta}_n$)
- therefore, if one is only interested in regression coefficients $\beta_1, \ldots, \beta_n$, the parameters describing the relations between the regressand y and the regressors $x_1, \ldots, x_n$, then one may calculate the estimates $\widehat{\beta}_1, \ldots, \widehat{\beta}_n$ using the demeaned versions of regressand and regressors

Simultaneous Equations I

- we can generalize the multiple linear regression model with respect to the number of dependent variables by allowing more than one dependent variable
- such models are called 'simultaneous equation models', they look as follows (with m denoting the number of dependent variables):

$$
\begin{aligned}
y_1 &= \alpha_1 + \beta_{12} y_2 + \ldots + \beta_{1m} y_m + \gamma_{11} x_1 + \ldots + \gamma_{1n} x_n + u_1 \\
y_2 &= \alpha_2 + \beta_{21} y_1 + \beta_{23} y_3 + \ldots + \beta_{2m} y_m + \gamma_{21} x_1 + \ldots + \gamma_{2n} x_n + u_2 \\
&\vdots \\
y_m &= \alpha_m + \beta_{m1} y_1 + \ldots + \beta_{m,m-1} y_{m-1} + \gamma_{m1} x_1 + \ldots + \gamma_{mn} x_n + u_m
\end{aligned}
$$

Simultaneous Equations II

- denoting $y := (y_1, \ldots, y_m)'$, $x := (x_1, \ldots, x_n)'$, $u := (u_1, \ldots, u_m)'$, $\alpha := (\alpha_1, \ldots, \alpha_m)'$,

$$B := \begin{pmatrix} 0 & \beta_{12} & \ldots & \ldots & \beta_{1m} \\ \beta_{21} & 0 & \beta_{23} & \ldots & \beta_{2m} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \beta_{m-1,1} & & \ddots & \ddots & \beta_{m-1,m} \\ \beta_{m1} & \ldots & \ldots & \beta_{m,m-1} & 0 \end{pmatrix} \in \mathbb{R}^{m \times m},$$

$\Gamma := (\gamma_{ij})_{i=1,\ldots,m, j=1,\ldots,n} \in \mathbb{R}^{m \times n}$, this can be written in compact form:

$$y = \alpha + B\,y + \Gamma\,x + u$$

Simultaneous Equations III

- estimating simultaneous equation models, though typically not done using OLS, usually results in $\widehat{\alpha} = \overline{y} - \widehat{B}\,\overline{y} - \widehat{\Gamma}\,\overline{x}$, while the estimates $\widehat{B}$ and $\widehat{\Gamma}$ can be calculated using the demeaned data $y - \overline{y}$ and $x - \overline{x}$
- $y = \alpha + B\,y + \Gamma\,x + u$ is called the structural form, while

$$y = \underbrace{(I - B)^{-1}\alpha}_{\widetilde{\alpha}} + \underbrace{(I - B)^{-1}\Gamma}_{\widetilde{\Gamma}}\,x + \underbrace{(I - B)^{-1}u}_{\widetilde{u}}$$

  is called the reduced form
- $y_1, \ldots, y_m$ are called endogeneous, $x_1, \ldots, x_n$ are called exogeneous

SEM or Simultaneous Equations with Latent Variables I

- an SEM model is 'simply' a simultaneous equations model where, unfortunately, the endogeneuos and exogeneous variables are latent, i.e. not observable
- in SEM models,
  - the m (latent) endogeneous variables are usually called η
  - the n (latent) exogeneous variables are usually called ξ
  - the m-dimensional error term is usually called ζ
  - the relations between the latent variables are modeled by

$$\eta = B\,\eta + \Gamma\,\xi + \zeta \quad \text{or} \quad \eta = \alpha_\eta + B\,\eta + \Gamma\,\xi + \zeta,$$

with the assumptions $E(\zeta) = 0$, $I - B$ non-singular, and ξ uncorrelated to ζ. The intercept term α_η is only included when means of the latent variables are to be considered, too.

SEM or Simultaneous Equations with Latent Variables II

- the so-called structural equation $\eta = (\alpha_\eta +) B\,\eta + \Gamma\,\xi + \zeta$ must be accompanied by so-called measurement equations which relate the latent variables ξ and η to their observable counterparts x and y:

  $$x = \Lambda_x\,\xi + \delta,\ y = \Lambda_y \eta + \epsilon, \quad \text{or} \quad x = \alpha_x + \Lambda_x\,\xi + \delta,\ y = \alpha_y + \Lambda_y \eta + \epsilon,$$

  where
  - x and y, often called indicators (for ξ and η), consist of q and p observable variables, respectively,
  - $\Lambda_x \in \mathbb{R}^{q \times n}$ and $\Lambda_y \in \mathbb{R}^{p \times m}$ are matrices which contain the so-called factor loadings
  - δ and ϵ are q- and p-dimensional error terms.
  - the intercept terms α_x and α_y are included only if means are to be considered, too.

SEM or Simultaneous Equations with Latent Variables III

- all error terms are assumed to have zero mean
- usually, ξ, x, η, y are also assumed to have zero mean: in this case, no intercept terms appear in the above equations, and demeaned data are used (this does not hold, though, for instance for so-called multi-group or latent curve models)
- it is assumed that both ϵ and δ are uncorrelated with ζ and ξ, and that the two error terms ϵ and δ are uncorrelated
- often, the covariance matrices are called Φ (for ξ), Ψ (for ζ), Θ_ϵ (for ϵ), and Θ_δ (for δ)
- however, the notations Σ_ξ, Σ_ζ, Σ_ϵ, and Σ_δ are much more intuitive and will be used in the sequel

Summary of SEM Equations, Quantities, and Assumptions

- structural equation: $\eta = (\alpha_\eta +) B\,\eta + \Gamma\,\xi + \zeta$
- measurement equations:
  $x = (\alpha_x +)\Lambda_x\,\xi + \delta$, $y = (\alpha_y +)\Lambda_y \eta + \epsilon$
- ξ: n latent exogeneous random variables with covariance matrix Σ_ξ
- η: m latent endogeneous random variables
- x, y: q- and p-dimensional observable random variables, indicators of ξ and η
- ζ, δ, ϵ: m-, q-, p-dimensional error terms with covariance matrices Σ_ζ, Σ_δ, Σ_ϵ
- it is assumed that $\Sigma_{\xi\zeta} = 0$, $\Sigma_{\epsilon\delta} = 0$, $\Sigma_{\xi\epsilon} = 0$, $\Sigma_{\xi\delta} = 0$, $\Sigma_{\zeta\epsilon} = 0$, $\Sigma_{\zeta\delta} = 0$
- fixed, but unknown quantities (model parameters): B, Γ, Λ_x, Λ_y; when means are considered, additionally α_η, α_x, α_y

Fundamental Theorem of SEM I

- under the assumptions stated above,
  - the covariance matrix Σ_x of the latent exogeneous variables' observed indicators x is given by:

$$\Sigma_x = \Lambda_x \Sigma_\xi \Lambda_x' + \Sigma_\delta$$

  - the covariance matrix Σ_{xy} between the latent exogeneous variables' obeserved indicators x and the latent endogeneous variables' observed indicators y is given by:

$$\Sigma_{xy} = \Lambda_x \Sigma_\xi \Gamma' (I - B')^{-1} \Lambda_y'$$

  - the covariance matrix Σ_x of the latent endogeneous variables' observed indicators y is given by:

$$\Sigma_y = \Lambda_y (I - B)^{-1} \left(\Gamma \Sigma_\xi \Gamma' + \Sigma_\zeta \right) (I - B')^{-1} \Lambda_y' + \Sigma_\epsilon$$

Fundamental Theorem of SEM II

- the unknown parameters, i.e. the structural parameters contained in B and Γ, the factor loadings contained in Λ_x and Λ_y, as well as the variances and covariances contained in Σ_ξ, Σ_ζ, Σ_δ, and Σ_ϵ are determined such that the differences between the model-implied covariances delivered by the fundamental theorem and the empirical estimates $\widehat{\Sigma}_x$, $\widehat{\Sigma}_{xy}$, and $\widehat{\Sigma}_y$ are as small as possible
- estimating the parameters by matching empirical and model-implied covariances can be done by using the empirical covariances of the observed variables only, therefore software packages often do not only accept raw data, but can also simply be given the empirical covariances
- for these reasons, one speaks of 'covariance-based' estimation of SEM models